REVIEW
For Bulletin of the London Mathematical Society, 24 (1992) pp. 87-96

By Aaron Sloman, University of Birmingham

Author: Roger Penrose

Title: The Emperor's New Mind
Concerning Computers, Minds and the Laws of Physics

Publisher: Oxford University Press

Date: 1989 Price: 20.00 pounds Length: 466p

Despite its flaws, this is a stimulating book, ranging over such varied topics as: philosophy of mind, theoretical computer science, artificial intelligence (AI), tiling theory, the Mandelbrot set, philosophy of mathematics (including the analysis of undecidability theorems), the main ideas of classical physics, quantum physics, cosmology (big bang, black holes and all), the nature of time, and neurophysiology. It has already attracted a great deal of attention, in reviews, on radio and television discussions, and lectures given by the author, (partly because there are many people who *want* to believe that AI must fail). The December 1990 issue of the *Behavioural and Brain Sciences* journal, includes a full treatment of the book, including comments by thirty seven reviewers and a reply by the author.

Penrose claims that there are aspects of consciousness that cannot be replicated within any computer model, no matter how sophisticated, as long as the model is based on an algorithm that could, in principle, be a program for a Turing machine. So he has to explain what computation is, produce an (alleged) example of human thought that is not amenable to computational modelling (i.e. following a proof of Godel's incompleteness theorem), and then, since he is no mystic, offer an alternative scientific theory according to which the human brain is not a computer, but is a physical system of a type that embodies super-computational mechanisms, which, unfortunately, can only be understood in terms of as yet unachieved advances in quantum gravity theory. He says (p. 438) "I am speculating that the action of conscious thinking is very much tied up with the resolving out of alternatives that were previously in linear superposition".

This reviewer is not competent to comment on the cosmology, quantum physics, or more abstruse mathematics, except to say that it all makes fascinating, reading, though some sections have to be treated as wishful speculation.

As regards AI, Penrose, like John Searle (1984), attacks what he describes as "the strong AI thesis", which states that there is some (undiscovered) algorithm whose instantiation would produce mental states and processes. Unfortunately, this thesis is so weak as to be hardly worth attacking, and does not relate to work actually done in AI. Being an instance of some (sufficiently complex) algorithm could not suffice for the production of mental states because many static objects and abstract objects that obviously are not minds, including sets of marks on paper and, via Godel-numbering, large numbers, can be construed as instances of algorithms, i.e. as computations. The only precise definition of computation amounts to the specification of an ordered set of structures satisfying certain formal relationships, no matter whether they are produced in time by some causal mechanism, or are abstract static sequences, or patterns of leaves blown in the wind. Thus being an instance of an algorithm is a structural property, satisfiable in all sorts of ways that have nothing to do with minds. So this version of the "Strong Strong" AI thesis is just a straw man. We need a "Weak Strong" AI thesis (Sloman 1986).

Almost as absurd is the the view "that mental activity is simply the carrying out of some well-defined sequence of operations, frequently referred to as an *algorithm"* (Penrose, p. 17) or "the *strong-AI* contention that the mere enaction of an algorithm would evoke consciousness". I don't know anyone who believes this claim. A mind requires many distinct, co-existing, asynchronously causally interacting, states and processes, performing various functions such as detecting information, interpreting it, storing it, reasoning, generating and analysing motives, forming plans, controlling actions, monitoring actions, learning, and many more, to do with feelings and emotions. This is nothing like the enaction of a single algorithm. No AI worker trying to design a complete intelligent robot would try to base it on one algorithm.

Perhaps a network of interacting computers would suffice: but that can't be settled till we have analysed the required functions. Replicating animal minds might need additional non-computational mechanisms, e.g. chemical processes for global control. This is not an objection to the AI programme.

Penrose thinks that "consciousness" refers to some entity "that is, on the one hand, evoked by the material world, and, on the other, can influence it" (page 405). If consciousness were a *thing* then we could ask why it evolved, or what "selective advantage" it confers (page 405), or whether its operation could be explained by quantum mechanisms (see page 399). The problem is that there is no unique thing: the concept is full of muddle and confusion. People feel they have direct insight into the nature of mental states, but this is just an illusion. Our brains include (limited) self-monitoring mechanisms, giving *some* information about internal processes. But no perceptual process, internal or external, gives sufficiently accurate information for scientific purposes. E.g. we don't see the constitution of material objects in the environment, such as clouds or trees. Perceptual mechanisms, whether internal or external, evolved to serve limited practical needs: they often simplify or even distort reality.

Penrose links consciousness with understanding Godel's incompleteness theorem. This is very odd, because many animals (and most people) are conscious without being able to follow Godel's proof. No doubt some mathematicians would like to believe that they have a higher form of mentality than others.

Penrose starts from the fact that for any formal system F rich enough to express the arithmetic of natural numbers, a construction using Godel-numbering will produce an arithmetical formula Pk(k), where the predicate Pk is apparently defined so that it is true of the integer k if and only if there is no proof in F of the formula for which k is the Godel number, which in this case is Pk(k) itself. So if F is consistent there can be no derivation in F of Pk(k) or of its negation. (For more details see p.105-8, or Nagel and Newman 1958). That it is not provable in F is exactly what Pk(k) apparently *asserts.* Therefore what it asserts must be *true.* Hence Penrose can see something to be true which cannot be derived in F even if F is meant to be the formal system defining how Penrose works. Hence no formal system can define how he works, and there is no algorithmic explanation of his thinking.

However the formula says only that a certain very complex number has a very complex arithmetical property. This could be true or could be false. Either way the formula is not derivable in F if F is consistent. But why is Penrose convinced it is *true?* This depends on Godel's mapping, such that (1) the number "k" corresponds to a formula, and (2) the predicate Pk corresponds to a property of that formula. So Pk(k) seems to *assert* something that has been proved (if F is consistent).

But k is, after all, just a numeral: it denotes a number, not a formula. Similarly, Pk is but an arithmetical predicate about numbers and functions on numbers, not a predicate about formulas in F. That we can map it onto an assertion about formulas in F, does not prove that it makes that assertion. In fact, because the formula is neither refutable nor derivable in F there will be models of F in which it is true and models in which it is false. So Penrose can't have 'seen' that it *must* be true. The idea that Pk(k) expresses some definite true proposition about formulas in F is erroneous: it is merely an assertion about numbers, an assertion that has not been proved.

The book is very stimulating, but weak as an attack on AI. The speculations about the relevance of quantum mechanics are unconvincing. Penrose seems to need them only because he has not (yet) seriously tried doing AI.

References

E. Nagel and J.R. Newman *Godel's Proof* Routledge and Kegan Paul Ltd, 1958.

Searle, John, *Minds Brains and Science*, (The Reith lectures) BBC Publications, 1984.

Aaron Sloman, 'Did Searle attack strong strong or weak strong AI' in A.G. Cohn and J.R. Thomas (eds) *Artificial Intelligence and Its Applications*, John Wiley and Sons 1986.

Aaron Sloman, 'Motives Mechanisms Emotions' in *Emotion and Cognition* 1,3, pp 217-234 1987, reprinted in M.A. Boden (ed) *The Philosophy of Artificial Intelligence* "Oxford Readings in Philosophy" Series Oxford University Press, 1990.