1. General Remarks

1.1 I am glad to have this opportunity to address some of the criticisms that have been aimed at arguments in my book *Shadows of the Mind* (henceforth *Shadows*). I hope that in the following remarks I am able to remove some of the confusions and misunderstandings that still surround the arguments that I tried to make in that book - and also that we may be able to move forward from there.

1.2 In the accompanying PSYCHE articles, the great majority of the commentators' specific criticisms have been concerned with the purely logical arguments given in Part 1 of *Shadows*, with comparatively little reference being made to the physical arguments given in Part 2 - and virtually none at all to the biological ones.<1> This is not unreasonable if it is regarded that the entire rationale for my physical and biological arguments stands or falls with my purely logical arguments. Although I do not entirely agree with this position - since I believe that there are strong motivations from other directions for the kinds of physical and biological action that I have been promoting in *Shadows* - I am prepared to go along with it for the moment. Thus, most of my remarks here will be concerned with the implications of Gödel's theorem, and with the claims made by many of my critics that my arguments do not actually establish that there must be a noncomputational ingredient in human conscious thinking.

1.3 In replying to these arguments, I should first point out that, very surprisingly, almost none of the commentators actually addresses what I had regarded as the central (new) core argument against the computational modelling of mathematical understanding! Only Chalmers actually draws attention to it, and comments in detail on this argument, remarking that "most commentators seem to have missed it".<2> Chalmers also remarks that "it is unfortunate that this argument was so deeply buried". I apologize if this appears to have been the case; but I am also very puzzled, since its essentials are summarized in the final arguments of "Reductio ad absurdum - a fantasy dialogue", which is the section of *Shadows* (namely Section 3.23) that readers are particularly directed towards. This section is referred to also by McDermott and by Moravec, but neither of these commentators actually addresses this central argument explicitly, and nor do any of the other commentators. This is particularly surprising in the case of McCullough, as he is concerned with some of the subtleties of the logic involved, and also of Feferman, in view of his very carefully considered logical discussion.

1.4 It would appear, therefore, that I have an easy solution to the problem of replying to all nine commentators. All I need do is show why the ingenious argument put forward by Chalmers (based partly on McCullough's very general considerations) as a counter to my central argument is in fact (subtly) invalid! However, I am sure that this mode of procedure would satisfy none of the other commentators, and many of them also have interesting other points to make which need commenting upon. Accordingly, in the following remarks, I shall attempt to address all the serious points that they do bring up. My reply to this main argument of Chalmers (partly dependent upon that of McCullough) will be given in Section 3, but it will be helpful first to precede this by addressing, in Section 2, the significant logical points that are raised by Feferman in his careful commentary.

2. Some Technical Slips in *Shadows*

2.1 Feferman quite correctly draws attention to some inaccuracies in *Shadows* with regard to certain logical technicalities. The most significant of these (in fact, the only really significant one for my actual arguments) concerns a misunderstanding on my part with regard to the assertion of omega-consistency of a formal system F, which I had chosen to denote by the symbols Omega(F), and its relation to Gödel's first incompleteness theorem. (As it happens, two others before Feferman had

also pointed out this particular error to me.) As Feferman says, the assertion that some particular formal system is "omega-consistent" is certainly not of the form of a PI_1-sentence (i.e. not of the form of an assertion: "such-and-such a Turing computation never halts" - I call these "P-sentences" from here on). This much I should have been (and essentially was) aware of, despite the fact that in the first two printings of *Shadows*, p.96 I made the assertion that Omega(F) is a P-sentence. The fact of the matter was that I had somehow (erroneously) picked up the belief that the statement that Gödel originally exhibited in his famous first incompleteness theorem was *equivalent* to the omega-consistency of the formal system in question, not that it merely *followed* from this omega-consistency. Accordingly, I had imagined that for some technical reason I did not know of, this omega-consistency must actually be equivalent (for sufficiently extensive systems F) to the particular assertion "C_k(k)" that I had exhibited in Section 2.5, when the rules of the formal system F are translated into the algorithm A. Accordingly, I had mistakenly believed that Omega(F) must, for some subtle reason (unknown to me), be equivalent to the P-sentence C_k(k) (at least for sufficiently extensive systems F).

2.2 This error affects none of the essential arguments of the book but it is unfortunate that in various parts of Chapter 3, and most particularly in the "fantasy dialogue" in Section 3.23, the notation "Omega(F)" is used in circumstances where I had intended this to stand for the actual P-sentence C_k(k). In later printings of *Shadows*, this error has been corrected: I use the Gödel sentence G(F) (which asserts the consistency of F and is a P-sentence) in place of Omega(F). It is in any case much more appropriate to use G(F) in the arguments of Chapter 3, rather than Omega(F), and I agree with Feferman that the introduction of "Omega(F)" was essentially a red herring. In fact, the presentation in *Shadows* would have usefully simplified if omega-consistency had not even been mentioned.

2.3 The next most significant point of inaccuracy - or rather imprecision - in *Shadows* that Feferman brings up is that there is a discrepancy between different notions of the term "sound" that I allude to in different parts of the book. (This is actually quite an important issue, in relation to some of the discussion to follow, and I shall need to return to it later in Section 3.) His point is, essentially, that in some places I need make use of the soundness of a formal system only in the limited sense of its capacity to assert the truth of certain P-sentences, whereas in other places I am actually referring to soundness in a more comprehensive sense, where it applies to other types of assertion as well. I agree that I should have been more careful about such distinctions. In fact, it is the weaker notion of soundness that would be sufficient for all the "Gödelian" arguments that I actually use in Part 1 of *Shadows*, though for some of the more philosophical discussions, I had in mind soundness in a stronger sense. (This stronger sense is not needed on pp. 90-92 if omega-consistency is dropped; nor is it needed on p.112, the weaker notion of soundness now being equivalent to consistency.)

2.4 Basically, I am happy to agree with all the technical criticisms and corrections that Feferman refers to in his section discussing my treatment of the logical facts". (I should attempt a point of clarification concerning his puzzlement as to why I should make the "strange" and "trivial" assertions he refers to on p.112. No doubt I expressed myself badly. The point that I was attempting to make concerned the issue of the relationship between the formal string of symbols that constitute "G(F)" and "Omega(F)" and the actual *meanings* that these strings are supposed to represent. I was merely trying to argue that meanings are essential - a point with which Feferman strongly concurs, in his commentary.) It should be made clear that none of these corrections affects the arguments of Chapter 3 in any way (so long as Omega(F) is replaced by G(F) throughout), as Feferman himself appears to affirm in his last paragraph of the aforementioned section.

2.5 I find it unfortunate, however, that he does not offer any critique of the arguments of Chapter 3. I would have found it very valuable to have had the comments of a first-rate logician such as

himself on some of the specifics of the discussions in Chapter 3. Feferman seems to be led to having some unease about the arguments presented there, not because of specific errors that he has detected, but merely because my "slapdash scholarship" may be "stretched perilously thin in areas different from [my] own expertise". A related point is made by McCarthy, McDermott and Baars in connection with my evidently inadequate referencing of the literature on AI, and on other theories that relate to consciousness, either in its computational, biological, or psychological respects.

2.6 I think that a few words of explanation, from my own vantage point, are necessary here. An ability to search thoroughly through the literature has never been one of my strong points, even in my own subject (whatever that might be!). My method of working has tended to be that I would gather some key points from the work of others and then spend most of my time working entirely on my own. Only at a much later stage would I return to the literature to see how my evolved views might relate to those of others, and in what respects I had been anticipated or perhaps contradicted. Inevitably I shall miss things and get some things wrong. The most likely source of error tends to be with second-hand information, where I might misunderstand what someone else tells me when reporting on the work of a third person. Gradually these things sort themselves out, but it takes time.

2.7 My reason for mentioning this is to emphasize that errors of the nature of those pointed out by Feferman are concerned essentially with this link of communication with the outside (scientific, philosophical, mathematical, etc.) world, and not with the internal reasonings that constitute the essential Gödelian arguments of *Shadows*. Most specifically, the main parts of Chapter 3 (particularly 3.2, 3.3 and 3.5-3.24) are entirely arguments that I thought through on my own, and are therefore independent of however "slapdash" my scholarship might happen to be! I trust that these arguments will be judged entirely on their intrinsic merits.

3. The Central New Argument of *Shadows*

3.1 Chalmers provides a succinct summary of the central new argument that I presented in *Shadows* (Section 3.16, and also 3.23 and 3.24 - but recall that my Omega(F) should be replaced by G(F) throughout Section 3.16 and 3.23). Let me repeat the essentials of Chalmers's presentation here - but with one important distinction, the significance of which I shall explain in a moment.

3.2 We try to suppose that the totality of methods of (unassailable) mathematical reasoning that are in principle humanly accessible can be encapsulated in some (not necessarily computational) sound formal system F. A human mathematician, if presented with F, could argue as follows (bearing in mind that the phrase "I am F" is merely a shorthand for "F encapsulates all the humanly accessible methods of mathematical proof"):

(A) "Though I don't know that I necessarily am F, I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F' is F supplemented by the further assertion "I am F". I perceive that it follows from the assumption that I am F that the Gödel statement G(F') would have to be true and, furthermore, that it would not be a consequence of F'. But I have just perceived that "if I happened to be F, then G(F') would have to be true", and perceptions of this nature would be precisely what F' is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of F', I deduce that, I cannot be F after all. Moreover, this applies to any other (Gödelizable) system, in place of F."

3.3 (Of course, one might worry about how an assertion like "I am F" might be made use of in a logical formal system. In effect, this is discussed with some care in *Shadows*, Sections 3.16 and 3.24, in relation to the Sections leading up to 3.16, although the mode of presentation there is

somewhat different from that given above, and less succinct.)

3.4 The essential distinction between the above presentation and that of Chalmers is that he makes use (in the first (2) of his Section 2) of the stronger conditional assumption "I know that I am F", rather than merely "I am F", the latter being all that I need for the above. Thus, if we accept the validity of the above argument, the conclusion is considerably stronger than the "strong" conclusion that Chalmers draws ("threatening to the prospects of AI") to the effect that it "would rule out even the possibility that we could empirically discover that we were identical to some system F".

3.5 In fact, it *was* this stronger version (A) that I presented in *Shadows*, from which we would conclude that we cannot be identical to any knowable (Gödelizable) system F whatever, whether we might empirically come to believe in it or not! I am sure that this stronger conclusion would provide an even greater motivation for people (whether AI supporters or not) to find a flaw in the argument. So let me address the particular objection that Chalmers (and, in effect, also McCullough) raises against it.

3.6 The problem, according to Chalmers, is that it is contradictory to "know that we are sound". Accordingly, he argues, it would be invalid to deduce the soundness of F, let alone that of F', from the assumption "I am F". On the face of it, to a mathematician, this seems an unlikely let-out, since in all the above discussions we are referring simply to the notion of *mathematical proof*. Moreover, the "I" in the above discussion refers to an idealized human mathematician. (The problems that this notion raises, such as those referred to by McDermott, are not my concern at the moment. I shall return to such matters later; cf. Section 6.) Suppose that F indeed represents the totality of the procedures of mathematical proof that are in principle humanly accessible. Suppose, also, that we happen to come across F and actually entertain this possibility that we might "be" F, in this sense (without actually knowing, for sure, whether or not we are indeed F). Then, under the assumption that it *is* F that encapsulates all the procedures of valid mathematical proof, we must surely conclude that F is sound. The whole point of the procedures of mathematical proof is that they instil *belief*. And the whole point of the Gödel argument, as I have been employing it, is that a belief in the conclusions that can be obtained using some system H entails, also, a belief in the soundness and consistency of that system, together with a belief (for a Gödelizable H) that this consistency cannot be derived using H alone.

3.7 This notwithstanding, Chalmers and McCullough argue for an inconsistency of the very notion of a "belief system" (which, as I have pointed out above, simply means a system of procedures for mathematical proof) which can believe in itself (which means that mathematicians actually trust their proof procedures). In fact, this conclusion of inconsistency is far too drastic, as I shall show in a moment. The key issue is not that belief systems are inconsistent, or incapable of trusting themselves, but that they must be restricted as to what kind of assertion they are competent to address.

3.8 To show that "a belief system which believes in itself" need not be inconsistent, consider the following. We shall be concerned just with P-sentences (which, we recall, are assertions that specified Turing machine actions do not halt). The belief system B, in question, is simply the one which "believes" (and is prepared to assert as "unassailably perceived") a P-sentence S if and only if S happens to be true. B is not allowed to "output" anything other than a decision as to whether or not a suggested P-sentence is true or false - or else it may prattle on, as is its whim, generating P-sentences together with their correct truth values. However, as part of its internal musings, it is allowed to contemplate other kinds of thing, such as the fact that it *does*, indeed, produce only truths in its decisions about P-sentences. Of course, B is not a computational system - it is a Turing *oracle* system, as far as its output is concerned - but that should not matter to the argument. Is there

anything wrong in B "believing in the soundness of B"? Nothing whatever, if we interpret this in the right way. The important thing is that B is allowed only to make assertions about P-sentences. It can use whatever procedures it likes in its internal musings, but all its *outputs* must be assertions as to the validity of particular P-sentences. If we apply the diagonal procedure that Chalmers and McCullough refer to, then we get something which is not a P-sentence, and is accordingly not allowed to be part of this belief system's output.

3.9 It may be felt that this is a pretty limited kind of "belief system", where it can make assertions only about the truth or falsity of P-sentences. Perhaps it is limited; but it is precisely a belief system of this very kind that comes into the arguments of Chapter 3 of *Shadows*. In that discussion, I was careful, in the key Section 3.16 of *Shadows*, to limit the mathematical assertions under consideration to P-sentences. This avoids many difficult issues that can arise without such restrictions. However, the robots described in that section are allowed to think in very general terms - as human mathematicians may do - about non-computable systems and uncountable cardinals, etc. Nevertheless, the *-assertions under consideration must always be P-sentences, and it is *only* in relation to such sentences (as outputs) that the formal systems Q(M) and Q_M(M) are constructed. In this circumstance the argument serves to show that the robots' "belief system" cannot, after all, be a computational one, provided that it is broad enough to encompass Gödelian reasoning - which is a contradiction with the notion of "robot" that was being used.

3.10 This is not to say that the diagonalization procedure that McCullough and Chalmers refer to need apply only to computational belief systems F. As they both argue (particularly McCullough), there is no requirement that F be computational in their discussions. Indeed, in Section 7.9 of *Shadows* (which is in Part 2, so it is easy to miss, if one is concerned only with the logical arguments of that book - and neither McCullough nor Chalmers actually mention it), I explicitly referred to the fact that the Gödel-type diagonalization arguments of Part 1 will apply much more generally than merely to computational systems. For example, if Turing's oracle-computation notions are adopted, then the diagonalization procedures are quite straight-forward. However, in any specific application, it is necessary to restrict the class of sentences to which the notion of "unassailable belief" can be applied. If we do not do this, we can land in paradox, which is exactlythe situation that McCullough and Chalmers find themselves in.

3.11 Indeed, McCullough actually carries through such paradoxical reasoning in his Section 2.1, seeming to be presenting this parody of my own reasoning as though it were actually my own reasoning. This is beneath his usual standards. It would have been more helpful if he had addressed the arguments as I actually presented them.

3.12 Returning to the argument (A), we now see how to avoid the inherent difficulties that occur with a belief system with an unrestricted domain. A sufficient thing to do is to make sure that the word "sound" is interpreted in the restricted sense which applies only to P-sentences - as was indeed done in *Shadows*, Section 3.16. (Recall the discussion of Section 2, above, in which Feferman draws attention to possible differences of interpretation of that word.) This provides the needed argument against computationalism, and it is not subject to the objection brought forward by Chalmers in his discussion of my "second argument" in his Section 2.

3.13 Of course, as in Section 7.9 of *Shadows* and as in McCullough's discussion, it is possible to repeat this argument at a higher level. Rather than restricting attention to P-sentences (that is, PI_1 sentences), we could use PI_2-sentences, say (cf. Feferman's commentary). The diagonal process can be applied, but it does not yield a PI_2-sentence, so contradiction (of the Chalmers/McCullough type - to a self-believing belief system) is again avoided. The same argument applies to higher-order sentences. However, it is important to put *some* restriction on the type of sentence to which

the belief system is applied. This kind of thing is very familiar in mathematical logic. One may reason about sets, and about sets of sets, and sets of sets of sets, etc., but one cannot reliably reason about the set of *all* sets. That leads immediately to a contradiction, as Cantor and Russell pointed out long ago. Likewise, a self-believing belief system cannot consistently operate if it is allowed to apply itself to unrestricted mathematical systems. In Section 3.24 of *Shadows*, I tried to explore the tantalizing closeness that my Gödelian reasoning of Section 3.16 seemed at first to have with the Russell-type reasoning that leads to paradox. My conclusion was that the argument of Section 3.16, as I presented it, was not actually of the same nature at all, since the domain of consideration (P-sentences) was indeed sufficiently restricted. I am well aware that the argument can be taken much further than this, and it would be interesting to know how far. Moreover, it would be interesting to have a professional logician's commentary on all these lines of thinking.

4. The "Bare" Gödelian Case

4.1 Although I have concentrated, in the previous section, on what I have referred to as the "central new argument" of *Shadows*, I do not regard that as the "real" Gödelian reason for disbelieving that computationalism could ever provide an explanation for the mind - or even for the behaviour of a conscious brain.

4.2 Perhaps a little bit of personal history on this point would not be amiss. I first heard about the details of Gödel's theorem as part of a course on mathematical logic (from which I also learned about Turing machines) given by the Cambridge logician Steen. As far as I can recall, I was in my first year as a graduate student (studying algebraic geometry) at Cambridge University in 1952/53, and was merely sitting in on the course as a matter of general education (as I did with courses in quantum mechanics by Dirac and general relativity by Bondi). I had vaguely heard of Gödel's theorem prior to that time, and had been a little unsettled by my impressions of it. My viewpoint before that would probably have been rather close to what we now call "strong AI". However, I had been disturbed by the possibility that there might be true mathematical propositions that were in principle inaccessible to human reason. Upon learning the true form of Gödel's theorem (in the way that Steen presented it), I was enormously gratified to hear that it asserted no such thing; for it established, instead, that the powers of human reason could not be limited to any accepted preassigned system of formalized rules. What Gödel showed was how to transcend any such system of rules, so long as those rules could themselves be trusted.

4.3 In addition to that, there was a definite close relationship between the notion of a formal system and Turing's notion of effective computability. This was sufficient for me. Clearly, human thought and human understanding must be something beyond computation. Nevertheless, I remained a strong believer in scientific method and scientific realism. I must have found some reconciliation at the time which was close to my present views - in spirit if not in detail.

4.4 My reason for presenting this bit of personal history is that I wanted to demonstrate that even the "weak" form of the Gödel argument was already strong enough to turn at least one strong-AI supporter away from computationalism. It was not a question of looking for support for a previously held "mystical" standpoint. (You could not have asked for a more rationalistic atheistic anti-mystic than myself at that time!) But the very force of Gödel's logic was sufficient to turn me from the computational standpoint with regard not only to human mentality, but also to the very workings of the physical universe.

4.5 The many arguments that computationalists and other people have presented for wriggling around Gödel's original argument have become known to me only comparatively recently: perhaps we act and perceive according to an unknowable algorithm; perhaps our mathematical

understanding is intrinsically unsound; perhaps we could know the algorithms according to which we understand mathematics, but are incapable of knowing the actual roles that these algorithms play. All right, these are logical possibilities. But are they really plausible explanations?

4.6 For those who are wedded to computationalism, explanations of this nature may indeed seem plausible. But why *should* we be wedded to computationalism? I do not know why so many people seem to be. Yet, some apparently hold to such a view with almost religious fervour. (Indeed, they may often resort to unreasonable rudeness when they feel this position to be threatened!) Perhaps computationalism can indeed explain the facts of human mentality - but perhaps it cannot. It is a matter for dispassionate discussion, and certainly not for abuse!

4.7 I find it curious, also, that even those who argue dispassionately may take for granted that computationalism in some form - at least for the workings of the objective physical universe - *has* to be correct. Accordingly, any argument which seems to show otherwise *must* have a "flaw" in it. Even Chalmers, in his carefully reasoned commentary, seeks out "the deepest flaw in the Gödelian arguments". There seems to be the presumption that whatever form of the argument is presented, it just *has* to be flawed. Very few people seem to take seriously the slightest possibility that the argument might perhaps, at root, be correct! This I certainly find puzzling.

4.8 Nevertheless, I know of many who (like myself) do find the simple "bare" form of the Gödelian argument to be very persuasive. To such people, the long and sometimes tortuous arguments that I provided in *Shadows* may not add much to the case - in fact, some have told me that they think that they effectively weaken it! It might seem that if I need to go to lengths such as that, the case must surely be a flimsy one. (Even Feferman, from his own particular non-computational standpoint, argues that my diligent efforts may be "largely wasted".) Yet, I would claim that some progress has been made. I am struck by the fact that none of the present commentators has chosen to dispute my conclusion G (in *Shadows*, p.76) that "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth". I doubt that any will admit to being persuaded by any of the replies to my queries Q1, ..., Q20, in Section 2.6 and Section 2.10, but it should be remarked that many of these queries represented precisely the kinds of misunderstandings and objections that people had raised against my earlier use of the bare Gödelian argument (and its conclusion G) in *The Emperor's New Mind*, particularly in the many commentaries on that book in *Behavioral and Brain Sciences* (and, in particular, one by McDermott 1990). Perhaps some progress has been made after all!

5. Gödel's "Theorem-Proving Machine"

5.1 Before addressing the important issue of possible errors in human reasoning or the possible "unknowability" of the putative algorithm underlying human mathematical reasoning (which provide the counter-arguments that so many computationalists pin faith on), I should briefly refer to the discussion of Section 3.3 in *Shadows*, which Chalmers regards as "one of the least convincing sections in the book". This is the first of the two arguments of mine that he addresses, but I am not sure that he (or any other of the commentators) has appreciated what I was trying to express. In that section (and also Section 3.8, cf. figure 3.1 on p. 148), I was attempting to show the actual absurdity of the possibility that human understanding (with regard to P-sentences, say) might be encapsulated in what I have referred to as a "Gödel's theorem-proving machine". As quoted on p. 128 of *Shadows*, Gödel seemed not to have been able to rule out the possibility that mathematical understanding might be encapsulated in terms of the action of an algorithm - his "theorem-proving machine" - which, although sound, could not be humanly (unassailably) perceived to be sound. Yet it might be possible to come across this algorithm empirically. I shall refer to this putative "machine" (or algorithm) here as T.

5.2 In Section 3.3, I was concerned with a mathematical algorithm, of the type that might be considered seriously by logicians or mathematicians, so it is not unreasonable to think of T as formulated in the kind of terms which mathematical logicians are familiar with. Of course, even if T were not initially formulated in such terms, it could be if desired. It is sufficient to restrict Gödel's hypothetical theorem-proving machine to be concerned only with P-sentences. Then T would be an algorithmic procedure that generates precisely all the true P-sentences that are perceivably true, in principle, by human mathematicians. Gödel argues that although T might be empirically discoverable, the perception of its soundness would be beyond the powers of human insight. In Sections 3.3 and 3.8, I merely try to make the case that the existence of T is a very far-fetched possibility indeed, especially if we try to imagine how it might have come about (either by natural selection or by deliberate AI construction). But I did not argue that it was an entirely illogical possibility.

5.3 In Feferman's commentary, he refers to Boolos's "cautious" interpretation of the implications of Gödel's theorem that a let-out for computationalism would be the existence of "absolutely unsolvable diophantine problems". Such an absolutely unsolvable problem could be constructed, by well understood procedures, from the algorithm T, if T were to exist. Phrased in these terms, it does not seem at all out of the question that such a T might exist. In Section 3.3, my intention was merely to point out some of the improbable-sounding implications of the existence of T. It seems to me that this does go somewhat beyond what Feferman refers to at the end of his commentary. Moreover, the arguments referred to in Section 2 above (concerning Section 3.16 of *Shadows* that most commentators appear to have missed) certainly do proceed well beyond this interpretation.

5.4 Later in *Shadows* (cf. Sections 3.5-3.23, and especially 3.8), I argue that it is extremely hard to see how an extraordinarily sophisticated algorithm of the nature of T could come about by natural selection (or by deliberate AI construction), even if it could exist. It has to be already capable of correctly dealing with subtle mathematical issues that are, for example, far beyond the capabilities of the Zermelo-Fraenkel axiom system ZF (for example, the Gödel procedure can be applied to ZF to obtain humanly accessible P-sentences that are indeed beyond the scope of ZF). Yet issues of this nature played no role in the selective processes that were operative with our remote ancestors. I would argue that there is nothing wrong with natural selection having been the driving force, so long as it is a non-specific *non-computational* quality such as "understanding" that natural selection has favoured, rather than some improbable algorithm, such as T.<3>

5.5 Even if we do not worry about how T might possibly have come about, there is a distinct implausibility in its very existence, if T were to be an algorithm that could be humanly understood (or "knowable", in the terminology of *Shadows*). This is basically "case II" of *Shadows* (cf. p. 131), where the soundness of T, and certainly its specific role, would not be humanly knowable. The implausibility of such a T was the main point that I was trying to make in Section 3.3. I think Chalmers is arguing that such a T might come about by some bottom-up AI procedures and, if so, it might not look at all like a mathematical formal system. However, in the absence of some strongly held computationalist belief - to the effect that it *must* have been by procedures of this very kind that Nature was able to produce human mathematicians - there is no good reason to expect that this would be a good way of finding such a T (as I argue in *Shadows* Section 3.27), nor is there any reason to expect such a T actually to exist. It was the burden of later sections of *Shadows*, not of Section 3.3, to argue that such bottom-up procedures do not do what is required either. In effect, in these later sections, I argue that if merely the (partly bottom-up) computational *mechanisms* for ultimately leading to a T could be known, then we would indeed be able to construct the formal system that T represents. This will be discussed further in Section 7, below.

6. The Issue of Errors

6.1 Some commentators (particularly McDermott and, in effect, Baars) try to argue that the fact that human mathematicians make errors allows the computational model of the mind to escape the Gödel-type arguments. (This was also apparently Turing's let-out, as illustrated in the quote in *Shadows*, p.129.) I have stressed in many places in *Shadows* that the main arguments of that book (certainly those in Chapter 2) are concerned with what mathematicians are able to perceive *in principle*, by their methods of mathematical proof - and that these methods need not be necessarily constrained to operate within the confines of some preassigned formal system.

6.2 I fully accept that individual mathematicians can frequently make errors, as do human beings in many other activities of their lives. This is not the point. Mathematical errors are in principle correctable, and I was concerned mainly with the ideal of what can indeed be perceived in principle by mathematical understanding and insight. Most particularly, I was concerned with those P-sentences that can be humanly perceived, in principle, i.e., with those which are in principle humanly accessible. The arguments given above, in Sections 3 and 5, were also concerned with this ideal notion only. The position that I have been strongly arguing for is that this ideal notion of human mathematical understanding is something beyond computation.

6.3 Of course, individual mathematicians may well not accord at all closely with this ideal. Even the mathematical community as a whole may significantly fall short of it. We must ask whether it is conceivable that this mathematical community, or its individual members, could be entirely computational entities even though the ideal for which they strive is beyond computation. Put in this way, it may perhaps seem not unreasonable that this could be the case. However, there remains the problem of what the human mathematicians are indeed doing when they seem able to "strive for", and thereby approximate, this non-computational ideal. It is the *abstract idea* underlying a line of proof that they seem able to perceive. They then try to express these abstract notions in terms of symbols that can be written on a page. But the particular collections of symbols that ultimately appear on the pages of their notes and articles are far less important than are the ideas themselves. Often the particular symbols used are quite arbitrary. With time, both the ideas and the symbols describing them may become refined and sometimes corrected. It may not always be very easy to reconstruct the ideas from the symbols, but it is the *ideas* that the mathematicians are really concerned with. These are the basic ingredients that they employ in their search for idealized mathematical proofs. (These matters have relevance to the question of how mathematicians *actually* think,<4> as raised by Feferman in his commentary, and they are related also to issues raised also by Baars and McCullough.)

6.4 Sometimes there may be errors, but the errors are correctable. What is important is the fact is that there *is* an impersonal (ideal) standard against which the errors can be measured. Human mathematicians have capabilities for perceiving this standard and they can normally tell, given enough time and perseverance, whether their arguments are indeed correct. How is it, if they themselves are mere computational entities, that they seem to have access to these non-computational ideal concepts? Indeed, the ultimate criterion as to mathematical correctness is measured in relation to this ideal. And it is an ideal that seems to require use of their conscious minds in order for them to relate to it.

6.5 However, some AI proponents seem to argue against the very existence of such an ideal, a position that Moravec (if his robot is to be trusted as espousing Moravec's own views) seems to be taking in his commentary. Moreover, Chalmers comments: "an advocate of AI might take [the position] that our reasoning is fundamentally unsound, even in idealization". There are others, such as Baars ("I do not believe in the absolute nature of mathematical thought"), who also have

difficulty with this notion, perhaps because their professional interests have more to do with examining the ways in which particular individuals may deviate from such ideals than with the ideal notions themselves. It is common for such people to point to errors that have persisted in the mathematical literature for some while (such as McDermott 's reference to Kempe's erroneous attempt at a proof of the four-colour theorem - which, incidentally provided an important ingredient in the actual proof that was finally arrived at in 1976 by Appel and Haken; cf. Devlin (1988) - or to Frege's inconsistent attempt at building up a formal set theory - which was a good deal more influential, in a very positive sense). But these errors are more in the nature of "correctable errors", and do not really argue against the very existence of a mathematical ideal.

6.6 In *Shadows*, Section 3.2, I did examine, in a serious way, the possibility that mathematical reasoning might be fundamentally unsound. But one should bear in mind that the presumption of mathematical unsoundness is an extremely dangerous position for anyone purporting to be a scientist to take. If our mathematical reasoning were indeed fundamentally unsound, then the whole edifice of scientific understanding would come crashing to the ground! For virtually all of science, at least detailed science, depends upon mathematics in one respect or another. I find it remarkable how frequently attacks on the Gödelian argument seem to degenerate into attacks upon the very basis of mathematics.<5> To attack the notion of "ideal" mathematical concepts or idealized mathematical reasoning is, indeed, to attack the very basis of mathematics. People who do so should at least pause to contemplate the implications of what they are contending.

6.7 While it is true that there are different philosophical standpoints that may be adopted by different mathematicians, this has little effect on the basic Gödelian argument, especially if we restrict attention to P-sentences; see responses to queries Q9-Q13 in Sections 3.6, 3.10 of *Shadows*. For the remainder of my arguments here, I shall take it as read that there *is* an ideal notion of (in principle) humanly accessible mathematical proof, at least with respect to P-sentences, and that this ideal notion of proof is *sound*. (And I am not against there being *more* than one, provided that they are not in contradiction with one another with regard to P-sentences; see *Shadows* Section 3.10, response to Q11.) The question, then, is how serious are the errors which undoubtedly occur when actual human mathematicians attempt toemulate this ideal.

6.8 For the arguments of Chapter 3 of *Shadows*, particularly Sections 3.4, 3.17, 3.19, 3.20, and 3.21, I try to address the issue of errors in purported mathematical arguments, and the question of constructing an error-free formal system from the *actual* output of a manifestly computational system - the hypothetical mathematical robots that I consider for the purpose. The arguments are quite intricate in places, and I do not blame some of the commentators for balking at those sections. On the other hand, it would have been helpful to have had a dispassionate discussion of these arguments in their essential points. McDermott does at least address some of the more technical arguments concerning errors - though I feel it is not altogether appropriate to refer to his account as "dispassionate". More importantly, he does not answer the essential point of my conclusions. If it is to be *errors* that provide the key escape route from the Gödel conundrum, we need to explain the seeming necessity for a "conspiracy" preventing any kind of computational procedure for weeding out all the errors in the merely *finite* set that arises in accordance of the discussion of Section 3.20 (see 3.21 and also the second paragraph of 3.28). In his commentary McDermott does not actually address the argument as I gave it, but goes off on a tangent (about a "computerized Gauss" and the like) which has very little to do with the specific argument provided in *Shadows*. (The same applies to most of his other arguments which, he contends, have "torn [my] argument to shreds". His discussion might have been more convincing had it referred to my actual arguments! I shall make some further comments concerning these matters in Section 7 below.)

6.9 McDermott does, however, come close to expressing the central dilemma presented by the

Gödelian insight - although apparently unwittingly. He has a hard time coming to terms with the fact that mathematical unassailability needs "to be both informal and guaranteed accurate". Although he is unable to "see how that's possible", it is basically this conflict that forces us into a non-computational viewpoint. If by a "guaranteed accurate" notion of unassailability he means something that has been validated by a procedure that is computationally checkable, then this notion would basically have to be one that can indeed be encompassed by a formal system in the ordinary sense. We must bear in mind that the guarantee must apply not only to the correctness of *carrying out* the rules of the procedure (which is where the "computational checkability" of the procedure might have importance), but also to the validity, or *soundness* of the very rules themselves. But if we can guarantee that the rules are sound, we can also guarantee something beyond those rules. The rules would be subject to Gödel's theorem, so there would also be certain P-sentences, such as the Gödel sentence asserting the consistency of the "guaranteeing system", that would be just as "guaranteed" as the things that have already been previously "guaranteed". If McDermott is requiring that "formal" implies "computational", and that "guaranteeable" also implies computational, then he has a problem encompassing certain things that mathematicians are actually capable of guaranteeing, namely the passing from a given guaranteeing system to the implied guaranteeing of its Gödel sentence.

6.10 One of the key points of the discussion of Chapter 3 of *Shadows* was to exhibit the importance of this conflict within the context of an entirely computational framework. If we accept that the putative robots described there are entirely computational entities, then any "guaranteeing" system that they come up with must necessarily be computational also. Accepting that the robots must also guarantee their guaranteeing system (see Section 3 above) and that they appreciate Gödel's theorem - and also accepting that random elements play no fundamentally important role in their behaviour (see 3.18, 3.22) - we are driven to the remaining loophole for computationalism: errors. It was the thrust of Sections 3.17-3.21 to demonstrate the implausibility of this loophole also. For this discussion, one attempts to find computationally bounded safeguards against errors, and then shows that this is impossible.

6.11 In effect, though in a stronger form than usual, all this is saying is that it is impossible to "formalize" the informal notion of unassailable mathematical demonstration. In this sense McDermott is indeed right to fail to "see how that's possible". It's *not* possible if "formalize" indeed implies something *computational*. That's the whole point!

7. The "Unknowability" Issue

7.1 Several other commentators (Chalmers, Maudlin, Moravec - and also McDermott again!) prefer to attack the Gödel argument from the standpoint that the "algorithm" (or formal system) to which Gödel's theorem is to be applied is *unknowable* in some sense - or, at least, unknowable to the person attempting to apply the argument. (Indeed, Chalmers, for one, seems to be happy enough to accept "that we have an underlying sound competence, even if our performance sometimes goes astray"; so in his commentary on my "First Argument" - that given in *Shadows*, Section 3.3 - he seems to be resorting to the "unknowability" of the algorithm in question.)

7.2 There is an unfortunate tendency for some people (Chalmers, and some others excepted) to try to twist my use the Gödel argument away from the form in which I actually gave it, which refers to "mathematical understanding" in the abstract sense - or at least in the sense in which that term might apply to the mathematical community as a whole - to a more personal form. Such people seem to regard it as more impressively ridiculous that some individual mathematician could know his or her "personal algorithm", than that the principles underlying the proof procedures that are common to mathematicians as a whole might be accessible to the common understanding of the

mathematical community. And they apparently regard it as particularly evidently ridiculous that I myself should have such access (cf. commentaries by McCullough, Maudlin, and Moravec), so they phrase what they take to be my own Gödelian arguments in the form of what kind of a contradiction I might land myself in if I happened to come across my own personal algorithm! I suppose that in order to make "debating points", such procedures may seem effective, but I find it distinctly unhelpful to phrase the arguments in this way; for the arguments then become significantly changed from the ones that I actually put forward.

7.3 Particularly unhelpful are formulations like Moravec's "Penrose must err to believe this sentence." and McCullough's "This sentence is not an unassailable belief of Roger Penrose." Although there are ways of appreciating the nature of the particular sentence that Gödel originally put forward in terms that are not totally dissimilar from this, it is certainly a travesty to attempt to express the essentials of my own (or indeed Gödel's) argument in this way. Only marginally better would be "No mathematician can believe unassailably that this sentence is true." or "No conscious being can accept the truth of this sentence." - mainly because of their manifest similarity to the archetypal self-contradictory assertion: "This sentence is false." In Section 3.24 of *Shadows*, I explicitly addressed the possibility that the kind of reasoning that I had been using earlier in the book (basically the argument of 3.16, which is that of Section 3 above, but also 3.14) might be intrinsically self-contradictory in this kind of way. I do not think that it is, for reasons that I discussed in 3.24. None of the commentators has chosen to dispute me on this particular issue, so perhaps I may take it that they agree also!

7.4 Instead, the arguments, relevant to the present discussion, that Chalmers, Maudlin, McDermott, and Moravec are really putting forward (and which are greatly obscured by the above kind of formulation), is that the algorithm in question might be *unknowable*. They make the point that in order to provide an effective simulation of the thought processes of an individual mathematician, an almost unimaginably complicated algorithm would have to be envisaged. Of course, this point had not escaped me either(!), which is the main reason why I formulated my own discussion in quite different ways from this.

7.5 There are, in fact, two distinct broad lines of argument put forward in *Shadows*, the *simple* argument and the *complicated* argument. The simple argument (which has always been good enough for me) is basically the "bare" Gödelian reasoning referred to in Section 4 above (leading to the conclusion G of *Shadows*, p.76), as applied to the mathematicians' belief that they are "really doing what they think they are doing", rather than blindly following the rules of some unfathomable algorithm (see the opening discussion of 3.1 and the final one of 3.8). Accordingly, the procedures available to mathematicians *ought* all to be knowable! The only remark concerning any aspect of implications of this line of argument that I can find in these commentaries is that towards the end of McDermott's piece, in which he remarks that the quality of conscious understanding will, in his view, turn out to be something "quite simple" (because "consciousness is no big deal"). I remarked (*Shadows*, p.150) that "understanding has the appearance of being a simple and common-sense quality", but if it actually is something simple, it has to be something non-computational, because otherwise it becomes subject to the bare form of the Gödelian argument. I do not think that McDermott would be very happy with that, but he does not refer to this particular problem. (As an aside, I find it hard to see why some commentators, such as Maudlin, seem to argue that the slightest flaw in the discussion of Part 1 of *Shadows* would demolish the whole argument. In fact there are several different lines of argument presented there. All of them would have to be demolished independently!)

7.6 The complicated lines of argument are addressed more at those who take the view that mathematicians are not "really doing what they think they are doing", but are acting according to

some unconscious unfathomable algorithm. Since there is no way that we could know what this algorithm is (or what several distinct but effectively equivalent algorithms might be; cf. *Shadows*, Section 3.7), I adopt a completely different line of approach. This is to examine how such an unfathomable algorithm might conceivably come about. The issue of the role of natural selection (treated particularly in 3.8) was referred to in Section 5 above. The other possibility that I discussed was some form of deliberate AI construction, and that was the thrust of Sections 3.9 onwards, in Chapter 3 of *Shadows*.

7.7 Rather than trying to "know" whatever putative algorithm might now describe the physical action of the brain of some individual human mathematician - or else what complicated computer program might now control the actions of some putative intelligent mathematics-performing individual robot - I consider the general type of computational AI process that might underlie the evolution of such a robot. We do not need to know how the robot's computer-brain is actually supposed now to be wired up, since I am prepared to accept that the "bottom-up" procedures that are used (artificial neural networks, genetic algorithms, random inputs, even natural selection processes that might be applied to the robots themselves, etc. - and also adequately simulated environments, cf. *Shadows* Section 3.10, McCarthy, McCullough, and McDermott take note) could lead to a final product of almost unimaginable complication. Nevertheless, these very *mechanisms* that go into the ultimate construction of the robots would indeed be knowable - in fact, it might well be claimed (as I know that Moravec (1988) has actually claimed) that these mechanisms are, in effect, known already. The whole point of considering these mechanisms, rather than the "actual" algorithm that is supposed to be enacted by the computer-brain of our putative robot (a point apparently missed by Maudlin, McDermott, and Moravec), is that the former would be supposed to be knowable, so long as those aspects of the AI programme that are aimed at the construction of an actually intelligent robot - intelligent enough for it to be able to understand mathematics - are attainable within the general framework of present-day computer-driven ideas.

8. AI and MJC

8.1 A summary of this line of reasoning formed part of the "fantasy dialogue" given in Section 3.23 of *Shadows*. (In what follows, "MJC" refers to the robot, whereas "AI" refers to the subject of artificial intelligence.) Thus, when Maudlin ridicules the possibility that MJC might "easily 'digest' its own algorithm", he has missed the point. There is not supposed to be anything "unknowable" about the procedures of AI; otherwise there would be no point in people actually trying to *do* AI!

8.2 One of the aims of the discussion in the dialogue was to bring home the fact that, according to the "optimistic" school of AI, to which Moravec belongs, it need not be so far into the future when robots are actually constructed which could exceed all human mental capabilities. In particular, such a robot could perform feats of mathematical understanding that exceed those of any human mathematician. This indeed seems to be a corollary of such an optimistic stance with regard to AI, and is not particularly (as McDermott contends) "extravagant" from the point of view of the tenets of AI. My characterization of MJC was to set its abilities, with regard to mathematical understanding, just ahead of that of humanity, but with a particularly effective ability with regard to directly computational matters. Thus, it would have no difficulty at all in assimilating the purely computational aspects of the mechanisms concerning its original construction (since these were in any case already known to Albert Imperator, but they might be computationally very involved), whilst MJC might be relatively slower in appreciating the subtleties of certain logical points - although still a good deal faster than one might imagine a human mathematician would be.

8.3 This does not seem to me to be an "incongruity" in characterization, as Moravec seems to suggest. Of course MJC goes mad at the end - but why not? It has just been driven to the logical

conclusion that the only way in which it could have come about was by God implanting a Divine Algorithm into its mechanisms, through the "chance" elements that were part of those mechanisms. It is not a question of MJC suddenly realizing that its initials stand for "Messiah Jesus Christ", as McDermott seems to think. (The initials were just intended as a little joke for the reader, and not really part of the story.) In fact, McDermott seems extraordinarily slow in getting to the point of the story, if indeed he ever really gets to the point of it. (Actually, it seems that he does not, especially in view of his comments about "affixing a * to Omega(Q)", etc. He has not appreciated the central argument repeated in Section 3 above. It is clear that Maudlin misses the point also, since the dialogue has nothing to do with "a computer failing to pass the Turing test". But so also does Moravec's robot, so McDermott and Maudlin are in good company!) I certainly do *not* believe that a computationally controlled robot could achieve the kind of easy-flowing intelligent-sounding dialogue that MJC exhibits. That is the whole point of a reductio ad absurdum. One assumes that all the implications of the premise, that one intends ultimately to disprove, actually hold good. The final contradiction disproves the premise. Here, the premise is that the procedures of computational AI can ultimately lead to the construction of an intelligent mathematics-performing robot. Of course such a putative robot could be articulate and sound intelligent in other ways than just in mathematics. But it doesn't mean that I believe the premise.

8.4 On another point, the fantasy dialogue does not actually summarize *all* the arguments of Chapter 3 of *Shadows*. Most particularly, it does not summarize most of the arguments given in Sections 3.17-3.21 against the "errors" argument (cf. Section 6 above). I did not include these mainly because I felt that the discussion was already getting rather long and complicated; and since the "error" discussion was rather involved, I thought it best to leave most of it out. In addition to this, the way that the dialogue developed, it seemed appropriate for MJC to have a distinctly arrogant character. It would have changed the flavour of the story to allow MJC to acquire the humility that would have been needed in order to have it admit to being subject to serious error.

8.5 In some ways this was perhaps unfortunate, because it appears to give Maudlin and McDermott an easy way out by allowing their allegedly more "realistic" version of MJC to make the occasional mistake. This, however, would be to miss the point of the "errors" arguments, as given in *Shadows* and as referred to in Section 6 above.

9. Mathematical Platonism

9.1 I think that a few remarks in relation to my attitude to mathematical Platonism are appropriate at this stage. Indeed, certain aspects of my discussion of errors, as given in Section 6 above, might seem to some to be inappropriately "Platonistic", as they refer to idealized mathematical arguments as though they have some kind of existence independently of the thoughts of any particular mathematician. However, it is difficult to see how to discuss abstract concepts in any other way. Mathematical proofs are concerned with abstract ideas - ideas which can be conveyed from one person to another, and which are not specific to any one individual. All that I require is that it should make sense to speak of such "ideas" as real things (though not in themselves material things), independent of any particular concrete realization that some individual might happen to find convenient for them. This need not presuppose any very strong commitment to a "Platonistic" type of philosophy.

9.2 Moreover, in the particular Gödelian arguments that are needed for Part 1 of *Shadows*, there is no need to consider as "unassailable", any mathematical proposition other than a P-sentence (or perhaps the negation of such a sentence). Even in the very weakest form of Platonism, the truth or falsity of P-sentences is an absolute matter. I should be surprised if even Moravec's robot could make much of a case for alternative attitudes with regard to P-sentences (though it is true that some

strong intuitionists have troubles with unproved P-sentences). There is no problem of the type that Feferman is referring to, when he brings up the matter of whether, for example, Paul Cohen is or is not a Platonist. The issues that might raise doubts in the minds of people like Cohen - or Gödel, or Feferman, or myself, for that matter - have to do with questions as to the absolute nature of the truth of mathematical assertions which refer to large infinite sets. Such sets may be nebulously defined or have some other questionable aspect in relation to them. It is not very important to any of the arguments that are given in *Shadows* whether very large infinite sets of this nature actually exist or whether they do not or whether or not it is a conventional matter whether they exist or not. Feferman seems to be suggesting that the type of Platonism that I claimed for Cohen (or Gödel) would require that for *no* such set could its existence be a conventional matter. I am certainly not claiming that - at least my own form of Platonism does not demand that I need necessarily go to such extremes. (Incidentally, I was speaking to someone recently, who knows Cohen, and he told me that he would certainly describe him as a Platonist. I am not sure where that, in itself, would leave us; but it is my direct personal impression that the considerable majority of working mathematicians are at least "weak" Platonists - which is quite enough. I should also refer Feferman to the informal survey of mathematicians reported on by Davis and Hersch in their book The Mathematical Experience, 1982, which confirms this impression.)

9.3 The issue as to the "existence" of some very large set might occasionally have a bearing on the truth or otherwise of certain P-sentences. Accordingly, a mathematician's belief with regard to such a P-sentence might be influenced by that mathematician's particular attitude to the existence of such a set. Questions of this nature were discussed in *Shadows*, Section 2.10, response to Q11, where it is concluded that there is no great issue to disturb significantly the Gödelian conclusion G. Feferman has not chosen to comment on this matter, so I suppose that he has no strong objection to my line ofreasoning.

10. What has Gödel's Theorem to do with Physics?

10.1 Maudlin questions the very basis of my contention that one can indeed deduce something important and new about the nature of physical laws from the actual behaviour of certain physical objects: human mathematicians.<6> As far as I can make out, his basic claim is that the computability, or otherwise, of mathematicians has no externally observable consequences. I find this claim to be a very strange one. He refers to what he calls the "Strong Argument", which he says is "clearly unsound". The Strong Argument contends that "no computer could reliably produce the visible outward motions of a conscious person" and, consequently, there must be something beyond computation in the behaviour of physical objects (e.g. humans). Maudlin's objection seems to rest on the *finiteness* of the total output of a human being. Whatever the total output of some human being might be (and his "human being" is "Penrose", of course!), that output would indeed be finite. Therefore there would be some computer program which could, in principle at least, simulate that person's action.<7> This is a very odd line of reasoning, because it would invalidate any form of deduction about physical theory from observation whatsoever. The number of data points concerning observations of the solar system is finite, after all, so those data points could form the output of a sufficiently large computer, quite independently of any underlying physical theory. (Or they could be used to support a wrong theory with enough parameter freedom, such as the Ptolemaic theory, or even chariots in the sky.) I am tempted to reply to Maudlin by merely saying "be reasonable!"

10.2 Of course, canned answers could in principle provide any answer you want - even with infinite numbers of alternatives if the canning is allowed to be infinite. But the *whole point* of a Turing test (as Turing himself importantly understood) is that it takes the form of a question and answer session. It is simply not *practicable* to take into account all conceivable questions and follow-up

questions and follow-up follow-up questions, etc. simply by storing all possible alternatives. (Anyone who has contemplated the task of writing a comprehensive CD-ROM program - or even a book such as *Shadows* in which one attempts to "second guess" all readers' possible counter-arguments - will begin to appreciate what I mean. There can be a significant complexity explosion even in the relatively short reasoning chains that are involved in such things.) Maudlin refers to this matter of complexity explosion, but he does not draw the appropriate conclusion from it.

10.3 My contention is that without any *genuine understanding* on the part of the computer, it will (at least in most cases) eventually be found out, when subjected to sensitive enough questioning. Trying to simulate intelligent responses by having mountains and mountains of stored-up information, using the programmer's best attempts to assimilate all possible alternatives, would be hopelessly inefficient. It appears that Maudlin believes that he has made a decisive logical ("in principle") point by bringing in the finiteness argument. But he is allowing his computer to have an exponentially larger finite storage limit than the finite limit that he imposes on the human (which is a general feature of the "canned response" approach), and this is totally unreasonable. Indeed, this "exponential" relationship involved in a canned response (or in what is called a "look-up table") *is* a decisive logical ("in principle") response to Maudlin's proposal. This applies both in the finite and in the (idealized) infinite case; for we have 2^alpha > alpha whether or not alpha is finite, and this inequality comes from the same kind of diagonal argument (Cantor's original argument) as that used in the Gödel theorem.

10.4 In fact the finiteness issue was discussed in *Shadows* (in the responses to Q7 and Q8 in Section 2.6), though from a slightly different angle. Maudlin does refer to this discussion, but he appears to misunderstand it. (Baars, in expressing his somewhat muddled parallel/serial worries about the "infinite memory" of a Turing machine is in effect, also addressing the "finiteness" issue, but he does not refer to my discussion of it, nor to the relevant Section 1.5.) In that discussion, I addressed the problem of how one might provide answers to mathematical questions - of, say, deciding the truth of P-sentences - by simply listing all the correct answers. In my response to Q7, I pointed out that the very process of listing the answers required some means of forming reliable truth judgements. This matter has simply been ignored by Maudlin, yet it contains the whole point of the non-computability argument. In order to be able to list the *correct* answers to the P-sentences in his canned responses, Maudlin's *computer programmer* will need to possess the (non-computable) quality of understanding in order to provide what are actually the correct answers! When I said, in *Shadows* that "the odd against this are absurdly enormous", I was referring to the chances against providing the answers to mathematical problems of this nature without any understanding on the part of the programmer. Maudlin's situation is completely different, where he in effect presupposes that the programmer is allowed to have this understanding, and this completely begs the non-computability question.

10.5 There is, however, a somewhat related issue that has also been raised with me by other people: how could one actually *tell*, by observational means alone, whether or not the physical world behaves non-computably? (Here, I am leaving aside the question of the behaviour of extremely highly sophisticated physical objects like human beings; I am concerned with direct physical experiments and the like.) It seems to me that this issue is quite comparable to a somewhat related one, namely that of *determinism*. How could one tell by direct physical experiment whether or not the physical world is deterministic? Of course, one cannot tell - not just like that. Yet there is the common assertion that the classical behaviour of physical objects is indeed deterministic. What this means is that Newtonian *theory* (or Maxwell's *theory* or Einstein's *theory*) is deterministic; that can be shown mathematically. What one does is to design sophisticated experiments or observations to test the theory in other respects, and if the expectations of the theory are borne out, we conclude that various other things about that theory, such as the fact that it is indeed deterministic, ought also

to hold for the behaviour of the universe (to the appropriate degree of approximation as is implied by the limits within which the theory has been shown to be valid). And so it will be with the new theory of physics that unites the classical and quantum levels and which, I maintain, will turn out to be a non-computable theory. Of course, I am at a disadvantage here, since this theory has yet to be discovered! But the general point is the same.

11. How Could Physics Actually Help?

11.1 Several commentators (Baars, Chalmers, Feferman, Maudlin) question the competence of *any* physical theory ever having anything of importance to say about mind, consciousness, qualia, etc. and Klein asks for clarification on this issue. According to Feferman, for example, my attempts to push the consciousness discussion in the direction of physics would merely be to replace one "nothing but" theory with another, i.e. to replace "the conscious mind is nothing but a computer" with "the conscious mind is nothing but sub-atomic physics". Other commentators, in effect, express similar worries. In fact, to describe things in the aforementioned way is rather to miss the point of what I am trying to say. I certainly do not expect to find any answers in sub-atomic physics, for example. What I am arguing for is a *radical upheaval* in the very basis of physical theory.

11.2 In most respects, this upheaval would have to have no observable effects, however. This might seem odd, but we have an important precedent. Einstein's general relativity, as regards most (indeed, almost all) of its observational consequences, is identical with Newton's theory of gravity. Yet, it indeed provided a radical upheaval in the very basis of physical theory. The concept of gravitational force is gone. the concept of a flat background Euclidean space is gone. The very fabric of space-time is warped, and the density of energy and momentum, in whatever form, directly influences the measure of this warping. The precise way in which the warping occurs describes gravity and tells us how matter is to move under its influence. Self-propagating ripples in this space-time fabric can occur, and carry away energy in a mysterious non-local way. Although for many years observational support for Einstein's theory was rather marginal, it can now be said that, in a clear-cut sense, Einstein's theory is confirmed to a precision of one part in one hundred million million - better than any other physical theory (see *Shadows*, Section 4.5).

11.3 What I am asking for is a revolution of (at least) similar proportions. It should represent as much of a change in our present-day ways of looking at quantum theory as general relativity represents a change from Newtonian theory. Some will argue, however, that even the profound changes that I have described above, which overturn the very basis of Newtonian physics, will do nothing to help us come to terms with the puzzle of mentality within such a physically determined universe. I do not deny the significance of that argument. But we do not yet know the very form that this new theory must take. It might have a character so different from that which we have become accustomed to in physical theory that mentality itself may not seem so remote from its form and structure. Moreover, quite apart from any considerations of mentality, there are, in my opinion, very powerful reasons coming from within physics itself for believing that such a revolution is necessary. (Baars, in particular, fails to appreciate this point when he says "there is yet nothing to revolt against".)

11.4 Einstein's theory was to do with the issue of how to describe the phenomenon of gravity - in its action in guiding the planets and the stars and the galaxies, and in the shaping of the large-scale structure of the universe. These phenomena do not directly relate to the processes which control the behaviour of our brains and which presumably actually underlie our mentality. What I am now asking for is a revolution that would operate at the very scales relevant to mental processes. Yet I am also arguing that the physical revolution we seek should actually be dependent upon the

particular revolutionary changes that Einstein's theory already represented from the older Newtonian ideas about the nature of reality.

11.5 I know that this puzzles many people; in fact, it puzzles many *physicists* that I should seriously attempt to claim such a thing. For the scales at which gravitational interactions reign seem totally different from those which operate in the brain. A few words of explanation may well be helpful at this juncture. I am certainly not asking that gravitational interactions (or "forces") should have any significance for the physical processes that are going on in the brain. The point is quite a different one. I am referring, instead, to the influences that Einstein's viewpoint with regard to gravity will have upon the very structure of quantum theory. Instead of quantum superpositions persisting for all time - as standard quantum theory would have us believe - such superpositions constitute a state which is *unstable* (see Penrose 1996). Moreover, this decay time can be computed, at least in certain very clear-cut situations. Yet, many physicists might well take the view that the time-scales, distance-scales, mass-scales, and energy-scales that would arise in any framework that purports to embody the union of Einstein's general relativity with quantum theory must be hopelessly wrong. Indeed the relevant time-scale (~10^-43 seconds) is some twenty orders of magnitude shorter that the briefest processes that are considered to take place in particle physics; the relevant space-scale (~10^-13 cm) is some twenty orders of magnitude smaller than the diameter of a proton; the relevant mass-scale (~10^-5 grams) is about the mass of a flea, which seems much too big; and the relevant energy scale (~10^18 ergs) is about what would be released in the explosion of a can of petrol. However, when one comes to examine the details, these figures conspire together (some being individually too small but others correspondingly too big) to produce an effect that is indeed of an eminently appropriate magnitude. (For details, see a forthcoming paper by Stuart Hameroff and myself (Hameroff and Penrose 1996).)

11.6 Again, many would argue that we shall still have come no closer to an understanding of mentality in physical terms. Perhaps, indeed, we will not have come a great deal closer. But I believe that some progress will have been made in an appropriate direction. The picture of quantum state reduction that this viewpoint is concerned with ("OR": objective state-reduction) involves the bifurcation and then selection of one out of several choices for the very shape of space-time. Moreover, there are fundamental issues arising here as to the nature of time and the apparent flow of time (see Section 13 below, in relation to Klein's commentary). I am not arguing that these issues will, in themselves, resolve the puzzles of human mentality. But I do claim that they could well point us in new directions of relevance to them, and this could change the very nature of the questions that the problems of mentality raise.

11.7 I think that people in AI, and perhaps a good many philosophers also, have a tendency to underestimate the importance of the *specific* nature of the physical laws that actually govern the behaviour of our universe. What reason do we really have to assume that mentality does not need these particular laws? Could consciousness arise in a world controlled by some arbitrarily chosen set of rules? Could it arise within scope of John Conway's "game of life" (Gardner 1970, Poundstone 1985), for example, as Moravec (1988) has suggested? Although the Conway rules for a "toy universe" are ingenious, they do not have the subtle sophistication of Newtonian mechanics - whose sophistication people often take for granted. Yet despite the extraordinary fruitfulness of Newtonian ideas, even they cannot explain something so basic as the nature and stability of atoms. We need quantum theory for that. And even quantum theory does not fully account for the behaviour of atoms, because its explanations require that curious hybrid of procedures of unitary (Schroedinger) evolution and quantum state-vector reduction (denoted in *Shadows* by U and R, respectively) - procedures which are not really consistent with one another, I claim. Eventually, in order to explain even the stability and the specific nature of atoms, we shall need a better theory of physics than we have today, at the *fundamental* level.

11.8 There is no doubt that physics - and often the very detailed nature of the specific underlying physical laws - is essential to most of the sophisticated behaviour of the world we know. So why should the most sophisticated behaviour that we know of in the world, namely that of conscious living human beings, not also depend on the very detailed nature of those laws? As I have indicated above, we do not yet know the full nature of these laws, even in some of their most basic respects. A new theory is needed quite independently of any necessity for new laws to describe a universe that can support consciousness. However, physicists themselves often get carried away into thinking that they know everything that is needed - in principle, at least - for the behaviour of all things of relevance. There is a curious irony, here, in McDermott's quoting from *Shadows* p.373 "It is only the arrogance of the present age that leads so many to believe that we now know all the basic principles that can underlie all the subtleties of biological action." For he takes that remark to be aimed primarily at the AI community. In fact, the people I had primarily in mind were the (theoretical) *physicists*. I do not blame the biologists, or even AI researchers, when they take from the physicists a picture of the world commonly claimed to be almost final - bar some technical details that are irrelevant for the behaviour of macroscopic objects. But perhaps McDermott is right; some AI researchers seem to be nearly as arrogant as high-energy physicists (and with far less reason) - especially those AI researchers who claim that the deepest mystery of the physical world can be answered without any reference to the actual laws that govern that world!

11.9 I should make it clear, however, that I am certainly making no claim that the mystery of mentality can be resolved *merely* by finding the correct physical theory. I am sure that there are vital insights to be gained from psychology as well as from neuro-physiology and other aspects of biology. Baars seems to think that I am denying the existence of the *unconscious*, because there is no significant mention of it in *Shadows* (though there was some small reference to the unconscious mind in *The Emperor's New Mind*). I should like to reassure Baars that I fully accept both the existence of the unconscious and its importance to human behaviour. The only reason that the unconscious was not discussed in *Shadows* was that I had no contribution to make on the subject. I was concerned with the issue of *consciousness* directly, in particular in relation to the quality of understanding. However, I certainly agree that a complete picture cannot be obtained without the proper role of unconscious mentality being appreciated also.

12. State-Vector Reduction

12.1 Some commentators express worries in connection with my quantum state-vector proposals - whereby the quantum procedure R is to be replaced by some form of *objective* reduction, which I denote by OR. There are many misunderstandings here. Baars seems to think that I am taking the view that R has something to do with "observer paradoxes", which is explicitly not my view, as I thought I had made clear in *Shadows*, Chapter 6. Klein does not make this mistake, but seems to take the view that the measurement problem (R) has (or ought to have) something directly to do with metaphysics. This is certainly different from my own "objective" standpoint with regard to R.

12.2 Maudlin complains that my "objections to Bohm's theory" (a theory that, in a sense, incorporates R) "are impossible to decipher from the text" - which is not surprising since I did not give them there - and that my "objections to the GRW theory are clearly not decisive". My objections to GRW (the OR scheme of Ghirardi, Rimini, and Weber, 1986) were not meant to be decisive. In my opinion, this scheme is a very interesting one, but it suffers from being somewhat ad hoc. What one needs (and I am sure that the authors of this scheme would not disagree) is some way of fitting the scheme in more convincingly with the rest of physics. In fact Diosi made a proposal in 1989 that could be regarded as a GRW-type model in which the ad hoc nature of the GRW parameters was removed by fixing them to be provided by the *quantum gravity* quantities

referred to in Section 11, above. Diosi's model encountered difficulties, as was pointed out by Ghirardi, Grassi, and Rimini (1990), who also suggested a remedy, but at the expense of re-introducing another parameter. It should be said that in fact the Diosi-Ghirardi-Grassi-Rimini proposal is extremely close to the OR scheme that I was proposing in *Shadows* (and in Penrose 1995). These other authors do not mention non-computability (their proposal being entirely stochastic), but there is no essential incompatibility between both sets of ideas. In Sections 7.8 and 7.10 of *Shadows*, I give some reasons (admittedly far from conclusive) for anticipating that a full quantum gravity scheme of this nature might indeed be non-computable.

12.3 In his final paragraph, Maudlin seems to be complaining that the tentative OR proposals that are being promoted in *Shadows* do not solve all the problems of uniting quantum theory with relativity, and of explaining the problems of human cognition. He's not asking for much! These proposals were hardly intended to provide a complete theory (as anyone reading Section 7.12 of *Shadows* would surely appreciate) but merely to give some idea of the orders of magnitude involved in the collapse rate for such an OR theory - if such a theory *were* to be found.

12.4 Klein refers to the excellent little book QED of Feynman (1985) which introduces the basic rules of quantum theory (and quantum fieldtheory) with the minimum of fuss. However, Feynman never attempts to address the measurement problem in this book - which amounts to the issue of why (and when) do the quantum-level complex-valued amplitudes become classical-level real-valued probabilities, in the process of having there moduli squared. It might be worth mentioning that I read QED for ideas, just before embarking on writing my chapter on quantum mechanics in *The Emperor's New Mind*. However, I found that Feynman's approach was not altogether suitable for me because I needed to address the measurement problem in some detail, which Feynman avoided completely. Feynman has certainly worried about this problem, but he preferred not to emphasize it in his writings. There is a historical point of interest, here. For it was actually Feynman's early worrying about the nature of the union of Einstein's general relativity with quantum mechanics (expressed in Feynman's contribution to the conference held in Chapel Hill in the 1950s, cf. Feynman et al 1995, Section 1.4, p. 15) that originally motivated Karolyhazy (1966) to seek an explanation of state-vector reduction in terms of gravitational effects (and Feynman also influenced me in the same way). Diosi's particular approach arose from the work of Karolyhazy's Budapest school.

12.5 Questions to do with "the overlap of states" referred to by Klein do not really resolve the measurement issue, and von Neumann's point about the difficulty of locating exactly where (or when) R takes place just emphasizes the subtlety of the R phenomenon. However Klein is completely right in pointing to the biological difficulties involved in maintaining quantum coherence within microtubules and, more seriously, in allowing this coherence to "leap the synaptic barrier". To see how this might be achieved is a fundamental problem for the type of scheme that I (in conjunction with Stuart Hameroff) have been proposing. Clearly more understanding is needed. (See Section 14 below, for a tentative suggestion in relation to this.)

12.6 I should point out a misunderstanding on the part of Maudlin. He seems to think that my "collapse theory offers a stochastic collapse postulate" and that it is concerned with "the exact timing of the collapses". I have nowhere said either of these things. I get the impression that Maudlin has been confused by the comparisons that I have been making between the suggested OR process and the phenomenon of the decay of an unstable particle (or unstable nucleus). But at no stage did I suggest that it would be in the precise *timing* of the decay of the quantum superposition that significant non-computability would occur. (But on rereading the relevant parts of *Shadows*, I realize that I was not at all specific there as to what I *did* mean.) Of course, since the detailed theory is not known, it is possible that there would be relevant non-computability in the timing also, but

what I had in mind was something quite different, and certainly more relevant. The idea is that when collections of tubulin conformational movements become involved in coherent quantum superposition, there will come a point when the mass movements within the tubulin molecules are sufficient for OR to come into effect (without significant environmental disturbance). When that happens Nature must make a choice between the various collections of conformational states under superposition. It is not so much a question of *when*, but of *which* among the collection of superposed states Nature indeed chooses. Choices of this kind could actually influence the behaviour of a synapse. (There are various possibilities for this; for example, the particular collections of conformational states of tubulins in a microtubule might influence a dendritic spine, via the actin within the spine. Moreover, a great number of microtubules would be expected to act in concert - since a single OR state-selection process would act within many microtubules all at once. However, there is no point in trying to be too specific at this stage.) It would be in the *particular* choice that Nature makes that the non-computability could enter significantly, and this particular choice (global over a significantly large area of the brain, probably involving at least thousands of neurons) could result in subtle collective changes of synapse strengths all at once. (See Hameroff and Penrose 1996.)

13. Free Will

13.1 What kind of a theory might it be that determines these choices? Many people who are unhappy with computationalism would be just as unhappy with any other type of mathematical scheme for determining them. For they might argue that it is here that "free will" makes its entry, and they would be unhappy that their free-will choices could be determined by *any* kind of mathematics. My own view would be to wait and see what kind of non-computable scheme ultimately emerges. Perhaps a sophisticated enough mathematical scheme will turn out not to be so incompatible with our (feelings of) free will. However, McCarthy takes the view that I am "quite confused" about free will, and that my ideas are "not repairable". I am not really clear about which of my confused ideas McCarthy is referring to. In *Shadows*, I did not say much about the issue of free will, except to raise certain issues. Indeed, I am not at all sure what my views on the subject actually are. Perhaps that means that I *am* confused, but I do not see that these ideas are remotely well enough defined to be irreparable!

13.2 As I remarked above, most people would probably take the view that if there is *any* kind of mathematical theory precisely determining the way we behave, then there is no free will. But, as I have indicated, I am not so sure about this. The answer could depend on the very nature of this mathematical theory. The theory would certainly have to be non-computable (according to my own considerations), but much more than this. We recall from the discussion given in Section 3 above (McCullough, Chalmers, and Section 7.9 of *Shadows*) that the Gödel diagonalization procedure can be applied to systems much more general than merely computational ones. Thus, my arguments would equally imply that our missing theory must be not just non-computational, but also beyond (or at least different from) Turing's notion of oracle computation. (An oracle computation is what is achieved by a Turing machine to which an additional command is appended: "if the pth Turing machine acting on the number q eventually halts, do A; if it doesn't, do B".) Again, we can consider second-order oracle machines (which can assess whether first-order oracle machines ever halt), and the diagonalization still applies. So the missing theory is not a second-order oracle theory either. The same applies to even higher-order oracles. Indeed the missing theory cannot be an oracle theory of order alpha for any computable ordinal number alpha. As far as I can make out, this is not the limit of it either. Diagonalization can be applied in very general circumstances indeed. We enter very nebulous areas of mathematical logic. It seems that the quality of "understanding" - which is what this discussion is effectively all about - is something very mysterious. Consequently, any theory of the physical world which is capable of accommodating beings that are capable of genuine

understanding must itself be in a position to cope with such subtleties.

13.3 As a side comment, I should remark that this form of "repeated Gödelization" is somewhat related to, but not at all the same as, that referred to by McCarthy and McCullough, who both describe the process whereby sound extensions of a sound formal system can be obtained, corresponding to any computable ordinal alpha. This procedure was described in *Shadows* Section 2.10, answer to query Q19. I am not quite sure why they go to the trouble to repeat this argument, with no reference to my own discussion. The conclusion, noted in *Shadows*, that "repeated Gödelization does not provide us with a mechanical procedure for establishing the truth of P-sentences" is confirmed by Feferman. (As far as I can make out, Feferman's comments about how his work extended that of Turing's are related to the considerations of the previous paragraph, above.)

13.4 The issue of free will is related also to the experiments of Libet (1992), (and to earlier experiments of Deeke, Groetzinger, and Kornhuber (1976) and also Grey Walter) that are referred to by Klein. These experiments suggest a delay of the order of a second, in an entirely volitional act, between the first indications of mental activity (as evidenced by brainwave studies) and the final willed (say) finger movement. Klein calls into question my puzzlements, expressed in *Shadows* Section 7.11, concerning the seeming slowness of consciousness - but as far as I can make out, he has misunderstood my point (as, I believe, did Ian Glynn before him, in the 1990 article that Klein refers to). Klein says that there are "no surprises" in the fact that there is "substantial unconscious processing" (in fact, of the order of a second's worth) before "the subjective awareness of the decision to act" takes place, and that only about one fifth of a second's delay occurs before the motor act. But it was *precisely* the length of time involved in the unconscious processing that was worrying me. (Or at least this was *one* of the two things that was worrying me; the other had to do with the *passive* half-second delay that Klein also regards as an over-estimate - to which I shall return below.)

13.5 If consciousness has any active role to play in a response to an external stimulus, then it is no good for the unconscious to have already lined up the action a second ahead of time unless the unconscious already "knew" what decision that the conscious mind was going to take. Klein asserts that I am referring to a "stimulus-response situation", for which the response could be much more rapid and essentially entirely *unconscious*. In fact I am *not* considering situations of this kind, but those in which it is necessary for consciousness to come into play in order that it can actively influence the outcome. If the "free will" of the conscious mind is allowed to come into play, then surely it would be necessary for *all* processes that need to be involved, whether they be the conscious ones or the preparatory unconscious activities, to take place *after* the external stimulus occurs. This leads us to a about a second's delay for a consciously influenced response.

13.6 The extra half second comes from Libet's other (passive) experiment (Libet et al 1979), which Klein argues may be too long for stimuli that are significantly greater than threshold. I do not wish to argue this point, since I am not aware of the relevant figures (the 100 msec figure referred to by Klein being not relevant to the situation I was considering, as far as I can see). A one-second's delay for a consciously controlled free-willed response seems already an inordinately long time.

13.7 I am not claiming that these considerations are decisive, in any way, as indications that the quantum/relativity puzzles concerning the nature of time and causality have significance for our consciousness and perception of "the flow of time". However, it seems to me that it is quite possible that there is something very odd going on concerning the timing of conscious events, if only for the reasons indicated in *Shadows* Section 7.11 that the role of time with respect to consciousness is quite different from its role in physics - in that it is only with the phenomenon of consciousness that

time seems to "flow". I certainly hope that more experiments of the types that Libet and his associates have been performing will be carried out in the future. I suspect that there may be further surprises in store.

14. Some Remarks on Biology

14.1 Many people have expressed reservations (of widely differing degrees) concerning the biological speculations put forward in *Shadows*. I have referred (in Section 12 above) to Klein's worries about the difficulty of maintaining quantum coherence within individual microtubules and, moreover, of this quantum coherence straddling, simultaneously, a great number of microtubules within collections of separated neurons. I agree with Klein that it would be an extraordinary challenge to see how such organization might be achieved. Yet, I maintain that somehow Nature must indeed have accomplished this extremely remarkable task. In this section, I shall try to address this issue further, and also address some of the other objections that have come my way. I shall also relate several new things, in relation to these issues, that I have learned since I wrote *Shadows*.

14.2 One complaint that I have heard is that the biological purpose of microtubules within cells is "already known", namely that they are there to provide the "tracks" along which molecules (generally known as "organelles") are transported from one part of a cell to another; and they grow, shrink, or bend in ways that are designed to influence the movement of cells. Moreover, so these arguments might continue, their tubelike construction is to give them structural strength - so there is no need to ask for a separate purpose for the tubes, such as to isolate some kind of quantum-coherent activity taking place within the tubes, from the outside environment. I do not doubt that microtubules indeed perform the tasks that they are presently believed to perform, and many more besides. But that is no argument against their *also* serving the additional purposes that I require of them. We know of many instances where Nature uses the same structure for many different purposes. We know that mammalian noses, for example, filter substances from the air before it reaches the lungs (not to mention their importance to the sense of smell). Yet this is no argument against elephants *also* using their noses in delicate ways to pick up objects from the ground!

14.3 A more serious argument is the lack of direct evidence for the type of "cellular automaton" activity that Hameroff and his colleagues have been arguing for in patterns of tubulin conformations along microtubules. The existence of some kind of activity of this general nature is indeed part of the general picture that Hameroff and I would require for our model of the physical processes underlying consciousness. To obtain direct experimental support for this kind of activity would be a key issue, and I certainly hope that it will be possible to design experiments to test it. Experimental support for the existence of some kind of quantum coherence within microtubules is a matter of even greater importance for the ideas that I have been promoting in *Shadows*. There is no doubt that definitive experiments would be difficult to perform, especially since there is a distinct possibility that the relevant effects might require microtubules *in vivo* rather than *in vitro*. I have been informed by Guenther Albrecht-Buehler that there is some kind of coating (analogous to the myelin sheaths of neurons) that microtubules have *in vivo* which tends not to be present *in vitro*.

14.4 On the theoretical side, some progress has been made. Work by Tuszynski et al (1996) gives theoretical support for information processing (of the Hameroff type) to be possible, within an appropriate temperature range, provided that the microtubule possesses the structure of what is known as the "A-lattice", which is indeed the structure depicted in *Shadows* figures 7.4, 7.8, 7.9, on pp. 359, 363. However, the work of Mandelkow and Mandelkow (1994) indicates that many (perhaps most) microtubules seem to have a somewhat different structure, known as the "B-lattice", in which there is a "seam" running the length of the microtubule. Tuszynski et al argue that the B-lattice is not capable of sustaining Hameroff-type information processing, but it may well be

appropriate for the transporting of organelles. It would be extremely interesting to have information about which kind of lattice structure is prevalent in axons, dendrites, non-neuronal cells, etc.

14.5 With regard to the theoretical possibility of quantum coherence within microtubules, the model of Jibu et al (1994) seems well-founded, in which super-radiance effects are anticipated within microtubules (analogous to the activity of a laser), where the electromagnetic field interacts with ordered water. For this process to occur, it would be necessary for the water within the tubes actually to adopt this ordered structure, and to be appropriately free of the wrong kind of impurities, such as chloride ions. (Apparently, sodium, calcium, magnesium, and potassium ions, in low enough concentrations should not disturb the ordering.) It should be mentioned, however, that the type of coherent activity that is anticipated in the model of Jibu et al may not be sufficient for my purposes. Though it is a necessarily quantum effect, it is not, as it stands, a *quantum-coherent* effect of the type that my arguments require. Genuine quantum coherence seems to be necessary in order that the quantum/classical borderline can be probed, where the (non-computable) effect of the missing OR theory can significantly make its mark. (This comment has relevance to Klein's query at the end of his Section 1. *Classical* coherence in the brain may well occur, but it does not provide an opening for non-computational activity, which I argue is a characteristic feature of consciousness.) The Jibu et al mechanism may be part (though not all) of what is needed.

14.6 An interesting possibility has come my way, which may conceivably have relevance to the question of how quantum coherence might get conveyed between one neuron and another (a question raised by Klein). As noted in *Shadows*, Figs. 7.11, 7.12 on pp. 365, 366, there are some particular molecules (clathrins) that inhabit synaptic boutons, which have the highly symmetrical structure of a *truncated icosahedron* (like a modern soccer ball). These clathrin molecules have importance in the release of neurotransmitter chemicals at synapses (whereby the nerve signals are transmitted from neuron to neuron). Although I do not have specific suggestions to make here, I am struck by the extraordinary symmetry of these molecules. It has been brought to my attention (by Roy Douglas, cf. Douglas and Rutherford 1995) that, according to the Jahn-Teller effect, such highly symmetrical molecules would have a large energy gap between the lowest quantum energy level and the next. This lowest level would be highly degenerate, and there would be interesting quantum-mechanical effects when this degeneracy is broken.

14.7 Energy gaps and symmetry breaking, of this general nature, are central to the understanding of superconductivity - and superconductivity is one of the few clear phenomena in which large-scale quantum coherence takes place. Known observationally since 1911, and explained quantum-mechanically in 1957, superconductivity had been thought originally to be an exclusively very low-temperature phenomenon, occurring only at a few degrees above absolute zero. It is now known to occur at much higher temperatures of -158 degrees Celsius, or perhaps even -23 degrees (although this is not properly explained). It does not seem to be out of the question that there might be similar effects at the somewhat higher temperatures of microtubules. Perhaps there are understandings to be obtained about the behaviour of microtubules from the experimental insights gained from such high-temperature superconductors.

14.8 Another question frequently asked is: what's so special about neuronal microtubules, as opposed to those, say, in liver cells? In other words, why isn't your liver conscious? In answer to this, it should be said that the organization of microtubules in neurons is quite different from that in other cells. In most cells, microtubules are organized radially, from a central region (close to the nucleus) called the *centrosome*. In neurons, this is not the case, and they lie essentially parallel with one another along the axons and dendrites. The total mass of microtubules within neurons seems to be much greater than in other cells, and they are mainly stable structures, rather than in most cells, where they continually polymerize and depolymerize (grow and shrink). Of course, there is much to

be learned about the respective roles of microtubules in neurons and in other cells, but there does seem to be clear enough evidence for an essentially distinct role for (some of) those in neurons. (The A-lattice/B-lattice question would seem to be of importance here also.)

14.9 In this connection, I should mention something of considerable interest and relevance that I learned recently from Guenther Albrecht-Buehler (1981, 1991), which concerns the role of the *centriole*, that curious "T" structure (roughly illustrated in *Shadows*, Fig. 7.5, on p.360), consisting of two cylinders resembling rolled-up venetian blinds, constructed from microtubules and other connectingsubstances, which lies within the centrosome. In *Shadows*, I had adopted the common view that the centrosome acts in some way as the "control centre" of the cytoskeleton of an ordinary cell (not a neuron), and that it initiates cell division. However Albrecht-Buehler's idea about the role of the centriole is very different. He argues, convincingly, in my opinion, that the centriole is the *eye* of the cell, and that it is sensitive to infra-red light with very good directional capabilities. (Two angular coordinates are needed for identifying the direction of a source. Each of the two cylinders provides one angular coordinate.) Impressive videos of fibroblast cells provide a convincing demonstration of the ability of these cells to pinpoint the direction of an infra-red light source. This also provides some remarkable evidence for individual cells having considerable information-processing abilities, which is at variance with current dogma. One may well ask where the "brain" of a single cell might be located. Perhaps its structure of microtubules can serve such a purpose, but it does seem that the centrosome itself must have some central organizing role. In a single (non-neuronal) cell, the microtubules emanate from the centrosome. I gather from Albrecht-Buehler that the specific contents of the centrosome are not known. It seems that it would be important to know what indeed is going on in the centrosome. Does it have some information-processing capabilities? Is there conceivably some structure there that is capable of sustaining quantum coherence in any form? The answers to questions of this nature could have considerable importance.

14.10 I should make clear that I am not arguing for any consciousness (or consciousness of any significant degree) to be present for individual cells. But according to the views that I have been putting forward, some of the ingredients that are needed for actual consciousness ought already to be present at the cellular level. Individual cells can behave in strikingly sophisticated ways, and I find it very hard to see how their behaviour can be explained along entirely conventional (classical) lines.

14.11 All this notwithstanding, there is the question of whether microtubules are indeed necessary for consciousness to be present in human beings or other animals. An argument that I have heard presented - as though it were a conclusive refutation of this contention (cf. Grush and Churchland 1995, Edelman 1995) - is that the drug colchicine, which is given as a treatment for gout, depolymerizes microtubules, yet it does not influence the mental state; moreover, when colchicine is delivered directly to the brains of experimental animals, they appear to remain conscious. This can be answered (cf. Penrose and Hameroff 1995 and references contained therein) by pointing out that: (1) the blood-brain barrier is not significantly breached in the gout patients, so their neuronal microtubules are not disturbed, and (2) in any case, most brain microtubules, unlike those in non-neuronal cells, are stable structures that do not undergo cycles of polymerization and depolymerization, and are *resistant* to colchicine. A minority of neuronal microtubules - those involved in restructuring synaptic connections - are involved in such activity, however, and these could be affected by colchicine. Indeed, the experimental animals referred to above suffer a kind of "dementia", similar to Alzheimer's disease (Bensimon and Chernat 1991), a disease which has itself been linked to microtubule disruption (e.g. Matsuyama and Jarvik 1992).

14.12 Of course, there is the additional issue of how we could know whether a demented rat is or is not conscious. We must return to the question of what consciousness is and what are its external

manifestations.

15. What is Consciousness?

15.1 In *Shadows* (cf. Section 1.12) I concentrated specifically on the quality of "understanding" as a particular manifestation of consciousness - a quality that would make its characteristic mark on external behaviour as well as being an internal manifestation of mentality. Only with respect to the quality of understanding have I been able to argue for non-computable ingredients being necessary. But, in my view, non-computable physical processes must also be essential for other aspects of conscious mentality.

15.2 Consciousness has its active aspects - basically the "free-will" issues that were considered in Section 13 above - and it has its passive aspects, which have to do with awareness and the vexed issue of *qualia*. Understanding fits somewhere between the two. In my view, anything that sheds some light on the problem of how a physical system can exhibit understanding must inevitably also shed some light on the "free-will" and "qualia" problems. Moreover, the issue of "understanding" seems to me to be one of the more tangible aspects of consciousness. I do not see how to say much that is scientifically useful about the qualities of "free will" or "awareness", but "understanding" is something that we can work with. Klein raises the issue of Wilczek's challenge of "looking for perceptual feats that humans can do more efficiently than robots". My answer would be: anything in which the quality of *understanding* is important. A good example is the chess problem presented in *Shadows* Fig. 1.7 on p.46. This problem was one of a series composed by William Hartston and David Norwood, consisting of chess problems, some of which were designed to be easy for humans but hard for computers (best solved using "understanding") and others, the other way about (best solved by "trying all possibilities"). This "Turing test" showed a virtually complete separation between humans and computers.

15.3 As must be clear from the preceding remarks, I do not believe that any real progress will be achieved towards solving the mysteries of how mental phenomena fit in with the physical universe until there are some important changes in our picture of physical reality. Perhaps these developments will lead to a theory in which "consciousness" finds some place within the purely physical descriptions of the world. One is reminded of such ideas as "panpsychism" (like those of Leibniz, Spinoza, or Whitehead), where consciousness may play its part within the processes of physical action at its deepest levels. I do not have strong opinions as to the significance of such ideas, mainly because I have not studied them in detail. But I suspect the truth to have a much more compelling grandeur to it than any set of ideas that I have seen so far.

15.4 Thus, I certainly do not go along with the "no big deal" viewpoint of McDermott, a viewpoint which is, in effect shared by Baars when he takes the view that the solutions are all to be found within psycho-biology, without any significant change in our physical world-view being necessary. I think that Baars grossly underestimates the force of the arguments from logic. (And although, as a mathematician, I frequently make use of "variables", I have no idea what Baars means by "treating consciousness as a variable".) I fully accept that there are invaluable insights to be gained from the areas of psychology and biology. But, important though these areas are, I was not so much concerned with them as with physics in *Shadows*. For I believe that it is *also* fundamentally important to see whether our present physical world-view is in fact adequate for accommodating, in any way, the phenomenon of consciousness. I have extensively put the case here, and in *Shadows*, that it is not. The arguments from logic and from physics are not counter to those from psychology and biology, but complementary to them.

15.5 Likewise Moravec and McCarthy appear to belong to the "no big deal" school. McCarthy puts

forward various suggestions for the circumstances under which he would consider that "consciousness" occurs. These are all within the computational model, so it is clear from this that I am not in agreement with him that his computer systems, acting according to his criteria, are *actually* conscious (in the sense that one could actually *be* such a system). Again, I fear that McCarthy does not appreciate the force of the logical arguments that I have given, which inform us that the quality of "understanding" cannot be accommodated within the computational model. It is easy to suggest *definitions* within the computational model (as McCarthy does) of such things as "consciousness", "awareness", "self-awareness", "intentions", "beliefs", "understanding", and "free will". But such definitions need not convey to us (and do not convey to me) any conviction that the corresponding mental qualities that humans actually possess are in any real sense captured by computational definitions of this nature. As I have argued extensively above, the actual quality of human *understanding* cannot be captured within *any* purely computational scheme. So it is clear that I cannot be in agreement with all of McCarthy's definitions.

15.6 Chalmers raises the issue of the distinction between "simulating", "evoking", and "explaining". I agree with him that there are indeed distinctions to be made. But I feel that these are distinctions that have more importance to a philosopher than to a scientist. Whilst I did myself distinguish between the possibility of "simulating" and of "evoking" consciousness in my "viewpoint B" and "viewpoint A" distinction in *Shadows* Section 1.3, I think that the normal scientific (as opposed to philosophical) stance would be to concentrate on what can be externally observed, so a system which succeeds in simulating the outward effects of consciousness would be suspected also of evoking them. Thus, my "viewpoint B" might not be a happy one for a hard-nosed scientist.

15.7 Likewise, a modern scientist might have trouble producing an "explanation" for something other than by producing a theory that in principle provides a (mathematical) "simulation" of that thing. It seems to me, therefore, that the "scientific" viewpoints (in this sense) are those who hold to the same position, with regard to computationalism, in respect of all three of the simulate/evoke/explain issues. And since Chalmers's "N" represents a denial of the competence of physics with regard to conscious mentality, this implies that we would be left with just "CCC" (computationalism all down the line) or "PPP" (non-computational physics all down the line). Thus, when I wear my scientist's hat, I am unable to understand why someone (such as Chalmers) can hold to different positions with regard to the simulate/evoke/explain issues, although when I wear my philosopher's hat, I can partly appreciate his point. However, I wear my scientist's hat much more frequently than my philosopher's hat!

15.8 But sometimes I try to wear both hats at once. The arguments in *Shadows* were concerned almost entirely with the "simulate" issue with regard to human mentality. I hope that those who study, with a genuinely open mind, the arguments given there (and the further discussions above) will come to accept that a non-computational physics will be needed in order even to simulate the actions of a conscious being. There are, in any case, powerful reasons for believing that profound changes in our physical world-view are in the offing. For the resulting science to be non-computable to the degree that seems to be required, we may well find the need for a science that is so different from the science of today that the *evoke* and *explain* issues with regard to mentality may finally find natural explanations.

Acknowledgement

I am grateful to the National Science Foundation for support under contract PHY 93-96246.

Notes

<1> As I understand it, there was to have been a tenth commentary, explicitly addressing a number of biological points, but unfortunately this article did not materialize. Nevertheless, in Section 14, I shall find it helpful to address a number of biological criticisms that have come to my attention.

<2> Most book-reviewers seem to have missed this argument too. In particular, Hilary Putnam, in his widely quoted review of *Shadows* in the Sunday New York Times Book Section (Putnam (1994) and reprinted, for some reason, in the Bulletin of the American Mathematical Society, Putnam (1995)) not only completely missed this argument, but tries to claim that I have not considered other issues that I have, in fact, discussed in great detail. The matters are thoroughly discussed again in Sections 6 and 7 of this reply.

<3> Dennett, in his 1995 book *Darwin's Dangerous Idea*, seems to be trying to make out that I do not believe that human abilities can have arisen by the process of natural selection, since I do not believe in computationalism. This is a very strange contention. Provided that the non-computational ingredients are present in nature, there is nothing against natural selection having made use of them - and it is my own contention that this is indeed how things were. If Dennett is arguing that I do not believe that natural selection provides the sole explanation of the origin of human mentality, then he is right. In particular, some specific laws of physics are also needed, within whose scope natural selection must operate. But that, in itself, is hardly a radical position, nor an unscientific one!

<4> It appears that some people, on reading the section entitled "Contact with Plato's world" in Chapter 10 of *The Emperor's New Mind*, have picked up the curious view that I believe that mathematicians obtain their mathematical knowledge by use of some direct mystical quality not possessed by ordinary mortals (see Grush and Churchland 1995, for example), and even that I may be claiming for myself a particularly unique such quality! This is a complete misreading of what I had intended in that section; for I was simply trying to find some explanation of the fact that different mathematicians can communicate a mathematical truth from one to another even though their modes of thinking may be totally dissimilar. I was arguing merely that the mathematical truths that each mathematicians may be groping for are "external" to each of them - these truths being "inhabitants of Plato's timeless world". I was certainly not arguing for a fundamentally *particular* quality of "direct Platonic contact" to be possessed only by certain individuals. I was referring simply to the general qualities of "understanding" (or "insight") which are in principle available to all thinking individuals (though they may perhaps come somewhat more *easily* to come individuals than to others). These qualities are not mystical - but as Gödel's theorem shows, there is indeed something rather mysterious about them.

<5> See, for example, Grush and Churchland (1995), and for a reply, Penrose and Hameroff (1995).

<6> John Searle, in his interesting recent review of *Shadows* in the *New York Review of Books* (November 2, 1995), seems to be making a somewhat similar point. However, he does not appear to have grasped, properly, the key notion of *non-computability* (and the fact that it has observational manifestations). See, in particular, the discussions given here (Section 7, 10, 11, 15) and also Sections 3.2 onward of *Shadows*, which explicitly address the conscious/unconscious issue that Searle raises, with regard to an algorothmic basis for mathematical understanding. The aforementioned sections may serve to clarify my own position concerning what he claims are my "fallacies".

<7> F.J.Tipler, in a review of *The Emperor's New Mind*, used a similar "finiteness" argument, specifically referring to the (absurdly large) Bekenstein bound on the information that can be stored within an object of a given (say human) size. I explicitly addressed this argument from finiteness in my responses to Q7 and Q8 in Section 2.6 of *Shadows*. However, I had deliberately desisted from

indicating the personal individuals whom I had in mind (in this case Tipler) as putting forward the various specific "queries" that I was responding to. There is a certain irony for me in this, because in his review of *Shadows* in *Physics World*, Tipler chastised me, claiming that I did "not even mention ... the Bekenstein bound" (Tipler 1994), not having even noticed this section of the book that had been specifically aimed at his own arguments!

References

Albrecht-Buehler, G. (1981). Does the geometric design of centrioles imply their function? *Cell Motility, 1,* 237-245.

Albrecht-Buehler, G. (1991). Surface extensions of 3T3 cells towards distant infrared light sources. *Journal of Cell Biology, 114,* 493-502.

Bensimon, G. and Chernat, R. (1991). Microtubule disruption and cognitive defects: effects of colchicine on learning behavior in rats. *Pharmacology and Biochemistry of Behavior, 38,* 141-5.

Davis, P.J. and Hersch, R. (1982). *The Mathematical Experience*. Harvester Press.

Deeke, L., Groetzinger, B., and Kornhuber. H.H. (1976). Voluntary finger movements in man: Cerebral potentials and theory. *Biology and Cybernetics, 23,* 99.

Dennett, D.C. (1995). *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Devlin, K. (1988). *Mathematics: The New Golden Age*. London: Penguin Books.

Diosi, L. (1989). Models for universal reduction of macroscopic quantum fluctuations. *Physical Review, A40,* 1165-74.

Douglas, R.R. and Rutherford, A.R. (1995). Pseudorotations in molecules I: Electronic triplets. To appear.

Edelman, G. (1995). Quoted on p. 323 of *Frontiers of Complexity*, by Peter Coveney and Roger Highfield. London: Faber and Faber.

Feynman, R.P. (1985). *QED: The Strange Theory of Light and Matter*. Princeton: Princeton University Press.

Feynman, R.P., Morinogo, F.B., and Wagner, W.G. (1995). *Feynman Lectures on Gravitation*. Reading, MA: Addison-Wesley.

Gardner, M. (1970). Mathematical games: the fantastic combinations of John Conway's new solitaire game 'Life'. *Scientific American, 223,* 120-123.

Ghirardi, G.C., Rimini, A., & Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review, D34,* 470.

Ghirardi, G.C., Grassi, R., & Rimini, A. (1990). Continuous-spontaneous-reduction model involving gravity. *Physical Review, A42,* 1057-64.

Grush, R. and Churchland, P.S. (1995). Gaps in Penrose's toilings. *Journal of Consciousness*

Studies, 2, 10-29.

Hameroff, S.R. (1987). *Ultimate Computing: Biomolecular Consciousness and Nano-Technology*. Amsterdam: North Holland.

Hameroff, S.R. and Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules - a model for consciousness. In S Hameroff, A. Kaszniak and A. Scott (Eds.) *Toward a Science of Consciousness*. Cambridge, MA: MIT Press.

Hameroff, S.R. and Watt, R.C. (1982). Information processing in microtubules. *Journal of Theoretical Biology, 98,* 549-61.

Jibu, M., Hagan, S., Hameroff, S.R., Pribram, K.H., Yasue, K. (1994). Quantum optical coherence in cytoskeletal microtubules: implications for brain function. *BioSystems, 32,* 195-209.

Karolyhazy, F. (1966). *Nuovo Cim., A42,* 390.

Karolyhazy, F. (1974). Gravitation and quantum mechanics of macroscopic bodies. *Magyar Fizikai Polyoirat, 12,* 24.

Libet, B. (1992). The neural time-factor in perception, volition and free will. *Review de Metaphysique et de Morale, 2,* 255-72.

Libet, B., Wright, E.W. Jr., Feinstein, B. and Pearl, D.K. (1979). Subjective referral of the timing for a conscious sensory experience. *Brain, 102,* 193-224.

Mandelkow, E-M. and Mandelkow, E. (1994) Microtubule structure. *Current Opinions Structural Biology, 4,* 171-179.

Matsuyama, S.S. and Jarvik. L.F. (1992). Hypothesis: Microtubules, a key to Alzheimer's disease. *Proceedings of the National Academy of Science USA, 86,* 8152-6.

McDermott, D. (1990). Computation and consciousness. *Behavioral and Brain Sciences, 13(4),* 676.

Moravec, H. (1988) *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind.* Oxford: Oxford University Press.

Penrose, R. (in press). On gravity's role in quantum state reduction. *General Relativity and Gravitation*.

Penrose, R. and Hameroff, S. (1995). What 'gaps'? - reply to Grush and Churchland. *Journal of Consciousness Studies, 2,* 99-112.

Poundstone, W. (1985). *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge.* Oxford: Oxford University Press.

Putnam, H. (1994). The best of all possible brains? *New York Times Book Review, Nov. 20,* 7-8.

Putnam, H. (1995). Review of *Shadows of the Mind,* by Roger Penrose. *Bulletin of the American Mathematical Society, 32,* 370-373.

Tipler, F.J. (1994) Can a computer think? Part II. *Physics World, December,* 51-52

Tusznyski, J., Trpisova, B., Sept, D., and Sataric, M.V. (1996). Microtubular self-organization and information processing capabilities. In S. Hameroff, A. Kaszniak, and A. Scott (Eds.) *Toward a Science of Consciousness*. Cambridge, MA: MIT Press.