

GAPS IN PENROSE'S TOILINGS

Rick Grush and Patricia Smith Churchland

Philosophy Department, UCSD

Abstract: Using the Gödel Incompleteness Result for leverage, Roger Penrose has argued that the mechanism for consciousness involves quantum gravitational phenomena, acting through microtubules in neurons. We show that this hypothesis is implausible. First, the Gödel Result does *not* imply that human thought is in fact non algorithmic. Second, whether or not non algorithmic quantum gravitational phenomena actually exist, and if they did how that could conceivably implicate microtubules, and if microtubules were involved, how that could conceivably implicate consciousness, is entirely speculative. Third, cytoplasmic ions such as calcium and sodium are almost certainly present in the microtubule pore, barring the quantum mechanical effects Penrose envisages. Finally, physiological evidence indicates that consciousness does not directly depend on microtubule properties in any case, rendering doubtful any theory according to which consciousness is generated in the microtubules.

I. Introduction

Consciousness is almost certainly a property of the physical brain. The major mystery, however, is how neurons achieve effects such as being aware of a toothache or the smell of cinnamon. Neuroscience has not reached the stage where we can satisfactorily answer these questions. Intriguing data and promising research programs do exist,¹ but no one would say we pretty much understand the neurobiological mechanisms of awareness.² Much more work, both experimental and theoretical, needs to be done. What available data do suggest is that awareness and subjectivity are probably *network* effects, involving many millions of neurons in thalamic and cortical structures. But there are other possibilities. Dualism aside, a different possibility is that consciousness emerges from quantum mechanical goings-on in subneuronal structures.

¹Cf. Logothetis 1989; Llinas and Pare 1993; Llinas and Ribary (1993); Crick (1994); Damasio (1994); Damasio & Damasio (forthcoming); P. M. Churchland 1995; Bogen (1995).

²Although Dennett's book title, Consciousness Explained (Dennett 1990), rather misleadingly suggests otherwise.

Quite a lot is known about how single neurons work. The biophysics of the synapse, the neuronal membrane, neuron-to-neuron interactions, enzyme-gene interactions, and organelle behavior (e.g. mitochondria, microtubules), is known in impressive, though not complete, detail.³ Given what is known, "very remote" is the label typically stamped on the possibility that quantum mechanical effects play any explanatorily significant role in neuronal function.⁴ "Very remote" is not equivalent to "certainly not", of course. The possibility that quantum mechanical effects give rise to conscious awareness remains alive, especially in certain quarters of physics and mathematics. An otherworldly, gaze-averting fondness for Platonism in mathematics, twinned to a fascination for the counter-intuitive aspects of quantum physics, can foster the hunch that something really uncanny -- and completely unkenneled -- is going on in the brain. Hoisting its status from "campfire possibility" to "scientific possibility" is problematic, however, given that quantum level effects are generally agreed to be washed out at the neuron level. Roger Penrose, however, has gallantly taken up the challenge in his widely discussed book, "The Emperor's New Mind" (Penrose 1989; henceforth 'EMPEROR') and in its successor, "Shadows of the Mind" (Penrose 1994b; henceforth 'SHADOWS').

The crux of the Penrose idea is that quantum mechanical effects exist at the *subneuronal* level -- the level of cell organelles, in particular, in microtubules, whose pore diameter of about 14 nanometers, as well as other physical properties, make them candidates for the possibility of harnessing quantum level effects. (FIGURE 1) This is the tactic to avoid the aforementioned "washed out" objection. Why expect a quantum level effect *anywhere* in the brain? Because, avers Penrose, certain cognitive processes -- including those responsible for mathematical knowledge -- are nonalgorithmic (in a sense to be discussed below), while all classical level biochemical processes are algorithmic. The central motivation, therefore, underpinning Penrose's whole argument structure is a problem in the *epistemology of mathematics*. It is the problem of how we understand mathematics if understanding does not *consist* in following a rule, but involves understanding the meaning of mathematical concepts (see below). This is a rather more arcane starting point than, say, the awareness of a toothache or the sleeping/dreaming/wakeness cycle, which are pretty robust phenomena with a nontrivial log of well-researched data from psychology and neuroscience.⁵

³Cf. the classic by D. Bray (1992); Z. Hall, Ed. (1992); Sossin, Fisher and Scheller (1989)

⁴"Quantal release" at the synapse means that either a vesicle of transmitter is released or it is not. It means that when there is release, the whole vesicle opens and all its transmitter is dumped into the synaptic cleft. This is in no sense a quantum mechanical effect. Also, by 'explanatorily significant' we mean any effect which is not capturable by a classical biochemical explanation.

⁵There is a very simple -- and fallacious -- argument sometimes tugging at the physicist's intuitions. Its clearest formulation is, to our knowledge, owed to the philosopher of science and mathematics, Itamar

In what follows we present a compact version of Penrose's argument as we understand it from EMPEROR and SHADOWS. Support for the quantum-consciousness connection draws chiefly from three sources: (1) Gödel's Incompleteness Result and the fact that mathematical understanding exists; (2) the properties of microtubules, long protein structures found in *all* cells, including neurons, and (3) heretofore unrecognized physical processes, perhaps exemplified in a kind of physical structure known as a "quasicrystal". Bringing these three together in such a way as to make a case for the role of quantum level processes in consciousness is the task Penrose sets himself. Our task will be to analyze and evaluate the argument.

To avoid standing in shadowy places, we wish to state at the outset that in our view, the argument consists of merest possibility piled upon merest possibility teetering upon a tippy foundation of "might-be-for-all-we-know's". It also rests on some highly dubious assumptions about the nature of mathematical knowledge. Our assessment is not that the Penrose hypothesis is demonstrably false, in the way that molecular biologists can confidently say that hereditary material is not a protein; it is DNA. Neurobiology is simply not far enough along to rule out the Penrose possibility by virtue of having a well established, well tested, neurobiological explanation in hand. Rather, we judge it to be completely unconvincing and probably false.

II. Compact Version of the Penrose Argument⁶

Part A: Nonalgorithmicity of human conscious thought.

A1) Human thought, at least in some instances, is sound⁷, yet nonalgorithmic (i.e. noncomputational). (Hypothesis based on the Gödel result.)

Pitowski, in honor of whom we call this the *Pitowski Syllogism*: (1) We really do not understand the nature of consciousness. (2) The only thing in the physical world we really do not understand are quantum level phenomena. (3) Therefore, these are probably the same mystery. As Pitowski is quick to point out, Premise (2) is clearly false.

⁶At least, this is our best, most sympathetic, and long-pondered shot at it. We do acknowledge that the presentation of the argument in SHADOWS is complex, the issues are complex, and that the compact account is a simplification. We have no interest in merely setting up a straw man, but there is value in having the crux of the argument clear.

⁷For purposes of this paper, soundness is a property of procedures or mechanisms or the exercise of capacities. Soundness can be taken to be roughly equivalent to 'truth producing', meaning that given true premises, when the normal functioning of the procedure or mechanism or capacity produces conclusions, they are true ones, not false ones. Soundness should thus not be thought of as necessarily tied to algorithmic or syntactic procedures, though of course these too might be sound. More generally: statements can be true or false, arguments can be valid or invalid, procedures etc. can be sound or not. Though our expression, "truth-producing", is less conventional than "truth-preserving" in this context, it is called for in order to accommodate Penrose's very interesting idea that things other than algorithms, formal systems, and the like, might be sound.

A2) In these instances, the human thinker is aware of or conscious of the contents of these thoughts.

A3) The only recognized instances of nonalgorithmic processes in the universe are perhaps certain kinds of randomness; e.g. the reduction of the quantum mechanical state vector. (Based on accepted physical theories.)⁸

A4) Randomness is not promising as the source of the nonalgorithmicity needed to account for (1). (Otherwise mathematical understanding would be magical.)

Therefore:

A5) *Conscious human thought, at least in some cases, perhaps in all cases, relies on principles which are beyond current physical understanding, though not in principle beyond any (e.g. some future) scientific physical understanding. (Via A1 - A4)*

Part B: Inadequacy of Current Physical Theory, and How to Fix It.

B1) There is no current adequate theory concerning the 'collapse' of the quantum mechanical wave function, but an additional theory of *quantum gravity* might be useful to this end.

B2) A more adequate theory of wave function collapse (a part, perhaps, of a quantum gravity theory) could incorporate nonalgorithmic, yet nonrandom, processes. (Penrose hypothesis.)

B3) The existence of quasicrystals is evidence for some such currently unrecognized, nonalgorithmic physical process.

Therefore:

B4) *Future theories of physics, in particular quantum gravity, can be expected to incorporate nonalgorithmic processes. (via B1 - B3)*

Part C: Microtubules as the means of harnessing quantum gravity.

C1) Microtubules have properties which make certain quantum mechanical phenomena (e.g. super-radiance) possible. (Hameroff/Penrose hypothesis.)

C2) These nonalgorithmic nonrandom processes will be sufficient, in some sense, to account for A5. (Penrose hypothesis.)

C3) Microtubules play a key role in neuron function.

C4) Neurons play a key role in cognition and consciousness.

C5) Microtubules play a key role in consciousness/cognition (by C3, C4 and transitivity).

Therefore:

⁸ Penrose does argue that chaotic processes are themselves algorithmic (Penrose 1994: 177-179), in that they can be simulated to any desired degree of accuracy by digital computational mechanisms. While some may feel that this subsumption of chaotic systems to algorithmic ones is unjustified, we propose to grant this premise.

C6) Microtubules, because they have one foot in quantum mechanics and the other in conscious thought, provide a window for nonalgorithmicity in human cognition.

THEREFORE:

D) Quantum gravity, or something similar, via microtubules, must play a key role in consciousness and cognition.

Briefly, our analysis of this argument indicates that A1 is most likely false, and Section III below provides some reasons for denying it. This undercuts the case for A5, and hence Part A. B3 is almost certainly false (this is the subject of Section IV), and given its falsity, B2 is entirely speculative as well. This undercuts the case for B4, and hence the conclusions of Part B are exposed as entirely speculative. C1 is quite speculative, C2 is no more than a guess, and C5 is simply a bad inference (these are discussed in Section V), and hence Part C looks tenuous. In short, it appears to us that even if D did happen to be true, the argument embodied in parts A, B and C provides no reason to believe that it is. In Section VI, we provide independent reasons for thinking that D is probably false.

III. Are there instances of conscious human reasoning that are sound and nonalgorithmic? Analysis of Premise A1

A. Insight, Pattern Recognition, and Artificial Neural Nets

The Gödel result forms the springboard of the reasoning underlying Penrose's premise A1, and we restate his argument below.⁹ As a preliminary, note that by nonalgorithmic is meant "noncomputable". This implies that the performance of the system could not be produced by any algorithmic procedure; more, it could not be *approximated* by an algorithmic procedure.¹⁰ The behavior of a river eddy is weakly nonalgorithmic insofar as it is a complex system and its states are continuous. It could, however, be *approximated* by an algorithm, *to any desired degree of accuracy*, and given Penrose's conventions, this entails that its behavior is algorithmic. "Nonalgorithmic" in the sense Penrose quite reasonably intends is, therefore, a *very* strong constraint; so strong in fact, that whether there exist any physical systems in the real world whose behavior is noncomputable in this

⁹For a brief but powerful criticism of Penrose on the Gödel result, see Putnam's review of SHADOWS in the New York Times Book Review, 1994.

¹⁰ Thus we will follow Penrose in using the term nonalgorithmic only for those processes which cannot even be approximated with an algorithm. This will contrast with 'weakly' nonalgorithmic, which just means not following explicit rules.

strong sense remains very much an open question. Penrose conjectures¹¹ that quasicrystals may be such phenomena. As for consciousness, the Penrose hypothesis is that given human mathematical performance, we can tell that the brain must be such a (deeply) noncomputable system.

Here is Penrose's own summary of the argument:

It is a kind of *reductio ad absurdum* argument, in which I try to show what would happen if we tried to construct robots with the kind of ability to understand that we have. Because the Gödel argument is basically about understanding; it tells us how to move from one formal system to a system outside that, from the understanding of what that system is trying to say. It is concerned with the question of the *meanings* of the symbols, which is a dimension that a computational system does not have; a computational system just has the rules which it follows. What one can do in mathematics is, by understanding the meanings of the symbols, one can go beyond the formal rules, and see what new rules must apply from those things, and one does this by understanding their meanings. (Penrose 1994a: 19.)

Granting that some instances of understanding and extending understanding do not involve *explicit* rule following, is there any framework for approaching such cognition? Indeed there is, and famously so. Artificial neural nets (ANNs) are capable of learning complex pattern recognition tasks as well as sensorimotor integration tasks¹². Once the weights have been set, typically by training on a range of cases, networks can perform very well on new cases, even giving "good" answers in cases that are nonstandard or missing bits or presented in unusual conditions. Given recurrent connections between units, ANNs can recognize sequences, for example sequences of sounds. Training a network is not programming with an explicit algorithm. The only algorithms in the neighborhood are relatively simple ones used to adjust the weights, in reinforcement learning or via a Hebbian learning rule, for example. In any case, the ANN has no *explicit* rules that govern its performance, any more than a child does when it successfully extends 'dog' beyond the family retriever to the neighbor's poodle and grandma's Great Dane. Ditto for concepts such as 'chair', 'cold front', 'promising student', 'fair', 'reasonable', and so forth.

Pattern recognition has been argued¹³ to be the key cognitive function of nervous systems, underlying not merely capacities such as recognizing a dog or a chair, but also, in the cognoscenti, recognizing a chess configuration and a theorem in the predicate calculus. In logic or mathematics, insight-cum-recognition can be followed up with a proof to

¹¹See footnote 34 below.

¹²Cf. Churchland and Sejnowski 1992; P.M. Churchland 1995; Jordan 1989.

¹³Cf. P.M. Churchland 1995.

determine whether one's insight was correct. On the other hand, for other highly complex patterns such as instances of injustice or insanity, for example, verifying insight may involve nothing so straightforward as application of a proof procedure. The general point, however, is that what gets called 'insight' and 'intuition' could very well be complex pattern recognition performed by recurrent neural networks.

The ANN processes are analog and parallel; the machine is flexible and plastic.¹⁴ These are indeed very striking capacities, and they are what make ANNs so exciting to robotics, artificial vision, and so forth, and what makes them relevant to real nervous systems. "Computing without rules" was indeed the popular watchword in early stages of connectionist research¹⁵. That ANNs can learn, rather than be programmed, that they have analog properties, that they are flexible, fault tolerant and can give answers to degraded inputs are, inter alia, what make ANNs far more suitable and powerful than classical programming techniques for many problems in the simulation of nervous system capacities. But uncanny they are not. Are their input-output functions *noncomputable* in the weak sense of not being instances of explicit rule following or discreet state transitions? Yes. Can the behavior of neural nets (artificial and otherwise) be *approximated by an algorithm*? So far as anyone knows, yes. This may, certainly, be a strained, semi-fictional sense of computable if no algorithm can execute the function in real time. Nevertheless, the "approximability" does mean that they fail to have the property Penrose is after, namely being noncomputable *and* "nonapproximatable". That is, they *are indeed computable and algorithmic, in the sense Penrose intends*. Now *if* Penrose is right in supposing human thought cannot be even approximated by an algorithm, then the success of ANNs is not, by itself, enough to subvert Penrose. At the risk of repeating ourselves, we do emphasize that it is not known whether *any* physical processes exist that are strongly noncomputable in the sense Penrose seeks. (See also below, Section IV on quasi-crystals.) *Even if they are not counterexamples to Penrose's hypothesis*, the success of ANNs teaches us that phenomena that appear intractable to conventional programming on a classical machine might very well be managed elegantly by a nonclassical, analog device.

B. What is Penrose's argument for Premise A1?

A1a) In order to ascertain mathematical truth, human mathematicians are not using a knowably sound algorithm.¹⁶

¹⁴Cf. Mead 1989.

¹⁵Cf. Churchland and Sejnowski 1992; P.M. Churchland 1995.

¹⁶ In brief, what Gödel showed was that for any sufficiently powerful consistent formal axiomatic system F, there will be true statements, expressible in F, yet not provable in F. In particular the statement that F is consistent, call this G(F), will not be provable in F, provided that F is in fact consistent. Curiously,

- A1b) The brain procedure that does underlie this 'ascertaining mathematical truth' is sound.
- A1c) If human mathematicians were using a sound *algorithm*, this algorithm would be knowable.
- A1d) Therefore, human mathematicians do not use an algorithm in order to ascertain mathematical truth.
- A1e) The understanding mathematicians employ is not different in kind from everyday human understanding and conscious thought.

Therefore:

Premise A1) Human thought, at least in some instances, perhaps in all, is sound, yet nonalgorithmic.

C. Critical Analysis of Penrose's Argument for Premise A1.

Penrose's arguments in favor of A1a and A1c are where most of the technical machinery is brought to bear. It may therefore be a relief that we propose, for convenience of strategy, to grant both of these premises, and focus rather on A1b.

Our point is this: *even if* humans are using some sort of algorithm, *and* that algorithm is knowable, A1a presents a problem *only if we assume that this algorithm is sound*. This is supposed to be not worth considering because an unsound procedure would license entailment to anything, and we can be sure that $(p \ \& \ \sim p)$ is not true. Matters are not quite so simple, however, because these are questions about human knowledge, not about what Eternal Immutable Truths really are on display in Plato's Heaven. By definition, Plato's Realm has only Truths. What is in the human mind/brain is a matter not of definition but of empirical fact. Our only access to Plato's Realm is through our brains, and our brains have to use cognitive procedures to figure things out. Were our knowledge system to contain an inferentially remote falsehood, some far-flung, plausible but false, proposition, we might have a hard time deploying the cognitive machinery to force it to the surface or to recognize it to be false. It is conceivable that one's mathematical understanding sequesters somewhere a false, but practically isolated, proposition masquerading as a truth. Thus we

G(F) will be provable in F if F is in fact inconsistent, i.e. if G(F) is false. In a nutshell, the argument for A1a is that since humans are sound (hence consistent), and since we can know that we are sound, we cannot be relying on any formal system for our knowledge. This is because if we were exclusively using some sound formal system, we could never know (prove) our own soundness. Penrose's treatment of this argument is much more complete (Penrose 1994b).

have to consider this possibility: humans could be using a cognitive procedure(s) which is unsound¹⁷ but *benignly* so -- perhaps because it includes the axiom of choice (explained below) -- or perhaps because it includes the negation of the axiom of choice, or for some other reasons altogether.¹⁸

To make Premise A1 plausible, Penrose must do three things: first, identify some range of phenomena -- some cognitive procedures, or set of insights, or whatever -- which are uncontroversially sound, and which are not known to be algorithmic. Let us call such a procedure, or insight, an S-procedure. Second, he must then invoke A1a to claim that in fact S-procedures cannot be supported by a knowable algorithm or approximated by a knowable algorithm. Third, he must argue for a presumed counterfactual: if S-procedures were supported by an algorithm, this algorithm *would in principle* be knowable. How successful is Penrose in satisfying the first of these three conditions? Not very, for the simple reason that there do not appear to be any S-procedures. Notice that anything which is knowably algorithmic cannot be an S-procedure, and so this rules out inference rules like *modus ponens*,¹⁹ anything formalizable in predicate logic or Zermelo-Fraenkel axiomatic set theory (ZF), and the like. And as we shall see, when these are excluded, there is reason to doubt that any procedures for mathematical deliberation are sound.

Mathematical thought, at least in some instances, is Penrose's prime candidate for a sound, noncomputable S-procedure. Errors in workaday human cognition are legion, belying any suggestion that sound procedures might be operative in nonmathematical reasoning, and Penrose clearly doesn't want to deny this²⁰. Is the case for mathematical thought better? Even in the domain of mathematics, mathematicians do make errors, errors that mathematicians themselves confess to be errors. Such errors are likely to be inobvious, and it can take months to determine whether a putative proof of, e.g. Fermat's Last

¹⁷ As Penrose himself notes, this seems to be what Turing thought was the real moral of the Gödel result. Turing is worth quoting at length: "It might be argued that there is a fundamental contradiction in the idea of a machine with intelligence. It is certainly true that 'acting like a machine', has come to be synonymous with lack of adaptability... It has for instance been shown that with certain logical systems there can be no machine which will distinguish provable formulae of the system from unprovable... Thus if a machine is made for this purpose it must in some cases fail to give an answer. On the other hand, if a mathematician is confronted with such a problem he would search around and find new methods of proof, so that he ought to be able to reach a decision about any given formula. Against it I would say that fair play must be given to the machine. Instead of it sometimes giving no answer we could arrange it so that it gives occasional wrong answers. But the human mathematician would likewise make blunders when trying out new techniques. It is easy for us to regard these blunders as not counting and give him another chance, but the machine would probably be allowed no mercy. In other words then, if a machine is expected to be infallible, it cannot also be intelligent. There are several mathematical theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretense at infallibility." (Turing (1986)).

¹⁸See footnote 24 below.

¹⁹*Modus ponens* has the form: if P then Q, P, therefore Q. *Modus tollens* has the form: if P then Q, not Q, therefore not P.

²⁰Cf. Penrose 1994b.

Theorem, is free of errors. To avoid impugning the underlying procedures, Penrose adopts the equivalent of a performance/competence distinction. Granted, the argument goes, mathematicians make mistakes, but these mistakes are merely performance mistakes, resulting from the misapplication of an *underlying sound competence*.

In linguistics, where Chomsky made famous the performance/competence distinction, many difficulties dog the task of rendering it precise. Crudely, one problem is that there does not appear to be any principled way to distinguish between these two cases: (1) The sentence Q is really grammatically correct but the native speakers have limitations on memory, attention etc. and hence conflict in judgment (2) Q is not really grammatical because native speakers conflict in judgment. The trouble arises when trying to adjudicate between competing theories of syntax, since the performance/competence distinction can always be invoked to insulate aspects of a formal syntactic theory from empirical disconfirmation. Restricting application of the distinction to unproblematic cases helps enormously.²¹ Thus it is important to identify sentences that both exhibit the relevant properties, but are simple enough so that judgment discrepancies cannot fairly be attributed to limitations in attention, memory, and so on. In other words, application is restricted to cases where the distinction does not obviously subvert experimental testing.

Comparably, to protect himself from begging any questions, Penrose must identify some set of mathematical capacities/abilities (our S-procedures) that are not only sound, but are simple enough, or short enough, such that performance errors are minimized or eliminated altogether. The secondary assumption is that performance on these tasks will result from deployment of the underlying sound procedures which support more complex and lengthy episodes of mathematical reasoning. So circumscribed, this set of capacities/abilities seem to be what Penrose intends by the phrase "The perception of unassailable mathematical truth." The danger in failing thus to circumscribe is that one may presume the actual truth of what is merely believed-to-be-true, and in consequence postulate capacities the brain does not actually have. These capacities/abilities and their underlying brain procedures are now our S-procedure candidates (or simply S-candidates). Do we know *they* are sound?

Trying to delimit instances of "the perception of unassailable mathematical truth" uncovers deep and troubling issues in the epistemology of mathematics generally. First, and most notoriously, what counts as unassailable in mathematics differs from century to century and from mathematician to mathematician. The brilliant 19th century mathematician Cauchy, for example, denied the existence of infinite sets. Infinite sets, he correctly

²¹Cf. Bates and MacWhinney 1989.

reasoned²² (and he was not alone), would have proper subsets with which they could be put into one to one correspondence. His mathematical intuitions led him to conclude that this would be a contradiction, and hence that the existence of such sets should be rejected as false. Thanks to Cantor, it can be shown that there is no contradiction, however much our intuitions bid us believe otherwise. The statement that Cauchy 'clearly perceived' as contradictory is now taken as a right and proper -- an unassailable, noncontradictory, teach-to-undergraduates -- *definition* of an infinite set. Why not view this as a development in mathematical understanding, not unlike progress in physics, chemistry, and biology?²³

Discrepancies in judgment (differences of opinion) are not merely a thing of the hoary old past, but exist today. One instance concerns the axiomatic status of the so-called axiom of choice, which is quite easy to state and understand: for any collection of non-empty sets, there exists another set that contains exactly one element from each set in the collection. Thus, as in the case of Cauchy's insight, when careful, sober, mathematical intuitions fail to coincide about the truth of the axiom of choice, the failure does not seem to be explainable on grounds of performance errors. Sound²⁴ procedures (to a first approximation, truth-producing procedures) cannot yield *both* a given result (S is true) and that result's denial (S is false), for if they do, they simply are not *truth*-producing. Nevertheless, some mathematicians' mathematical intuitions lead them to embrace the axiom of choice, while others' lead them to deny it²⁵. Some mathematicians, typically Platonists, have "perceived" the unassailable truth of propositions about transfinite sets, while constructivists "perceive" the contradictory nature of transfinite sets. For constructivists, the law of the excluded middle ($p \vee \sim p$) is not universally true and hence not a law; for *a priorists*, it is. And so on.²⁶

Penrose has a problem for his assumption regarding the "soundness" of human mathematical capacities when the careful, reflective performance of mathematicians on relatively simple statements, such as ($p \vee \sim p$), the axiom of choice, and the existence of infinite sets, is discrepant. He has essentially three options: (1) he could claim that some of the mathematicians (which ones? the constructivists?) are making performance errors in

²²Cf. Boyer 1959: 296.

²³Cf. Bloor, 1976: 131-156; Boyer 1959.

²⁴ Notice that the emphasis here is on *soundness*, and not just consistency. While it of course has been proven that the axiom of choice (and hence its denial) are independent of ZF, and thus consistent with it, it does not follow that both are *sound* when added to ZF. In fact, on any reasonable account of what soundness means, they cannot both be so.

²⁵ A particularly striking example of the tenuousness of the Penrose assumption concerning mathematical intuitions was encountered when one of us (Grush) was discussing these issues with a mathematics graduate student, who admitted, in good faith, the following: "I firmly believe that Zorn's lemma is true, and I'm convinced that Zorn's lemma is equivalent to the axiom of choice, and yet I am certain that the axiom of choice must be false."

²⁶ For a lucid and compelling discussion of a non-Platonist approach to mathematics, see Kitcher, 1984.

these and similar cases, care and IQ notwithstanding. This is unconvincing because the examples involved do not make heavy demands on attention and understanding, compared to many mathematical statements. (2) He could claim that *some* mathematicians do lack underlying sound procedures (competence), while others luckily have them. This is not an attractive option either, because those who allegedly depend on an unsound procedures (from a Platonists point of view that might be Cauchy or Dummett, for example) do in fact make significant insightful contributions to mathematics. At the very least, it is safe to say they understand Gödel's theorem. Now if some mathematicians (e.g. the constructivists and intuitionists) can do brilliant mathematics while having unsound mathematical understanding, how can we be sure this does not hold generally? (3) He could claim that these are not good examples of what he has in mind as S-candidates. This looks embarrassingly *ad hoc*, especially when something like the truth of an old saw like $(p \vee \sim p)$ can be reasonably assailed, albeit with complicated background argument. Moreover, if *these* examples will not do, if anything knowably algorithmic including simple inferences, such as *modus ponens* or anything else formalizable in first order logic or ZF will not do, what will?

Disappointingly, Penrose fails to provide any actual examples of background mathematical understanding which can be known to be sound. In fact, SHADOWS is rife with admissions of the form "...as mathematicians gain in experience, their viewpoints may well shift with regard to what they take to be unassailably true -- if they indeed ever take *anything* to be unassailably true,"²⁷. Given such admissions, why does Penrose still cling to the belief that mathematical performance is backed by sound procedures, as opposed to usually reliable, or heuristically useful procedures?

His core argument is "It would be an unreasonable mathematical standpoint that allows for a disbelief in the very basis of its unassailable belief system!"²⁸ That is, if a mathematician (or anybody else) has unassailable beliefs, then that person must believe that the procedures which support or result in those beliefs are sound. But does "A is unassailable" mean "A is certainly true" or does it mean "one is convinced that A is true"? Is $(p \vee \sim p)$ really unassailable for a Platonist, or it is rather that she is convinced that it is true while recognizing that some of her esteemed intuitionist colleagues do not take it to be true. She also knows full well that being utterly convinced that A is true is not a Divine Guarantee of the truth of A. This matters, because if there is any doubt at all, even a tiny doubt, that any of her beliefs really is true, then the inference to the soundness of their source is blocked. This does not mean she thinks that the underlying procedures must be hopelessly

²⁷Penrose 1994b: 103. Emphasis original.

²⁸Penrose 1994b: 131.

flawed, but only that there are some puzzling propositions where the reasonable person realizes there can be a difference of opinion.

Now it would seem a perfectly reasonable view, one often adopted by cautious, reflective humans, that some sources of information should be taken to be generally reliable, yet not entirely infallible, and that any of its entailments can be adopted with *very high confidence*, but not with the complete certainty a stroll in Plato's heaven might provide. Bus schedules, phone books, newspapers, college textbooks, mathematics textbooks, scholarly publications, eye witness testimony, opinions of experts, etc. etc., are all *known to be flawed from time to time, and hence known to be unsound*. Perhaps the *unreasonable* position is the one with standards so high as to *reject* any sources of information that might be flawed. On the contrary, it seems reasonable to admit the fallibility and retain some (perhaps quite high) measure of confidence in them.

Although respected mathematicians have undergone changes of mind, even on core foundational issues, Penrose's Platonism leads him to say, "...[it is] an unreasonable mathematical standpoint that allows for a disbelief in the very basis of its unassailable belief system" (1994b, p.131) Notice, finally, that "M does not believe that A is unassailably true" does *not* entail "M disbelieves A." It might merely mean that M is not certain, even though M takes A to be highly likely, and in daily life M acts as though it were unassailable. For example, consider the axiom of choice. One may be convinced of its truth, while not being prepared to stake one's life on it as an unassailable truth.

To make the point a bit more strongly, the inference from "I really believe that A is true," or "I don't see how A could *possibly* be false," to "A is unassailably true" is an inference rule known to be unsound. To take but one example, Kant thought it a necessary truth Euclidean geometry described physical space. That very rule *has*, in fact, led to falsehoods. Now either our mathematician does have beliefs that she takes to be absolutely unassailable, in which case she is unsound, because such a belief could only be the result of the unsound inference rule (or something similar) specified three sentences back, or the mathematician has beliefs held only with very great confidence, in which case her inference to the soundness (as opposed to reliability) of the source of those beliefs is not licensed.²⁹

Are there any alternatives to a Platonist ontology of mathematical objects (abstract, immutable objects -- truths, numbers etc.) and its usual companion, *an a priori* epistemology of mathematics (grasping with the intellect the absolute and immutable truths

²⁹Penrose might object that our argument relies on the variable or dicey mathematical competence of individual mathematicians, while his argument is couched in terms of the competence of the mathematical community as a whole (see discussion in Penrose 1994b: 97 - 101). This move would be rhetorically awkward, since presumably consciousness, in all its putative non-algorithmic glory, is supported by brains, and brains are supported by individuals.

in Plato's Heaven) ? Indeed there are³⁰. A major motivation for seeking a more satisfactory epistemology is just this: what are supposed to be the nature of the interactions between the Platonic Realm and the thinker's brain? If the denizens of the Realm cannot causally interact with anything, let alone human brains, how on earth can mathematical understanding be acquired by the human brain? Plato's own answer, namely, we do not learn, we only remember our soul's pre-birth observations with Truths in the Realm, is less than satisfactory.

Even supposing we retrofit Plato's account with supports from evolutionary biology, the idea still founders. Evolution may have wired in a variety of capacities, but in the life of primates and early hominids, there was no evolutionary pressure to acquire fancy mathematical knowledge (e.g. infinitesimal calculus), and some of mathematics surely had to be discovered by some and learned by others. Doubtless the struggle for survival meant that skills in planning, preparing, anticipating, communicating, and so forth would have had great value, and brain structures subserving such skills may well be deployable for culturally dependent skills such as reading, writing, and mathematics. It is harder to see, however, why *soundness*, as opposed to reliability, should be selected for. (Notice too that even if mathematical capacities are hard-wired in, this is no guarantee of the soundness of mathematical understanding.) Evolution is a satisficer, not an optimizer, and "approximately accurate" or "accurate for most of the likely cases", is often good enough.³¹ How humans come to have the conceptual and cognitive resources to develop formal systems, proof theory, and mathematical certainty is a puzzle, though not, perhaps, more intractable than how we have the resources to read and write, to compose and play music, to skate, hang glide and perform eye surgery. The idea, therefore, that mathematical capacity is an independent faculty of pure reason, whose exercise yields mathematical (or any other absolutely certain) knowledge by virtue of intuitive grasping of propositions and objects in Plato's Realm, is wanting in biological plausibility.

Cognitive models of higher cognitive processes in general, and mathematical cognition in particular, are not as far advanced as modeling of sensory processing and modeling of single neurons. Suffice it to say that the epistemology of mathematics has not kept pace with philosophical developments in other domains. Although the existence of the epistemological problems is well recognized amongst philosophers of mathematics, it is often of marginal interest to mathematicians themselves. One of the outstanding exceptions

³⁰Cf. Heyting 1956; Lakatos 1976; Benacerraf and Putnam 1983; Dummett 1991; Quine 1970; Kitcher 1984.

³¹Cf. Stich (1990) for an excellent critical discussion of the link between evolutionary pressures and truth.

is the recent work by Philip Kitcher ³², who does make a splendid attempt to bring the epistemology of mathematics up to date.

****FIGURE 2 ABOUT HERE ****

Platonism is surely a kind of convenient myth, rather like the way in which frictionless planes and ideal gases are convenient myths, or perhaps even as the "spirit of Christmas" or "zeitgeist" are convenient myths. As such, however, it cannot support grand metaphysical theories, such as that espoused by Penrose (Figure 2). Nor can it provide much in the way of significant constraints on the cognitive neurobiology of understanding and consciousness. When the "just-so" story is taken literally and used to constrain theories about natural phenomena, it positively gets in the way.³³

IV Quasicrystals: Argument against B3

A consequence of Penrose's Gödel-inspired arguments for strong (no-algorithm-can- even-approximate) noncomputability in the mind/brain is that current theories in physics are fundamentally incomplete. In Penrose's view, it ought to be augmented by a theory of quantum gravity. One might, however, prefer to *tollens* here rather than to *ponens*. In other words, one's confidence in current physics is a *prima facie* reason for denying the correctness of these Gödel-inspired arguments. The considerations given in the previous section means this option is not unattractive. From Penrose's perspective, therefore, it is important to have hard evidence for the existence of such postulated nonalgorithmic processes apart from the contentious case at hand, namely the mind/brain. There are sections in EMPEROR which could be construed as attempting to provide such evidence³⁴. The reasoning behind B3 we reconstruct as follows:

B3a) Because there exist sets of tiles which tile the plane only non-periodically, the question of whether a given set of tiles will tile the infinite Euclidean plane is not decidable (algorithmic).

³²Cf. Kitcher 1984. Also, see P.M. Churchland 1995.

³³ This is a point which we understand G. Kreisel to have made in correspondence with Francis Crick.

³⁴ These are pages 132-138 and 434-439. Fixing an adequate interpretation of these sections is difficult, as it is unclear if Penrose takes quasicrystals to be evidence for nonalgorithmic processes, or simply non-local but algorithmic ones. The paragraph bridging pps. 438 and 439 in EMPEROR seems to favor the reading where nonalgorithmicity is at issue. Curiously, there is no mention of quasicrystals in SHADOWS, perhaps because Penrose doesn't (any longer?) take them to provide evidence of such physical actions. If this is the case, then perhaps he would agree with the discussion in this section.

B3b) Some of these non-periodic sets tile the plane with a five-fold symmetry.

B3c) There exist 'quasicrystals' whose lattice structure exhibits a similar five-fold symmetry.

B3d) So the growth of these crystals depends on nonalgorithmic processes (maybe).

Therefore:

Premise B3) The existence of quasicrystals is evidence for some such currently unrecognized, nonalgorithmic physical process.

The argument is unconvincing. The chief problem is that B3d and hence B3 simply do not follow from B3a - B3c. First, even if the analogy between the structure of quasicrystals and the non-periodic tilings were close, the 'problems' whose computability is at issue in each case are different. In the case of the tiling problem, the undecidable feature is not how to put the tiles together in order to tile some region (putting the tiles together may very well be algorithmic), but whether or not given a set of tiles, and perhaps some way of putting them together, these tiles can cover the entire infinite Euclidean plane without gaps or overlaps. Second, the analogy simply does not hold. The reason is that *quasicrystals are in fact finite*, unlike the infinite Euclidean plane, and hence their growth undoubtedly is computable. Indeed, algorithms have been proposed³⁵ for the growth of just such crystals (as Penrose himself notes!³⁶). Although the information required to determine the appropriate arrangement might not be locally available to individual atoms in the lattice, that is not a problem as far as algorithmicity goes.

Without this premise, the whole of Part B of the Penrose argument -- that there are in fact nonalgorithmic (in the strong sense characterized earlier, p. 5) processes in the universe -- amounts to no more than unsupported speculation. Now it might be wondered whether we are uncharitably pillorying an argument that Penrose himself no longer favours nor deploys in SHADOWS. However that may be, the fact is that *the speculation that such processes really do exist* continues to be a crucial underpinning in the complex structure of the overarching argument. Naturally enough, identifying this claim as speculative does *not* entail that it is actually false, and we are prepared to admit that despite the absence of evidence, this speculation could turn out to be correct. Our point here is practical and simple: because investment of research time and energy is often made with the background probabilities in mind, it is important to recognize a bald speculation as a bald speculation.

³⁵Cf. Onoda et al. 1988; Sasajima et al. 1994.

³⁶Penrose 1989: 449 n.7.

Forewarned, one might nevertheless decide to throw one's lot in with the speculation anyhow.

V: Are Microtubules the Generators of Consciousness?

We come now to the specific hypothesis about how quantum gravity and consciousness are parts of the same mystery. As the idea that microtubules are the key originated with and is mainly articulated by Stuart Hameroff, later to be adopted by Penrose, we shall focus first on the story as it comes from Hameroff.³⁷

Here is Hameroff's summary statement of the conjecture:³⁸

To summarize, cytoskeletal microtubules are likely candidates for quantum coherence relevant to consciousness because:

- Microtubule individual subunit (tubulin) conformation may be coupled to quantum-level events (electronic movement, dipole, phonon) in hydrophobic protein regions.
- Microtubule paracrystalline lattice structure, symmetry, cylindrical configuration and parallel alignment promote long-range co-operativity and order.
- Hollow microtubule interiors appear capable of water-ordering, waveguide super-radiance and self-induced transparency.

First, we note that all body cells have microtubules, where a major function is to support cell division. In neurons, one of their known functions is transport, on their outside surface, of molecules such as neurotransmitters and various proteins, between the cell body and the axon, and between the cell body and the dendrites. (See Figure 3.) Now it is generally believed that some general anaesthetics (hydrophobic ones) alter the neuronal membrane receptors and protein channels, with an effect on protein water-binding. Hameroff builds on these data by arguing that the protein constituting the microtubules (tubulin) might also have *its* water-binding properties affected. *If* it did, and *if* tubulin structures did have the quantum mechanical properties of 'super-radiance'³⁹ and 'self-induced transparency', then, the argument goes, the loss of consciousness might be attributable to these changes in the microtubules.

³⁷Cf. Hameroff 1994; Hameroff, Rasmussen and Mansson 1989; Hameroff and Watt 1982; Hameroff et al. 1992; Jibu et al. 1994.

³⁸Hameroff 1994: 105.

³⁹In the context of the present discussion, super-radiance is a quantum mechanical effect in which water molecules within the microtubule might act in ways roughly analogous to a laser. Specifically these molecules, because they have dipole moments, could coherently emit radiation by shifting between angular momentum states. This effect, if it exists, would depend crucially on the purity of the water within the tubule, as well as on the properties of the tubule itself, such as its diameter and minute details of its electric field. See Jibu et al. 1994; Hameroff 1994.

**** Figure 3 about here ****

Briefly, here are some reservations:

A. The anesthesia/microtubule connection

(1) There is no direct evidence that changes in microtubules in neurons are responsible for the phenomenological effects of general anesthetics. What the most recent data do indicate is that at surgical concentrations, ligand-gated ion channels (proteins) in the neuronal membrane are the main sites of anesthetic effects. Relative to the quantum effects envisaged for microtubules, these are *large* effects. In their review article in *Nature* (1994), Franks and Lieb state that among possible receptor targets, GABA_A (γ -aminobutyric acid) receptor protein has been established by electrophysiological studies to be a major target. The general anesthetics that potentiate the GABA receptor protein include inhalational agents such as halothane, enflurane and isoflurane, as well as intravenous anesthetics such as pentobarbital, propofol and alphaxalone. GABA is the major inhibitory neurotransmitter in the brain, and potentiation of the GABA receptor probably up-regulates inhibition which effectively overrides excitation. This is consistent with the finding that anesthetic interactions with the GABA receptor protein extend its open time. The other receptor that has been shown to be a target is the excitatory voltage-dependent NMDA receptor. It is inhibited by the agent, ketamine.⁴⁰ To repeat, these effects on receptors are large effects in the millivolt range.

(2) Even if disruption of microtubule function were consistently correlated with general anesthetics, there is no reason to suppose that "normal" microtubule functioning is anything more than a necessary *background* condition -- one necessary condition among hosts of others (e.g. availability of oxygen, ATP, glucose etc.).

B. The microtubule/super-radiance/consciousness connection

(3) There is no evidence⁴¹ that quantum coherence involving super-radiance (or anything else for that matter) occurs in microtubules. At best, what Hameroff has done is to

⁴⁰Franks and Lieb 1994; for further discussion, see Bowdle, Horita and Kharasch 1994.

⁴¹In this paragraph, as well as in (1) above, we are in the awkward situation of merely stating that there is no direct evidence of such and such. This is not because we are simply discounting Hameroff's evidence, but because, so far as we can see, he doesn't offer any. His arguments take more the form of demonstrating how certain phenomena might be possible, not of providing direct evidence for them.

show that it might be possible. This should most definitely be distinguished from providing *evidence* that it is *actual*.

(4) It is highly unlikely that the pore of the tubulin tube contains nothing but pure water, since there is no known mechanism for keeping out common cytoplasmic ions such as calcium and sodium. This is a major problem for the hypothesis, because impurities are an obstacle to the postulated long-range co-operativity, especially super-radiance. It is therefore, highly speculative that quantum coherence involving super-radiance occurs in microtubules.

(5) An ancient and still common palliative for the disease known as gout is the drug, colchicine. This was introduced to American medical practice by Benjamin Franklin. For our purposes, colchicine has the highly interesting property that, *inter alia*, it disrupts microtubules by attaching to the "add-on" end and preventing repolymerization repairs. After a period of colchicine treatment for gout, therefore, the microtubules show considerable depolymerization.⁴² This is a major monkeywrench for super-radiance⁴³, a bit like the way in which breaks in fibers disrupt fiber optics transmission. The capacity of the microtubule to support the self-focusing soliton wave⁴⁴ (which 'transmits' the super-radiance effect down the length of the tube) depends crucially on the characteristics and the uniformity of the waveguide, which just is the interior of the microtubule in this case. So any disruption as radical as depolymerization of the tubule prevents any quantum phenomena that might have been. If Hameroff and Penrose are right, then microtubule depolymerization ought to block super-radiance and hence significantly impair consciousness. Does it impair consciousness? Not that anyone has noticed. And loss of awareness is a major thing to not notice. It is known, however, that prolonged use can

⁴² Goodman, Gilman, et al. 1990. This was brought to our attention by Chuck Stevens. For evidence that small, though 'effective' amounts of colchicine do pass the blood-brain barrier, see Bennett, Alberti and Flood (1981). A number of studies (such as Kolasa et al. (1992) and Emerich and Walsh (1991)) injected colchicine directly into the brain of rats and then ran various behavioral tests on them, and no such studies we have seen mention any problems associated with consciousness, including Bensimon and Chermat (1991), who injected rat brains directly with colchicine every day for 10 days, Conner and Varon (1992) who directly injected the brain at several different locations, and Ceccaldi et al. (1990) who pumped colchicine continuously into rat brains over extended periods of time with an osmotic pump. Note that all of these studies use colchicine *specifically because it depolymerizes microtubules*, disrupting their transport function. It is revealing to note that many of these studies mention that the rats were anaesthetized before being sacrificed, which implies that the microtubule disruption *did not* render them unconscious, and that the normal anaesthetic *did* render them unconscious (presumably in some manner other than effecting the already-depolymerized microtubules, contra Penrose/Hameroff).

⁴³ And probably any other putative quantum mechanical effect supported by microtubules as well. In fact, it seems also to be a problem for other accounts of the importance of microtubular computation (quantum mechanical or otherwise) as supporting consciousness or cognition, e.g. Hameroff 1994; Hameroff and Watt 1982; etc.

⁴⁴Cf. Jibu et al. 1994.

cause paralysis, beginning with the lower extremities. The colchicine data are surely an embarrassment for the Hameroff hypothesis.

(6) Hameroff envisages consciousness as involving a unity across diverse brain regions. Even if the interesting quantum events did occur in a single tubule, to play a role in consciousness the effect must be transmitted from one tubule to its microtubule neighbor within the cell. If big transmitter molecules are in the vicinity, making their way to and from the synapse, the prediction is that this would seriously inhibit the spread of the quantum coherence. The next-stage-up problem has the same form: to play a role in consciousness the effect must be transmitted from one *neuron* to other neurons. The problem with this step is that the principal neuromembrane effects involving voltage changes are big -- on the order of tens of microvolts -- and it is a fair prediction that these effects would wipe out the nanoeffects from the microtubules.⁴⁵ This is not to say it cannot be done, but no one has the slightest evidence that it is done, nor the slightest idea how it might be done.

(7) Given the Hameroff hypothesis, one might predict that disruption of microtubule function underlies other more routine changes in the state of awareness, such as the daily shift from being awake to being in deep sleep. This shift is generally regarded as a major change from being conscious of events to not being conscious of them.⁴⁶ There is no reason to suppose that microtubules alter function concordant with these state changes, for example by ceasing their water-ordering, wave-guided super-radiating and self-induced transparency. What we do know is that certain neurons do change in very specific ways with changes in sleep/wakeness -- in thalamic and brain stem structures in particular.⁴⁷

Penrose's preferred role for microtubules differs a bit from that of Hameroff. He sees them as altering neuronal signaling via modifying pre-synaptic efficacy. The idea seems to be this: if they can do this, then perhaps they can shape specific patterns of activation in the brain, and these particular patterns are what support consciousness. So what microtubules must do is to somehow encode the information derived from sensory structures, process it, and then modify the firing of neurons in such a way as to support consciousness of the stimulus, and perhaps a purposeful response as well. How plausible is this?

Here are some reservations specific to the Penrose version of the microtubule story:

(1) First, microtubules are seldom seen in close proximity to the business end of the

⁴⁵Of course, depending on the story told about exactly what kinds of quantum coherence are important, and how they are maintained in spatially separate regions, the presence of transmitter molecules and electric fields may not be a problem. The problem is that these details are not provided.

⁴⁶See Llinas and Ribabry (1993), and Flanagan (forthcoming).

⁴⁷Cf. Steriade et al. 1993.

synaptic complex. Generally, they appear to end about a micron from the synaptic complex, which means, for the kind of effect Penrose is talking about, the distance may as well be meters.⁴⁸

(2) How is the microtubule supposed to communicate with the synapse to have the Penrose effect? What precisely is supposed to be the effect on the neuronal membrane and how is it to be achieved? Penrose does not give us a clue. The release of neurotransmitter vesicles, for example, do not have any characteristic association with microtubules, so far as is known.

(3) Suppose Penrose has in mind that the alleged quantum goings-on in the microtubules modify neurotransmitter release. Is this reasonable? Not very. What is known is that release of a vesicle of neurotransmitter when a spike reaches the axon terminal is highly dependent on calcium channels and on the phosphorylation rate of membrane proteins. Neuromodulators (e.g. norepinephrine, any of the various neuropeptides, caffeine etc.) can affect calcium channels and hence affect transmitter release. Caffeine, for example, works by blocking the receptor adenosine, which acts with a second messenger to down-regulate the neuron. These are very big effects, in the world of neurons. Even if microtubules did display the quantum goings-on, the effects would seem to be trivial relative to the effects of neuromodulators.⁴⁹

(4) The encoding problem for microtubules is ignored by Penrose. If microtubules are to perform their Penrosian function, they must be able to *get* the information that they then (nonalgorithmically) process. Neurons depend, for *their* signals, mostly on neurotransmitters and neuromodulators which alter the neuron membrane's permeability to various ions. Is the idea that microtubules can use *these* ion concentrations as sources of information? Maybe, but this seems to conflict with the requirement that microtubules be insulated from their ionic environment in order to support quantum coherence, etc.⁵⁰ Even if that can be resolved, how is this supposed to work?

Even with generous conjecture-granting, Penrose is still not out of the woods. As Hameroff in (3) above has problems about getting a global effect out of a highly local effect, so of course does Penrose. How is it that microtubule clusters in different neurons coordinate their activities in order to shape the overall patterns of neural activity which support consciousness? For surely changes in a single neuron will not cause the generation or loss of consciousness. No answer. Are microtubules able to support some form of long range quantum coherence or entanglement? No answer. Quantum entanglement we assume

⁴⁸ See Peters, Palay, and Webster (1978).

⁴⁹ See Stevens and Wang (1994).

⁵⁰ Ions in close proximity to the tubulin dimers would effect their electrical properties, and hence their capacity to engage in or support these sorts of quantum phenomena. See Jibu et al. 1994; Hameroff 1994.

is ruled out, because it would require that the insulated interiors or surfaces of the microtubules interact in some way, and since they will be in separate neurons across wide stretches of the brain, its not clear how this could happen. Finally, how realistic is it to predict a quantum coherent state between the spatially separated microtubules (or their contents)? Penrose correctly assesses the prospects when he remarks (*Shadows* p. 373) "Such a feat would be a remarkable one -- almost an incredible one -- for Nature to achieve by biological means."

VI Conclusions

Let's take stock of where we are. The first part of the Penrose argument (Part A) is in trouble. There is reason to doubt that human cognition or consciousness must take advantage of nonalgorithmic processes, since unsound, albeit reliable, algorithmic processes escape Gödel's net. Second, Part B of the argument is entirely speculative -- there is no evidence at all (including quasicrystals) that there are any (strongly) nonalgorithmic processes anywhere in the universe. Indeed, even if Penrose's quantum gravity hunch were correct (another big 'if'), it need not incorporate any nonalgorithmic processes. Finally, the prospects for Part C look grim. There is no experimental evidence that microtubules do support interesting quantum phenomena, and save in the unlikely event that the water in and around them is pure, it is doubtful that they do. Water purity is not the only problem; other highly speculative conditions would have to obtain also. Even if they can support such phenomena, long range coherence seems quite far fetched. Furthermore, because microtubules can depolymerize without noticeable effect on consciousness, it seems unlikely that they support conscious thought.

Consciousness is a problem, but it must be remembered that we are still in the very early stages of understanding the nervous system. Many fundamental questions about basic phenomena -- such as the role of back projections, the nature of representation in sensory system, whether sensory systems are hierarchically organized, precisely how memory is stored and retrieved, how sensori-motor integration works, what sleep and dreaming are all about -- have not yet been satisfactorily answered. Is the problem of consciousness utterly different from all these other problems? Perhaps, but *qua* mystery it does not come with its degree or depth or style of mysteriousness pinned to its shirt. Sometimes problems that appear to be the really tough ones, such as the composition of the stars, turn out to be easier to solve than seemingly minor puzzles, such as the precession of the perihelion of Mercury.

Undoubtedly many surprises await us, and for all we know, some of them may involve surprises for (or from) physics. Nevertheless, before making a heavy research investment in a precarious and far-fetched hypothesis, it would be nice to have something solid to go on. This may be a matter of taste, however.

Despite the rather breathtaking flimsiness of the consciousness-quantum connection, the idea has enjoyed a surprisingly warm reception, at least outside of neuroscience. One cannot help groping about for some explanation for this rather odd fact. Is it not even *more* reductionist than explaining consciousness in terms of the properties of networks of neurons? Emotionally, it seems, the two reductionist strategies arouse quite different feelings. After some interviewing, in an admittedly haphazard fashion, we found the following story gathering credence.

Some people who, intellectually, are materialists, nevertheless have strong dualist hankerings -- especially hankerings about life after death. They have a negative "gut" reaction to the idea that neurons -- cells that you can see under a microscope and probe with electrodes, brains you can hold in one hand, and that rapidly rot without oxygen supply -- are the source of subjectivity and the "me-ness of me". The crucial feature of neurons that makes them capable of processing and storing information is just ions passing back and forth across neuronal membranes through protein channels. That seems, stacked again the "me-ness of me", to be disappointingly humdrum -- even if there are lots of ions and lots of neurons and lots of really complicated protein channels.

Quantum physics, on the other hand, seems more resonant with those residual dualist hankerings, perhaps by holding out the possibility that scientific realism and objectivity melt away in that domain, or even that thoughts and feelings are, in the end, the fundamental properties of the universe.⁵¹ Explanation of something as special as what makes me *me*, should really involve, the feeling is, something more "deep" and mysterious and "other worldly" than mere neurons. Perhaps what is comforting about quantum physics is that it can be invoked to 'explain' a mysterious phenomenon without removing much of the mystery, quantum physical explanations being highly mysterious themselves.

Now we are *not* for a moment suggesting that anything like this is behind Penrose's work, and whether our diagnosis is right or wrong has no bearing whatever on the strengths and weaknesses of his arguments. It may, however, help explain why the very possibility of a quantum physical explanation is often warmly greeted, whereas an explanation in terms of neurons may be considered "scary", "degrading" and even "inconceivable". Why should it be less scary, reductionist or counter-intuitive that "me-ness" emerges from collapse of a wave function than from neuronal activity?

⁵¹ Cf. Bennett, Hoffman and Prakash 1989.

Nothing we have said in this paper demonstrates the falsity of the quantum-consciousness connection. Our view is just that it is no better supported than any one of a gazillion caterpillar-with-hookah hypotheses.

Acknowledgments: We would like to thank the following people for valuable discussions, advice and insights: Oron Shagrir, David Chalmers, Paul Churchland, Francis Crick, Mark Ellisman, Andrew Hibbs, Brian Keeley, Christof Koch, Steve Quartz, Terry Sejnowski, Chuck Stevens, Timothy van Gelder, Hal White, Robin Zagone, the participants of EPL (Experimental Philosophy Lab, UCSD), and three helpful anonymous reviewers.

References:

- Bates, E., MacWhinney, B. (1989). Functionalism and the Competition Model. In *The Crosslinguistic Study of Sentence Processing*. (ed. Bates, E., MacWhinney, B.) Cambridge: Cambridge University Press.
- Benacerraf, P., and Putnam, H. (ed.) (1983). *Philosophy of mathematics; selected readings*. Englewood Cliffs, N.J.: Prentice-Hall
- Bennett, B.M., Hoffman, D.D., Prakash, C. (1989). *Observer mechanics*. San Diego: Academic Press.
- Bennett, E., Alberti, M.H., and Flood, J. (1981) Uptake of [³H]Colchicine into brain and liver of mouse, rat, and chick. *Pharmacology, Biochemistry and Behavior* 14(6):863-869.
- Bensimon, G. and Chermat, R. (1991) Microtubule disruption and cognitive defects: Effect of colchicine on learning behavior in rats. *Pharmacology, Biochemistry and Behavior* 38(1):141-145.
- Bloor, D. (1976). *Knowledge and Social Imagery*. Chicago: The University of Chicago Press.
- Bogen, J. E. (1995). On the neurophysiology of consciousness: I. An overview. *Consciousness and Cognition* 4: 52-62.
- Boyer, C. (1959). *The History of the Calculus and its Conceptual Development..* New York: Dover Publications.
- Bowdle. T. A., A. Horita, and E. D. Kharasch (1994). *The Pharmacological Basis of Anesthesiology*. New York: Churchill Livingstone.
- Bray, Dennis (1992). *Cell Movements*. Garland Publishing & Co. New York

Ceccaldi, P.E., Ermine, A., Tsiang, H. (1990) Continuous delivery of colchicine in the rat brain with osmotic pumps for inhibition of rabies virus transport. *Journal of Virological Methods*, 28(1):79-83.

Churchland, P.M. (1995). *The Engine of Reason, The Seat of the Soul*. Cambridge: MIT Press.

Churchland, P.S.& Sejnowski. T. (1992). *The Computational Brain*. Cambridge: MIT Press.

Conner, J.M. and Varon, S. (1992) Distribution of nerve growth factor-like immunoreactive neurons in the adult rat brain following colchicine treatment. *Journal of Comparative Neurology* 326(3):347-62.

Crick, Francis. (1994) *The Astonishing Hypothesis*. New York: Scribners

Damasio, A.R. (1994). *Descartes' Error*. New York: Putnam.

Damasio, A. R. & Damasio, H. (forthcoming). Images and subjectivity: Neurobiological trials and tribulations. In : *The Churchlands and their Critics*. (ed. Robert McCauley). Oxford: Blackwells.

Dennett, D.C. (1990). *Consciousness Explained*. Little, Brown and Company.

Dummett , M (1991). *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Emerich, D.F., and Walsh, T.J. (1991) Ganglioside AGF2 prevents the cognitive impairments and cholinergic cell loss following intraventricular colchicine. *Experimental Neurology* 112(3):328-37.

Flanagan, O. (forthcoming). Prospects for a unified theory of consciousness, or, what dreams are made of. In: *Scientific approaches to the question of consciousness:25th Carnegie Symposium on Cognition*. Ed. J. Cohen and J. Schooler. Hillsdale, N.J.: L. Earlbaum.

- Franks, N. P. and W. R. Lieb (1994). Molecular and cellular mechanisms of general anaesthesia. *Nature*. 367: 607-614.
- Goodman, L., Gilman, A., et al. eds. (1990). *Pharmacological basis of therapeutics*. 8th ed. New York : Pergamon Press.
- Hall, Z. W. ed. (1992). *An Introduction to Molecular Biology*. Sinauer.
- Hameroff, S.R. (1994). Quantum Coherence in Microtubules: a neural basis for emergent consciousness? *Journal of Consciousness Studies*, **1**, 98 - 118.
- Hameroff, S.R., Dayhoff, J.E., Lahoz-Beltra, R., Samsonovich, A., Rasmussen, S. (1992). Models for molecular computation: conformational automata in the cytoskeleton. *IEEE Computer (Special issue on molecular computing)*, 30-39.
- Hameroff, S.R., Rasmussen, S., and Mansson, B. (1989). Molecular automata in microtubules: basic computational logic for the living state? In *Artificial Life, SFI Studies in the sciences of complexity* (ed. C. Langton). New York: Addison-Wesley.
- Hameroff, S.R. and Watt, R.C. (1982). Information processing in microtubules. *Journal of Theoretical Biology*, **98**, 549-61.
- Heyting, A. (1956). *Intuitionism: an introduction*. Amsterdam, North-Holland Pub. Co.
- Jibu, M., Hagen, S., Hameroff, S.R., Pribram, K.H., and Yasue, K. (1994). Quantum optical coherence in cytoskeletal microtubules: implications for brain function. *BioSystems*, **32**, 195 - 209.
- Jordan, M.I. (1989) Serial order; a parallel distributed processing approach. in *Advances in connectionist theory* (ed Elman, J.L. and Rumelhart, D.E.) Hillsdale, NJ: Erlbaum.
- Kitcher, P. (1984). *The nature of mathematical knowledge*. Oxford: Oxford University Press.

- Kolasa, K., Jope, R.S., Baird, M.S., Johnson, G.V. (1992) Alterations of choline acetyltransferase, phosphoinositide hydrolysis, and cytoskeletal proteins in rat brain in response to colchicine administration. *Experimental Brain Research* **89**(3):496-500.
- Lakatos, I. (1976). *Proofs and Refutations*. Cambridge: Cambridge University Press.
- Llinas, R.R. and Ribary U., (1993). Coherent 40-Hz oscillation characterizes dream state in humans. *Proc. Natl. Acad. Sci.*, **90**, 2078-2081.
- Llinas, R.R. and Pare, D. (1993). Of dreaming and wakefulness. *Neuroscience*, **44**, 3:521 - 535.
- Logothetis, N. and Schall, J.D. (1989). Neural correlates of subjective visual perception. *Science*, **245**, 753-761.
- Mead, C., (1989). *Analog VLSI and neural systems*. Reading, MA. Addison-Wesley.
- Onoda, G.Y., Steinhardt, P.J., DiVincenzo, D.P., and Socolar, J.E.S., (1988). Growing perfect quasicrystals. *Phys. Rev. Letters*, **60**, 2688.
- Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.
- Penrose, R. (1994a). Interview with Jane Clark. *Journal of Consciousness Studies*, **1**, 1:17-24.
- Penrose, R. (1994b). *Shadows of the mind*. Oxford: Oxford University Press.
- Peters, A., S. L. Palay, and H. deF. Webster (1978). *The Fine Structure of the Nervous System*. Philadelphia: W. B. Saunders Ltd.
- Quine, W.V.O. (1970). *Philosophy of Logic*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Sasajima, Y., Adachi, K., Tanaka, H., Ichimura, M., et al. (1994). Computer simulation of the growth process of binary quasicrystals. *Japanese Journal of Applied Physics, Part 1 - Regular Papers, Short Notes and Review Papers*, **33**, 5A:2673-2674.

Sossin, W. S., Fisher, J. M., and Scheller, R. H. (1989). Cellular and molecular biology of neuropeptide processing and packaging. *Neuron* , **2**: 1407-1417.

Steriade, M., McCormick, D.A., and Sejnowski, T.J. (1993). Thalamocortical oscillations in the sleeping and aroused brain. *Science*, **262**, 5134:679-685.

Stevens, C.F., and Wang, Y.Y. (1994). Changes in reliability of synaptic function as a mechanism for plasticity. *Nature*, **371**, 6499:704-707.

Stich, S. (1990). *The Fragmentation of Reason*. Cambridge: MIT Press.

Turing, A. (1986). Lecture to the London Mathematical Society on 20 February 1947. In *A.M. Turing's ACE report of 1946 and other papers* (eds. B.E. Carpenter and R.W. Doran). The Charles Babbage Institute Reprint Series for the History of Computing, vol. 10. Cambridge, MIT Press.

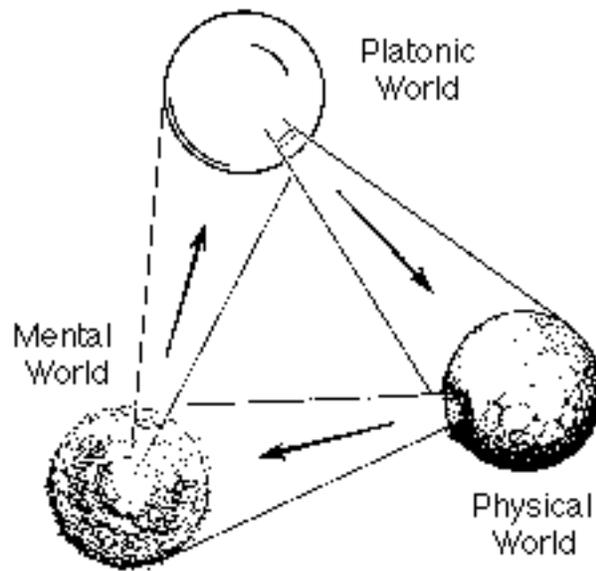


Figure 2: Penrose's Three Worlds. The drawing schematically illustrates the idea that the Physical World can be thought of as a projection from part of the Platonic World of eternal Truths, the Mental World arises from part of the Physical World (presumably the brain), and that the Platonic World is 'grasped' somehow during some mental activities. From Penrose 1994b.

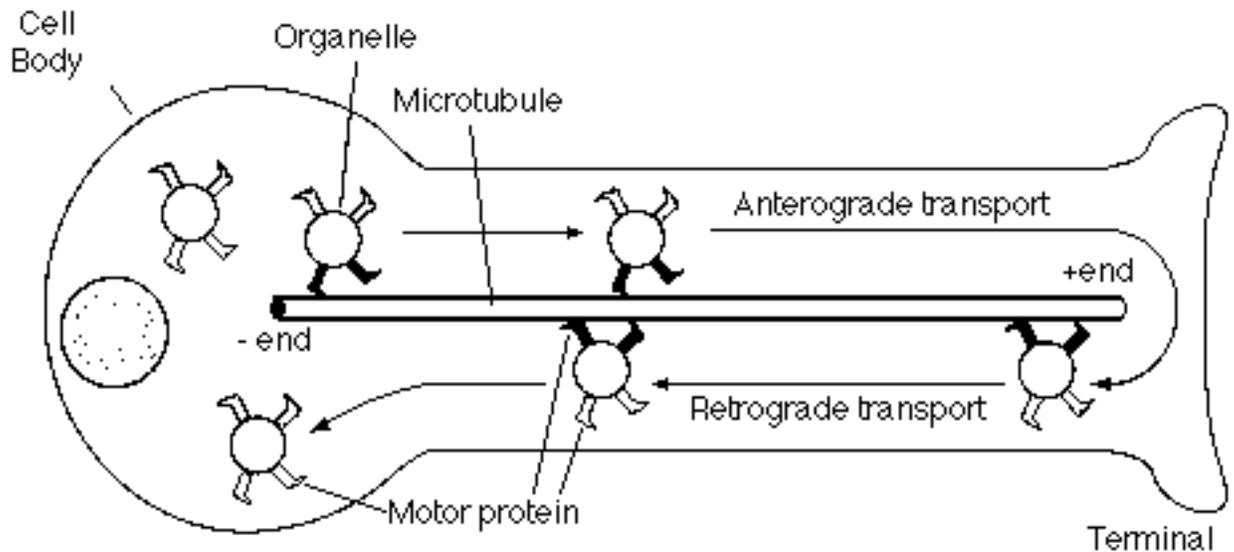


Figure 3: Very schematic drawing of a neuron, showing the microtubule's role as a structure along which organelles get transported from the cell body to the terminal and back. Adapted from Hall, Z. (ed.) 1992.

FIGURE 1 LEGEND

Electron micrograph of a longitudinally oriented ultrathin section (>500 Angstroms) through a myelinated axon near a node of Ranvier in rat optic nerve. The preparation is stained with Mg Uranyl Acetate and Pb citrate to highlight cytoskeletal structures and membranes. Microtubules are clearly delineated (arrow) as are the smaller neurofilaments

(arrowhead). Scale bar: 0.1 μ m. (This micrograph was kindly provided by Mark Ellisman at UCSD.)