Reset reproduction of article published in *Behavioral and Brain Sciences*, Vol. 13, No. 4 (December 1990), p. 661. Reprinted July 1994.

Roger Penrose's *The Emperor's New Mind*

Jon Doyle DOYLE@LCS.MIT.EDU
Massachusetts Institute of Technology, Laboratory for Computer Science
Cambridge, Massachusetts 02139, U.S.A.

Roger Penrose's book offers the reader a valuable perspective on the nature of physical reality and some of its possible implications for AI, computation, and the philosophy of mind. It is worth reading for the survey of physics alone. But the point of the book is to dispute the idea that "our thinking is basically the same as the action of some very complicated computer" by giving two arguments (one from observation, the other from physics) for the claim that "the *conscious* mind cannot work like a computer, even though much of what is actually involved in mental activity might do so" (his emphasis). This brief review confines attention to these two arguments. Though we find that our knowledge of physics and psychology is not yet complete enough to tell whether conscious mental processes are computable, one of the great virtues of this book is that it raises this question technically, clearly, and unavoidably.

Penrose's primary argument is that conscious thought involves seeing or intuiting necessary (mathematical) truths, and mathematical truth is not formalizable, hence it cannot be determined by computers. He claims that mathematicians have direct access to mathematical truth since many mathematicians (myself included) have the distinctive experience of mentally "seeing" mathematical objects laid out as a landscape before them.

Penrose's argument fails to differentiate the ultimate powers of people and machines because the relevant limitation of computers is that they cannot determine *all* mathematical truths, not that they cannot determine *any*. As Penrose admits, however, even the mathematician's conceptual vision is limited: not all truths are visible. Such limitations are not surprising, since most individual mathematical truths could not even be written down using paper the size of the universe and characters the size of protons. Penrose notes that mathematicians can use the method of reflection to resolve particular questions left open by specific formal theories, that is, by observing the results and limitations of the theories. He seems to think that such inferences are not mechanizable. But many of these reflective observations, which are epistemologically similar to observations of objects in the physical environment, can be automated as easily as ordinary deduction rules. (The problem Penrose cites of choosing the right reflections to perform is, as a practical matter, not more difficult than the problem of choosing the right ordinary inferences to draw. Both choices can be difficult.) If we are to suppose that ideal mathematicans can discover recursively enumerable sets of truths derived from finite sets of axioms, axiom schema, and inference rules, including reflection principles, we must suppose that computers can do this too.

Penrose believes humans are not limited to enumerable truths, however, and presents a *reductio ad absurdum* as the crux his argument that mathematical insight is not algorithmic. In short, the assumption that mathematical understanding is captured by some formal system conflicts with our ability to recognize the truth of a Gödel sentence unprovable in

that system. The critical hypothesis of Penrose's argument is that all mathematicians agree on a notion of mathematical truth and that this shared notion of truth does not change as they learn and reflect on proofs. But the only support he provides for this hypothesis is that mathematicians will generally agree on proofs once they learn of them (whether by thought or by communication): "*when* we understand [mathematical demonstrations], their truth is clear and agreed by all" (emphasis added). But this hardly rules out mathematical understanding evolving with new information and experience in universal, even algorithmic, ways. If this is possible (and it is almost an accepted axiom in studies of machine learning), there is no reason to assume that the formal system used to contemplate a Gödel sentence is still the one the sentence is about, and the argument falls apart. Indeed, intuitionist mathematicians contend that changes (not necessarily algorithmic ones) do occur in mathematical understanding, and Penrose's explicit dismissal of their views seems to beg the question. Perhaps the great and clear limitations of human mathematical vision are less limiting than the limitations suffered by computers, but Penrose does not demonstrate this.

Penrose's secondary argument for his thesis is indirect. He argues that thinking is the activity of physical brains, and nothing in the laws of physics as we understand them today ensures that this sort of physical activity is computable. His argument consists of a lengthy but superb survey of the major physical theories in which he points out the numerous ways they do not guarantee computability (or even determinism and locality). Penrose's silence on the topic of relative computability (that is, algorithms over operations other than Turing machine steps) is especially disappointing here, since his ideas suggest attempting to design specific physical mechanisms that realize simple Turing-uncomputable functions (for example, that solve Diophantine equations) for use as "oracles" by digital computers.

One need not accept Penrose's more speculative suggestions about physical reality to realize that there is a real possibility the dynamics of the brain is not computable. But Penrose does not demonstrate that brain dynamics is actually uncomputable. Even if it is uncomputable, he does not demonstrate that this entails uncomputability of any mental processes. The differential equations describing a flip-flop, for example, are probably uncomputable, but the digital computations performed by some systems built from flip-flops are perfectly computable nonetheless.

The book exhibits several minor flaws. Contrary to Penrose's belief, AI *does* employ most of the steps of cognitive processes he identifies as reasonable (including reflection and highly limited forms of "consciousness"), and his argument that human judgment is nonalgorithmic fails to compel because it does not restrict use of "nonalgorithmic" to the technical sense of "no algorithm exists," but mingles this sense with senses involving feasibility, discoverability, and comprehensibility.

Even though his main claim remains unsubstantiated, Penrose deserves our thanks for writing this book about the physical basis of psychology and computation. It is rare for a single book to open so many important questions to technical investigation.