Between The Motion And The Act...
A Review of *Shadows of the Mind* by Roger Penrose

Tim Maudlin
Department of Philosophy
Rutgers University
New Brunswick, NJ 08903
USA

maudlin@zodiac.rutgers.edu

PSYCHE, 2(2), April 1995
http://psyche.cs.monash.edu.au/v2/psyche-2-02-maudlin.html

KEYWORDS: artificial intelligence, computation, Gödel's theorem, Penrose, physics, quantum mechanics, relativity, Turing test.

REVIEW OF: Roger Penrose (1994) *Shadows of the Mind*. New York: Oxford University Press. 457 pp. Price: $US 25 hbk. ISBN 0-19-853978- 9.

1. Introduction

1.1 In these comments I want to leave aside entirely whether human mathematical understanding is achieved solely through the manipulation of linguistic symbols by syntactically specifiable rules, i.e. whether it is solely a matter of humans performing a computation. I also want to leave aside the problems that arise in interpreting quantum theory, in particular the measurement problem. Those problems stand on their own quite independent of Gödel's theorem. Rather, I want to focus explicitly on how Gödel's theorem, together with facts about human mathematical understanding, could conceivably have any bearing on physics, that is, on how the first part of *Shadows of the Mind* is related to the second. I want chiefly to argue the reflections arising from Gödel's theorem and human cognitive capacities do not, and could not, have any bearing on physics.

1.2 That there might be any connection at all would be surprising for the following reason. Ultimately, the empirical data of physics resolve themselves into claims about the positions of material bodies. Any physical theory that correctly predicts or accounts for the positions of bodies -- including the positions of needles on complicated scientific instruments, the positions of ink particles on computer printouts, and the positions of dots on photographic plates -- cannot be objected to on empirical grounds. One might object on aesthetic or other grounds (e.g. one might object in principle to a theory that postulates

unmediated action at a distance) but this would not be an empirical failure of the theory. So if Professor Penrose's argument somehow shows that classical physics or quantum physics cannot be complete and correct accounts of physical reality, then Gödel's theorem must somehow have implications about how material bodies can move.

1.3 The overall strategy for connecting Gödel's result to physics would have to be to show that some actual motion of bodies cannot in principle be accommodated within a physical theory of a certain kind. Just as analysis can show that the physical behavior of planets whose orbits precess cannot be accounted for by Newtonian gravitational theory, so Penrose seems to claim that all of classical and quantum physics (as well as a large class of possible extensions or emendations of those theories) cannot account for the physical motions of some known physical bodies: those of human mathematicians. How, in detail, could this connection between a mathematical theorem and physical action possibly be made?

2. The Strong Argument

2.1 In several places, Penrose seems to want to supply an argument that would quite directly connect Gödel's theorem to the motion of physical bodies. Consider, for example, the claim on p. 14 that according to his view, view C, no "fully effective simulation of a conscious person could ever be achieved merely by a computer-controlled robot. Thus, according to C, the robot's actual lack of consciousness ought ultimately to reveal itself, after a sufficiently long interrogation". Since a Turing test of the sort that Penrose endorses as a criterion of consciousness can be carried out via teletype, this amounts to the claim that there is a particular set of physical motions, viz. the motions involved in depressing keys on a keyboard to type out "responses" to given input questions, which Penrose himself (or any competent conscious mathematician) could perform but which could not, in principle, reliably be performed by any computer. Let us call this argument, to the effect that no computer could reliably produce the visible outward motions of a conscious person, the Strong Argument.

2.2 The Strong Argument has the logical form: such-and-such visible outward motions can reliably be performed by a person with human mathematical understanding but, as can be shown by appeal to Gödel's theorem, no computer can reliably produce such visible output given that input, ergo humans are not computers. Further, (and this is the important point for our purposes) the physics that gives rise to human behavior cannot even be simulated on a computer, else the simulation of a mathematician's brain run on a computer could give rise to the motions. Ergo that physics itself cannot be computable.

2.3 The Strong Argument is clearly valid, but just as clearly unsound. For whatever Gödel's theorem shows, it cannot possibly show that no computer can reliably mimic Penrose's own behavior in a Turing test. Under the mild assumption that Penrose cannot understand or respond to a sentence in English that takes (say) 100 millennia to pronounce, the number of questions (and follow-up questions) that can be asked of him in

a Turing test and to which he could intelligibly respond is strictly finite. So a computer could, in principle, be programmed to give completely "canned" responses to every possible set of Turing questions, which responses would match Penrose's own answers. There is no question of the computer understanding anything, or of simulating the underlying physics of Penrose's brain. The point is that the computer is certainly capable or producing exactly the same Turing test behavior that Penrose's brain, as a physical object, can. So the Strong Argument, directed at the outward behavior of mathematicians, cannot possibly be correct.

2.4 There is rather compelling reason to think that Penrose means to make the Strong Argument. Beside the passage just cited, consider the following (my underline added): "I shall shortly be giving some very strong reasons for believing that effects of (certain kinds of) understanding cannot be simulated in any kinds of computational terms" (p. 48); "Anyone who maintains that all the external manifestations of conscious thought can be properly computationally simulated... must find some way of coming to terms, in full detail, with the arguments that I shall give" (p. 49); "However, in the above discussion, it is not really necessary that the robot actually possess genuine mental qualities, provided that it is assumed possible for the robot to behave externally just as a human mathematician could...Thus, it is not necessary that the robot actually understand, perceive, or believe anything, provided that in its external pronouncements it behaves precisely as though it does possess these mental attributes" (p. 158); "The above arguments would seem to provide a powerful case against the computational model of the mind -- viewpoint A -- and equally against the possibility of an effective (but mindless) computational simulation of all the external manifestations of the activities of mind -- viewpoint B." (p. 202). Even more tellingly, the fictional dialogue in 3.23 features a computer which fails to act externally like a rational human. The dialogue which is supposed to illustrate the main point of Part I presents a computer failing to pass a Turing test. So it is plausible to assume that Penrose takes his arguments to show that no computer could pass a well designed Turing test.

2.5 The great advantage of the Strong Argument is that it actually is an argument, and it has a conclusion which bears on physics. So the Strong Argument, if sound, bridges the gap between Part I and Part II of the book. The even greater disadvantage, as mentioned above, is that it is clearly unsound. To repeat: Penrose himself could certainly pass a well designed Turing test, and Penrose himself is capable of comprehending and meaningfully responding only to questions of a finite fixed length, so the very same responses can be programmed into a computer (we may take the whole Turing test inquiry up to a given point as the complete question being asked at that point). Such a program with canned responses will certainly be practically impossible (due to combinatorial explosion of the possible questions) but is just as clearly possible in principle. It could carry on a conversation with the interrogator not only longer than the computer in the fantasy dialogue, but for, say, 1000 (or 100 million) years. Surely Penrose is not claiming that he could do better than that.

2.6 If it is Penrose's intent to establish the Strong Argument then we know that something has gone amiss, and the rest is just post mortem.

3. Backing Off

3.1 Given the ease with which the Strong Argument is defeated, it seems charitable to seek some other intent in the text, despite the passages cited above. And indeed, one can find indications that something less sweeping than the Strong Argument was aimed for. Let us consider some of the possibilities.

3.2 The clearest indication that, despite all appearances to the contrary, Penrose does not mean to establish the impossibility of a computer passing a Turing test occurs in the response to Q7 on p. 82. There, an objector raises a similar point about canned responses, noting that a computer could be programmed to produce the same output of mathematical theorems as all of humankind to date and well into the future. (This is not the same kind of canned response I sketched above, since this is apparently just a matter of reciting theorems rather than producing the requisite banter.) Penrose's answer to this question is doubly puzzling. First, he asserts that the question "ignores the central issue, which is how we (or computers) know which mathematical statements are true and which are false" (p. 83). The questioner should be absolved of blame for missing this central issue, given the passages cited above. Penrose goes on to assert that "The arguments that I am trying to make here do not say that an effective simulation of the output of human conscious activity (here mathematics) is impossible[!] , since purely by chance[?] the computer might 'happen' to get it right -- even without any understanding whatsoever. But the odds against this are absurdly enormous...". The first part of this sentence is quite a shock, and the second part fairly hard to understand. The odds against the "canned" computer getting the output right are nil, even though the computer has no understanding of anything. But this passage does show that, at least at some points, Penrose explicitly denies trying to show that computers can't pass Turing tests, and so implicitly denies using the Strong Argument to make the connection from Gödel to physics.

3.3 This leaves us with two problems. First, if the intent is not to show that computers can't produce some particular external behavior, what is it exactly that computers cannot do? Second, given that it is of central importance to the Strong Argument that the issue be one of external behavior, how is the connection to physics to be made?

3.4 One weakening of the claim that computers can't pass a Turing test is found in the statement of position B on the question of computers and consciousness. Position B holds that "Awareness is a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness" (p. 12) Note that position B concerns not the computational simulation of external behavior but the computational simulation of the physical action of the brain.

This passage suggests that Penrose means to establish only that the internal physics of the brain cannot be computationally simulated, not that a computer couldn't pass a Turing test or couldn't simulate the external behavior of a conscious human. Let us call this conclusion the Weak Conclusion. The Weak Conclusion does, of course, imply Penrose's claims about the inadequacy of contemporary physics, but it does not imply that a computer would eventually be "unmasked" by a clever interrogator.

3.5 The Weak Conclusion follows from the conclusion of the Strong Argument: if no computer can reliably produce the same output as a human brain, then no computer can simulate the physical action that gives rise to that output. The converse, or course, does not hold -- there is nothing absurd in the idea of a brain whose internal physical action cannot be computationally simulated but whose output can. Imagine, for example, the brain of a super-mathematician who is able to check whether every natural number is the sum of four squares by running through the entire set of natural numbers, checking each. No Turing machine can perform that feat, but it can arrive at the same answer ("yes") by other means. So Penrose needs an argument to the Weak Conclusion that does not give the Strong Conclusion (i.e. that no computer can reliably pass a Turing test), since the latter is indefensible. But there simply seems to be no way offered to the Weak Conclusion that does not go via the Strong one. External behavior is the only place where the motion of bodies, and hence physics, comes into play.

3.6 If not via the Strong Argument, how exactly do considerations of physics get into the game at all? I believe it is by conflating the claim that human brains don't understand mathematics by virtue of doing computations with the claim that they don't do so by virtue of computable physical action. These claims should be kept clearly distinct. The planets, for example, don't perform any computations at all, they do not manipulate symbols. In particular, they do not orbit the Sun in virtue of performing computations. They do, however, orbit the Sun in virtue of computable physical action. From the fact that we cannot understand the activity of the planets by ascribing computational structure to them, it does not follow that their activity is not the result of (and understandable in terms of) computable physics. Similarly, if the appeal to Gödel's theorem works, it shows at best that reflecting on mathematics is not a matter of just manipulating symbols by means of valid syntactic rules. But what could that observation prove about the underlying physics of the brain?

4. Simulating Brains

4.1 Let's try to be as concrete as possible about the situation. Suppose that I am interested in Penrose's brain as a purely physical object. I am not concerned with whether, much less how, he thinks about mathematics, or indeed whether he thinks at all. I am simply concerned, in the first place, to describe his brain as a collection of particles assembled in a particular configuration. I do not describe him, for the purposes of my physics, as using any sort of an algorithm or formal procedure. Nor is there any reason that I can think of for believing that, on the basis of the physical description, one could derive an algorithm

that he is using to solve mathematical questions. I just have a collection of particles in a given disposition.

4.2 Thought of in this way, Penrose's brain is unimaginably complex. Modeling its behavior using quantum theory would be unspeakably complicated, though theoretically possible. It is also theoretically possible to use a digital computer to simulate the physical action of that brain according to the laws of quantum mechanics (together with, say, GRW collapses<1>). Such a simulation would produce simulated output behaviors given simulated input (and simulated boundary condition to, e.g., keep simulated nutrients coming to the simulated neurons). And if we simulate stimulating his auditory nerves with a mathematical query, the computer would eventually produce simulated output to the voice box, which we could algorithmically translate as the answer to our query. What reason is there to believe that the simulated output would not be qualitatively indistinguishable from Penrose's actual behavior? And given that we have nowhere suggested that Penrose is using an algorithm to arrive at the answer, where could Gödel's theorem begin to get a grip on this question?

4.3 The only program around to which to apply Gödel's theorem is the program that simulates the action of Schroedinger's equation (and the GRW collapses) on the quantum state that describes Penrose's brain. But that program isn't even the kind of program that Gödel's theorem is concerned with -- it doesn't prove theorems or check whether a Turing machine ever stops! So how could that program be relevant to anything?

4.4 We can, however, append to our program another which would result in a "theorem proving" computer. First, write a program that translates English sentences into the sort of auditory stimulation that one would receive if the sentence were spoken. Then write a program that constantly checks the output to the vocal cords for the words "Ah, I am unassailably convinced that the Turing machine you just asked about will not halt", or words to that effect. If such words appear, the computer prints "does not halt" and shuts down. So now we can input questions about particular Turing machines, run the simulation of the physics of Penrose's brain, and wait to see if we get a simulated response. And we now have a purely mechanical device that will offer opinions about Turing machines. Let us call the algorithm this machine uses P.

4.5 Note first that this device will not, like the Mathematically Justified Cybersystem of the fantasy dialogue, claim to have some sort of superhuman mathematical ability. If it works as I claim it will, it will boast of no more mathematical ability than Penrose himself would. Indeed, perhaps it will refuse to say it is unassailably convinced of anything. In any case, it would be much easier for Penrose to write a fantasy dialogue with this robot -- he need only answer the questions as he himself would.

4.6 So how do we apply Gödel's theorem to this brain-simulating algorithm? The

conclusion that Penrose draws from Gödel's theorem is that human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth. And this conclusion is certainly correct: since the soundness of an algorithm (or at least its consistency) is a mathematical fact, mathematicians who only believe the theorems proved by an algorithm will only believe that the algorithm is sound if it proves itself to be sound. But Gödel showed that any formal system that can prove its own consistency isn't consistent, and hence not sound, and a fortiori not knowably sound. But how can we apply this conclusion in the situation sketched? Even if we have the physics right, Penrose himself is not using P to determine anything, that is, Penrose is not getting answers by imagining or reasoning about or employing an algorithm that simulates his brain activity. Indeed, right now Penrose has no idea at all of what P is. So it is only in a Pickwickian sense that one could say that the success of P in simulating Penrose's brain implies that he is using any algorithm at all. And if the success of P does not imply that Penrose is using an algorithm, then the success of P cannot possibly conflict with the conclusion Penrose draws from Gödel's argument.

4.7 But let's grant, for the sake of argument, this Pickwickian sense of "using an algorithm". If P is the algorithm that Penrose is "using", still, is it at all plausible that P is "knowably sound", and in particular, is it knowably sound by Penrose?

4.8 Again, think concretely about the situation. Penrose is not presented with some relatively short method or program, but with a quantum description of every single particle in his brain, together with a mechanical method of deriving time evolutions of that state, and a translation mechanism for input and output. He could obviously not hold the whole program in his head, since it has more information in it than he has neurons (or cytoskeletons!). He could not even read the program in his lifetime. He could not possibly determine whether the mathematical opinions offered by this machine will even be consistent. The only way that Penrose could conclude that this program constitutes a sound algorithm is by accepting that it is an accurate description of his brain, accepting that the physics is accurately depicted, and inferring that since his own methods are sound, so is this program. But that is not being "knowably sound" in the sense that Gödel's theorem requires, since it is not a matter of establishing the soundness by any mathematical or formal considerations. This would rather be an empirical argument, and fall entirely outside the bounds of Gödel's concerns. There is, for example, no sense in asking of such an empirical method if it is sound, much less knowably sound.

4.9 When Penrose discusses our resources for establishing the soundness of an algorithm (section 3.3), he addresses only the non-Pickwickian sense of "using an algorithm". That is, he discusses algorithms for manipulating symbols, whose axioms are translated as valid formulas and whose inference rules are recognizably sound (p. 133). But once we start thinking about directly modeling the physical action of the brain, rather than reducing the psychological processes of the thinker to manipulation of formal symbols, these resources for establishing the soundness of the process are lost. Our brain simulating algorithm doesn't have mathematical axioms: it has a description of the initial

physical state of the brain. And the "rules of procedure" of the algorithm are not inference rules defined over sentences, they are rules for evolving that physical state forward in time. As soon as we switch from the idea of an algorithm that manipulates mathematical symbols to one that manipulates representations of physical states, it becomes inescapable that the soundness of the algorithm (in terms of the sentences it eventually produces) is necessarily beyond the grasp of the person whose brain is being modeled.

5. The Final Escape Hatch

5.1 Even if we accept that Gödel's theorem proves that Penrose is not using a knowably sound algorithm to decide mathematical questions, that at best only implies that the unimaginably complex computer simulation P cannot be known, by inspection, to be sound. And indeed, Penrose could certainly not determine whether the proffered program was sound or not. (Compare this with the way the computer in the fantasy dialogue easily "digests" its own algorithm (p. 181) -- it could not similarly digest a complete description of its physical state!) But perhaps we are being blinded by merely accidental and contingent limitations on Penrose's insight. Perhaps Penrose could not "see" the soundness of the algorithm in practice, but he could nonetheless do so in principle.

5.2 The tricky qualifier "in principle" does appear at several junctures in the text. On p. 65: "...no such system of rules can ever be sufficient to prove even those propositions of arithmetic whose truth is accessible, in principle, to human intuition and insight..."; on p. 101: "For there certainly does appear to be a well-defined sense in which what is accessible in principle to one mathematician is the same....as what is accessible to another -- or, indeed, to any other thinking person"; and, in another context, on p. 48: "I shall maintain that a computer system's actual lack of general understanding should -- in principle, at least -- eventually reveal itself". Let us take the last of these first.

5.3 If we follow out the strategy of the fantasy dialogue to "unmask" the computer simulating Penrose's brain action, then, just as the Mathematically Justified Cybersystem was fed its own algorithm, so will we feed the computer its algorithm. But if the computer is doing its job well, it will mimic Penrose's own response to this input -- namely by expiring (and simulating a corpse) long before the input could even be read. It is therefore unclear what "in principle" means here. If it means that the questioning should be allowed to go on forever, with questions of unbounded complexity, with every question being answered, then the demand is completely unjustified. Penrose couldn't pass such a test -- so why should a computer simulating his brain action do better? Further, this whole line of argument only makes sense in the context of the Strong Argument, which we have long ago rejected. So this last "in principle" is of no help.

5.4 The other two "in principles" cited above look more promising. If the algorithm simulating Penrose's brain action is sound, and if he can become unassailably convinced it is sound, then there is a Gödel sentence for it which he can be unassailably convinced is true. But to become unassailably convinced that the algorithm is sound, he must analyze the algorithm and prove by uncontroversial mathematical methods that it is sound. And it certainly does seem impossible for Penrose, as he is, to ever prove the soundness of that

algorithm mathematically. But perhaps it is not impossible for someone to prove the soundness of the algorithm. And if Penrose can, in principle, know what anyone else can know, then he can, in principle, know the soundness of the algorithm.

5.5 If this argument has the air of a conjuring trick, that is because it is one. Penrose cannot prove the soundness of P, in part, because it has more lines of code than he has particles in his brain. If someone else (with a stupendously larger brain) could, somehow, inspect the algorithm and prove its soundness, it still doesn't follow that Penrose could. Perhaps Penrose could if his brain were bigger, but this leads to two problems. First, we don't know by what principle we are to enlarge his brain, as a physical object. Where do we add more neurons or cytoskeletons, and in what pattern? But worse than this, if we do enlarge his brain (as a physical object) then the original algorithm is no longer relevant -- it is not a simulation of the physics of his brain. There will be a new algorithm simulating the physical action of the new brain, an algorithm whose soundness will be beyond the new brain to prove. No progress has been made.

5.6 It is just here that the fundamental confusion in the argument of the book again rears its head. If we are concerned with the idea that mathematicians are using algorithms to come to mathematical conclusions, then several inferences are quite reasonable. One is that the algorithms are not terribly diverse or complex: after all, human abilities to follow out complex rules are limited. Another is that we can meaningfully discuss what such an algorithm could output in principle, i.e. if run on a Turing machine with infinite memory for an infinite time. And a third is that the discussion will be little altered if we discuss a community of mathematicians: the total number of algorithms being used will probably not much increase, and if we insist on allowing only algorithms that all of the members of the community endorse, then the number may well decrease. The idea of writing down the relevant algorithm and inspecting it does not seem absurd.

5.7 But if we are instead concerned with modeling the physical action of the mathematicians using computable dynamical equations, all of these inferences become invalid. The "algorithm" will be incomprehensibly complex. The only clear sense of what one person could do in principle is given by letting the program run on -- ending in their simulated death. And most importantly, the modeling of a community of mathematicians is necessarily orders of magnitude more complex than modeling just one. For the physics of a dozen brains is at least a dozen times more complex than the physics of one. So if we bring in some comrades to help Penrose out in proving the soundness of his (personal) algorithm, we change the problem. Modeling the physical behavior of Penrose's brain when he is conversing with his colleagues will require modeling the physics of their brains, and so engender a more complex algorithm. Similarly if he draws on the aid of computers, or even the lowly pencil and paper. All of the objects that play a physical role in the process that leads him to a conclusion must be modeled in the algorithm.

5.8 This fundamental fracture in *Shadows of the Mind*, the fracture that separates Part I from Part II and cannot be mended, is papered over by the single word "computational". Consider the following passage:

Of course, none of this will stop us from wanting to know what it is that is really going on in consciousness and intelligence. I want to know too. Basically, the arguments of this book are making the point that what is not going on is solely a great deal of computational activity -- as is commonly believed these days -- and what is going on will have no chance of being properly understood until we have a much more profound appreciation of the very nature of time, space, and the laws that govern them. (p. 395)

5.9 The conclusion of Part I is that mathematical understanding is not just a matter of using knowably sound algorithms. In that sense, there is more than a great deal of computational activity in the brain. But it simply does not follow that the physical action of the brain is not governed by dynamics that can be simulated on a computer. Certainly all possible computational activity -- all following of algorithms -- can be achieved in systems governed by computable physics. But it is simply affirming the consequent to conclude that all action in systems governed by computable physics is computational activity. This fallacy is masked by use of the same term, "computational activity", to denote doing a computation and doing something that can be simulated on a computer. Disambiguate the two meanings and the halves of the book fall neatly apart.

6. Collapse

6.1 Having argued that the conclusions of Part I cannot possibly have a bearing on the questions raised in Part II, I would like to end by simply registering my views of those two parts taken separately. As mentioned above, the conclusion G on p. 76 is certainly correct: Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth. I also agree that mathematical understanding, and indeed consciousness in general, is not simply a matter of doing computations or having a certain computational structure or being a Turing machine of a specified sort that is performing a computation. I have argued for this conclusion, on completely independent grounds, elsewhere (Maudlin, 1989).

6.2 On the physics side, there certainly are foundational problems in the quantum theory and seemingly intractable problems reconciling quantum theory with Relativity. With respect to the quantum theory alone, Penrose's objections to the GRW theory are clearly not decisive (once we see that being computable does not count against it), and his objections to Bohm's theory are impossible to decipher from the text. Reconciling any of these theories with Relativity does not look hopeful, but Penrose's own suggestion for a collapse theory does no better in this respect, despite the invocation of relativistic paraphernalia<2>. In particular, Penrose's proposal is couched in terms of a unique universal time function (cf. the "NOW" in figure 6.5 on p. 338), and so seems to presuppose a single preferred notion of simultaneity. Nothing in the proposal resolves the problem discussed on p. 295: different universal time functions will yield different accounts of how the collapses occur, at most one of which can be correct. How could one use the proposal to determine which side of the EPR experiment causes the first collapse, i.e. the collapse that causes the distant particle to go into a spin eigenstate? Different universal time functions will give different regions of space-time whose geometries are to be compared, and hence different predictions for collapse. If the collapses are real, then at

most one such time function is correct, yielding an absolute simultaneity function which cannot be reconciled with the relativistic account to space-time structure.

6.3 It is also notable that Penrose's collapse theory offers a stochastic collapse postulate. This is puzzling given the role that he suspects quantum computation to play in cognitive function. Recall the theorems which show that when a collapse occurs makes no difference (for all practical purposes!) once a quantum system has become sufficiently entangled with its environment. If this were true in the brain, then employing a computable collapse postulate (e.g. that of GRW) rather than Penrose's postulate would make no difference (for all practical purposes) in predicting the evolution of the brain state. Now the whole point of examining the physical structure of the cytoskeletons is to find a place in the brain where entanglement with the environment does not occur, and so where the exact timing of the collapses might make a noticeable difference to the evolution of the brain state. But if the collapses take place randomly, governed by a stochastic law, then the differences in brain state evolution that depend on the exact timing of the collapses will also be governed by a stochastic law: so mathematicians will disagree in their conclusions depending on just when the collapses in their brains occur. So in so far as the conclusions of mathematicians are sensitive to the timing of collapses they will disagree, and in so far as their conclusions do not depend on the exact timing of collapses, we can just as well use the GRW theory as Penrose's. If there is unanimity in the mathematical community, then (if we adopt a stochastic collapse theory) relevant evolution of brain state must be robust even under very different timings of the collapses, and so the exact timing of the collapses must be inconsequential.

6.4 It certainly seems plausible that "a much more profound appreciation of the very nature of time, space, and the laws that govern them" will be needed just to get the motion of electrons right, leave aside explaining consciousness. And perhaps folding gravitational effects into the quantum theory will lead us in the right direction. But the collapse proposal in *Shadows of the Mind* does not seem to resolve the tension between Relativity and quantum theory, nor does it fit very well with Penrose's own project of tying the fundamental laws of physics to the remarkable cognitive capacities of human brains.

Notes

<1> GRW = Ghirardi-Rimini-Weber. See Penrose, pp. 331-4.

<2> For a painfully extensive examination of this problem, see Maudlin (1994).

References

Maudlin, T. (1989) Computation and Consciousness. *Journal of Philosophy, 86,* 407-432.

Maudlin, T. (1994) *Quantum Non-Locality and Relativity.* Oxford: Blackwell.

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