The discomforts of dualism

A Review of Roger Penrose's *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*

(Appears in *Behavioral and Brain Sciences* **13**, 4, December 1990, 673–674)

Bruce MacLennan
Department of Computer Science,
University of Tennessee,
Knoxville, TN 37996-1301
Electronic mail: `maclennan@cs.utk.edu`

Penrose makes a Herculean attempt to give popular accounts of computability theory, special and general relativity, quantum mechanics and black holes, all with the aim of showing the relation of "computers, minds and the laws of physics." Unfortunately, I believe that he has failed in this attempt, and that the cause of this failure is a pervasive dualism. I'll discuss three issues where dualism gets Penrose into trouble.

1 Mathematical Thinking

Gödel's incompleteness theorem shows that any consistent formal system has an undecidable proposition, but that this very proposition can be shown to

be true by a metamathematical argument that appeals to the meaning of the proposition. Penrose attaches great significance to this fact, and uses it as the justification for many of his speculations. Since any formal system has such an undecidable proposition, and since it can always be proved by the metamathematical procedure, Penrose claims (p. 110) that this means that "mathematical truth is something that goes beyond mere formalism" (p. 111). From this he concludes that mathematical thought cannot be reduced to an algorithm, and hence that the mind cannot be equivalent to a computer (see also p. 118). I agree with his conclusions, but not with his argument.

First recall that 'metamathematical' refers only to the use of mathematical techniques to reason about mathematics; the metamathematical proof uses no esoteric techniques, nor does it depend on special, deep mathematical insights. In fact, the metamathematical proof is easily formalized. If Q is the undecidable proposition (constructed by the Gödel procedure) for a formal system $\mathcal{F}$, and if $\mathcal{F}'$ is a formal system powerful enough to talk about the truth of propositions in $\mathcal{F}$ (also easily constructed), then Q can be proved in $\mathcal{F}'$ by a formalized version of the metamathematical procedure. Thus the metamathematical proof does not make use of any inherently unformalizable procedures, and hence provides no evidence for nonalgorithmic mental powers.

It might be objected that the informal metamathematical proof can be carried out once for all formal systems, whereas the formal proof requires for each formal system $\mathcal{F}$ a new formal system $\mathcal{F}'$ in which to construct the proof. This is because informal mathematics can talk about the truth of all propositions, including those of informal mathematics. In fact we can accomplish the same thing formally by constructing a formal system $\mathcal{F}^*$ capable of expressing propositions about the truth of its own propositions. This system must be inconsistent, however, since it is also powerful enough to express the Liar Paradox. On the other hand, informal mathematics is no better off, since it can also express the Liar Paradox.

The mystery to be explained is not the *power* of informal reasoning, but the pragmatic *constraints* on it, which allow contradiction to be avoided most of the time. I expect that the explanation of mathematical truth is not to be found in the Platonic realm, but in a complex interaction of formal structures and mathematical practice (as a psychological and sociological phenomenon).

This is not a view that will be congenial to Penrose's Platonism, but the empirical evidence is against the Platonic realm. First, mathematicians

do not "see" the same mathematical reality. For example, standard analysis, constructive analysis (Bishop 1967) and nonstandard analysis (Robinson 1966) each have their own versions of the real numbers; whether we find noncomputable reals or infinitesmals in the Platonic realm seems to depend on whom we ask. Nor does mathematical truth seem to be changeless, as Penrose asserts (pp. 428, 445-446). If there is anything we would expect to find among the Platonic Forms it is polyhedra, yet Lakatos (1976) documents the long process by which the idea has evolved in interaction with Euler's theorem. There seems to be considerable Becoming in the land of Being.

2 Collapse of the Wavefunction

Remarkably, Penrose claims that understanding the brain will require a physical explanation of the collapse of the wavefunction, and that this will depend on a yet-to-be-discovered theory of quantum gravity. He speculates that "the action of conscious thinking is very much tied up with the resolving out of alternatives that were previously in linear superposition" (p. 438). We will see that he is forced to these conclusions by a dualist interpretation of quantum mechanics.

Quantum mechanics and logical positivism grew up hand in hand; the conventional (Copenhagen) interpretation of quantum mechanics is based on an extreme positivist instrumentalism: a physical system cannot be said to be in a definite state unless it has been observed. In other words, observation causes collapse of the wavefunction. This interpretation is essentially dualist, since it takes observation to be something done *to* the universe by an observer standing outside of it. The trouble is that when we acknowledge that the observer is part of the universe, the physical significance of the reduction of the wavefunction becomes problematic, and we find ourselves confronted with Schrödinger's cat and the like.

There is a simple way out of these paradoxes, but Penrose's dualist biases will not let him accept it. If the observer is part of the physical universe, then the results of observations can be in linear superposition just as the objects are. This of course is Everett's (1957) interpretation, the so-called "many worlds" model – an inaccurate name, for there is only one world: the wavefunction evolving in accord with the unitary operator **U**. Penrose vaguely alludes to the "problems and inadequacies" of Everett's interpreta-

tion (pp. 295-6, 432), but discusses only one. His objection is that we should be "aware" of the linear superposition of observational outcomes, but that we are not. This objection fails to take seriously consciousness as a *physical* phenomenon. If we do so, then we must conclude that conscious states, like other physical states, can exist in linear superposition, but that under normal conditions there is no reason to expect these states to interact. (Presumably we could – at least in principle – design experiments to test for particular superpositions of conscious states.)

In summary, a dualist view of consciousness leads Penrose to reject the Everett interpretation, which forces him to attribute physical reality to the reduction operator **R**, and lands him in need of a new theory of quantum gravity. Dualism comes at a heavy price![1]

3 The Nonalgorithmic Mind

Whether the mind is algorithmic is one of the central questions of Penrose's book, therefore we must consider the meaning of this question. Penrose takes 'algorithmic' to mean 'simulatable by a Turing machine' (p. 47). In this sense the brain is surely algorithmic, since it obeys electrochemical laws, which can be described by a huge system of differential equations, which can be – in principle – simulated by a Turing machine. This shows the irrelevance of this definition of 'algorithmic', since any physical system, including the entire universe, is algorithmic in this sense. Adopting this definition leads Penrose into great difficulties because he has already concluded from considerations of the metamathematical proof that the brain can't be simulatable by a Turing machine. Fortunately, we've seen that that conclusion doesn't follow, and therefore that there is no problem with accepting the brain as Turing-simulatable (at the level of a physical system).

Although we have concluded that the brain is algorithmic (in the sense of being Turing-simulatable), this isn't very interesting since by this standard virtually everything is algorithmic. On the other hand, one of the principal claims of connectionism (against traditional AI and cognitive science) is that the brain is nonalgorithmic. Is there no content to this claim?

The definition of 'algorithmic' that is relevant to these claims is 'a physical system operating by the formal manipulation of discrete symbol structures' (cf. Newell & Simon 1976). An interesting characteristic of this definition

is that it is a matter of degree – some systems are very algorithmic, others are quite nonalgorithmic, and yet others are in between. Connectionist researchers and others have made a good case that the brain is "quite nonalgorithmic," but this will not provide an escape hatch for attributing any special powers to the brain (we have seen that there is no need for them anyway). Rather, by asserting that the brain is nonalgorithmic we make the significant empirical claim that the brain operates on very different principles from a digital computer; whether one can simulate the other is irrelevant to this claim. No doubt the fuzziness of this sense of 'algorithmic' will exclude it from the Platonic realm, but that is often the price we must pay for having a *useful* category.

4 Conclusions

The brain *is* nonalgorithmic, but this doesn't mean that it isn't simulatable by a Turing machine. Rather, it means that formal symbol manipulation is not a fruitful account of its action. Gödel's theorem, and in particular the metamathematical proof, do not imply that the brain has any special powers of inference or insight that are inconsistent with classical physics. Likewise, quantum mechanics does not require any special interaction between minds and the rest of the universe to accomplish reduction of the wavefunction. It is Penrose's dualist biases that drive him to this unnecessarily esoteric account of the mind.

5 NOTES

1. Penrose suggests that quantum mechanics would allow the brain to carry on many simultaneous computations in linear superposition (p. 438). This kind of parallelism does not require quantum mechanics, however; it can be implemented by a variety of classical processes over linear spaces; see MacLennan (1987).

6 References

Bishop, E. (1967) *Foundations of Constructive Analysis*, McGraw-Hill.

Everett, H. (1957) 'Relative State' formulation of quantum mechanics. *Reviews of Modern Physics* 29: 454-62.

Lakatos, I. (1976) *Proofs and refutations: the logic of mathematical discovery.* Cambridge University Press.

MacLennan, B. J. (1987) Technology-independent design of neurocomputers: the universal field computer. *Proceedings of the IEEE First Annual International Conference on Neural Networks* III: 39-49.

Newell, A. & Simon, H. A. (1976) Computer science as empirical enquiry: symbols and search. *Communications of the ACM* 19: 113-126.

Robinson, A. (1966) *Non-standard analysis*, North-Holland.