

BOOK REVIEW

APPEARED IN BULLETIN OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 32, Number 3, July 1995, Pages 370-373
©1995 American Mathematical Society
0273-0979/95 \$1.00 + \$.25 per page

Shadows of the mind, by Roger Penrose. Oxford University Press, New York, 1994,
xvi + 457 pp., \$25.00. ISBN 0 19 853978 9

In this book, Roger Penrose claims to provide “a very clear-cut argument for a non-computational ingredient in our conscious thinking” (p. 49).

The background to Penrose’s claim is as follows: In 1961 John Lucas—an Oxford philosopher well known for espousing controversial views—published a paper in which he purported to show that a famous theorem of mathematical logic known as Gödel’s Second Incompleteness Theorem implies that human intelligence cannot be simulated by a computer. Roger Penrose is perhaps the only well-known present-day thinker to be convinced by Lucas’s argument. Lucas himself seems to have seen his argument as showing that the nature of the human mind is mysterious and has little to do with the physics and chemistry of the brain, but Penrose wishes to draw a very different conclusion. The right moral to draw from Lucas’s argument, according to Penrose, is that noncomputational processes do somehow go on in the brain, and we need a new kind of physics to account for them. In *Shadows of the mind*, Penrose not only provides a proof of Gödel’s Theorem and a defense of his own version of Lucas’s argument but also surveys present-day physics, biology, and brain science and speculates daringly as to the nature of the new sort of physics that he claims is needed.

Shadows of the mind will be hailed as a “controversial” book, and it will no doubt sell very well even though it includes explanations of difficult concepts from quantum mechanics and computational science. And yet this reviewer regards its appearance as a sad episode in our current intellectual life. Roger Penrose is the Rouse Ball Professor of Mathematics at Oxford University and has shared the prestigious Wolf Prize in physics with Stephen Hawking, but he is convinced by—and has produced this book as well as the earlier *The emperor’s new mind* to defend—an argument that all experts in mathematical logic have long rejected as fallacious. The fact that the experts all reject Lucas’s infamous argument counts for nothing in Penrose’s eyes. He mistakenly believes that he has a philosophical disagreement with the logical community, when in fact this is a straightforward case of a mathematical fallacy.

Gödel’s Second Incompleteness Theorem states that if a system S of formalized mathematics—that is, a set of axioms and rules so precisely described that a computer could be programmed to check proofs in the system for correctness—is

Copyright 1994 by the New York Times Company. Reprinted by permission.

strong enough for us to do number theory in it, then a certain well-formed statement of the system, one which implies that the system is consistent, cannot be proved within the system. In a popular (but dangerously sloppy) formulation, “If S is consistent, then that fact cannot be proved in S .” The fallacy in Lucas’s original argument is as follows: Lucas confused very different statements which could be called “the statement that S is consistent”. In particular, Lucas confused the ordinary language statement that the methods mathematicians use are consistent with the very complex mathematical statement which would arise if we applied the Gödel Theorem to a hypothetical formalization of those methods. (The latter statement, might, for example, be the statement that a certain primitive recursive function never takes the value zero.) However, Penrose uses a different form of the Gödel Theorem, and his mistake is less technical than Lucas’s.

The structure of Penrose’s argument is as follows: First Penrose provides the reader with a proof of a form of the Gödel Theorem due to Alan Turing, the father of the modern digital computer and the creator of the mathematical subject *recursion theory*, which analyzes what computers can and cannot in principle accomplish. From this proof, using certain not unreasonable assumptions, he concludes that no program that we can know to be sound can simulate all of our human mathematical competence. (Here *know* has a very strong sense: what we *know* has no chance of being false—probabilistic reasoning is not allowed—and we must, in a sense Penrose takes to be intuitively clear, be *aware* that we know what we know. It is reasonable to hold that mathematical knowledge has this character.)

So far, however, what Penrose has shown is quite compatible with the claim that a computer program could in principle successfully simulate our mathematical capacities. The possibility exists that each of the rules that a human mathematician explicitly relies on, or can be rationally persuaded to rely on, can be known to be sound and that the program generates all and only these rules but that the program itself cannot be rendered sufficiently “perspicuous” for us to know that that is what it does. Actual programs sometimes consist of thousands of lines of code, and it can happen that by the time a program has been tinkered with and debugged no one is really able to explain exactly how it works. A program which simulated the brain of an idealized mathematician might well consist of hundreds of thousands (or millions or billions) of lines of code. Imagine it given in the form of a volume the size of the New York telephone book. Then we might not be able to appreciate it in a perfectly conscious way, in the sense of understanding it or of being able to say whether it is plausible or implausible that it should output correct mathematical proofs and only correct mathematical proofs. (Penrose considers only programs which output theorems, but of course a mathematician’s brain outputs proofs as well as theorems.)

What Penrose considers (pp. 131–135) is not such a program but rather a *formal system* whose axioms are provable mathematical statements and whose rules of procedure lead only to mathematical statements that an idealized mathematician could prove. He claims that if any of these rules of procedure were such that we could not know it to be sound, the rule would be “essentially dubious”. It would be strange, he claims (pp. 133–135), if a rule we could not see any justification for always led “somewhat miraculously” to results which we could justify. But this has a shred of plausibility only if the rule in question is simple enough for us to thoroughly understand (and even then it may well be false!), and indeed Penrose limits his discussion at this point (p. 132) to rules which are “simple enough to

appreciate in a perfectly conscious way.”

It was in order to slide over precisely the possibility that our brains follow a program that is not “simple enough for us to appreciate in a perfectly conscious way” that Lucas blurred the question: exactly what consistency statement was he claiming a human mathematician can prove and a machine cannot?

Penrose completely misses this possibility. First, as just said, he discusses the case (his case “II”) in which the program which simulates our mathematical capacity is assumed to be simple enough for us to appreciate in a perfectly conscious way. That such a program might not be provably sound is, as we saw, a possibility Penrose dismisses as “implausible”. Then he considers the possibility (which he also regards as implausible) that the program might be so complex that we could not even write down its description (p. 142: “Here I shall mean that the algorithm is something whose very specification is beyond what can be achieved in practice”). This possibility is rejected because were it actual, then the program of “strong AI”—simulating our intelligence on a computer *in practice*—could not succeed (which is totally irrelevant to the question under discussion, whether *in principle* our brains can be simulated by a computer). In addition, it is argued that if the procedure itself has been formed by a learning algorithm, then that algorithm must either itself be “unknowable” (i.e., such that “even the specification of [its Turing-machine action] would be so complicated that it lies beyond all human activities”, p. 143) or else we are back in the previous case. But there is an obvious lacuna: the possibility of a program which we could write down but which is not “simple enough to appreciate in a perfectly conscious way” is overlooked!¹

We might try simply deleting the restriction of case II to programs which are simple enough to appreciate in a perfectly conscious way. But then the reference to formal systems becomes a red herring. One possible procedure is: “Run the program whose description is . . . (here imagine a telephone book full of code), and if the program outputs a sentence S , write down that sentence as a theorem.” To reject the possibility that such a formal system might simulate the output of a idealized mathematician (as involving something “somewhat miraculous” or “essentially dubious”) is to give no argument at all—and most certainly not “a very clear-cut argument for a non-computational ingredient in our conscious thinking.”

Last but not least, in section 2.9 Mr. Penrose sketches the proof that any program for generating theorems of mathematics is equivalent to a formal system, but the proof involves describing the entire program in a formal language. Thus if the program is too complicated to appreciate in a perfectly conscious way, the resulting formal system will also be too complicated to understand, and there will be nothing “essentially dubious” about the fact that we cannot prove its soundness.

These are the fallacies on which the whole book rests. But apart from the fallacies, there are some interesting questions of a quasiphilosophical rather than a mathematical kind that Penrose does discuss, and here he is interesting (if not always convincing). Is the notion of simulating the performance of a typical mathematician really so clear? Perhaps the question whether it is possible to build a machine that behaves as a typical human mathematician behaves is a meaningful empirical question, but a typical human mathematician makes mistakes. The

¹In the second of his two letters to the *New York Times* (Letters, January 15, 1995), Penrose tacitly admits that this case is not covered by his “possibility II”, but writes that, as I interpret it, it falls under his “possibility III”. Here Penrose forgets that he explicitly restricted “possibility III” to programs such that even writing them down is beyond human capability.

output of an actual mathematician contains inconsistencies (especially if we are to imagine that he or she goes on proving theorems forever, as the application of the Gödel Theorem requires), so the question of proving that the whole of this output is consistent may not even arise. To this, Penrose replies (p. 103) that the mathematician may make errors, but he or she corrects them upon reflection. This is true, but to simulate mathematicians who sometimes change their minds about what they have proved we would need a program which is also allowed to change its mind; there are such programs, but they are not of the kind to which Gödel's Theorem applies.

Again, in his masterwork, *Philosophical investigations*, Wittgenstein emphasized the importance of distinguishing between what an actual machine (or an actual person) can do and what an idealized machine (or an idealized person) can do. Amazingly, Penrose writes, "I am not concerned with what detailed arguments a mathematician might *in practice* be able to follow" (p. 101, emphasis is original). Thus he admits that he is talking about an idealized mathematician, not an actual one. It would be a great feat to discover that a certain program is the one that the brain of an actual mathematician "runs", but it would be quite a different feat to discover that a program is the one that the brain of an idealized mathematician would run. To confuse these questions is to miss the normativity of the question: what is *ideal* mathematics like? Penrose worries that if we say that our (idealized) mathematical output is not describable as the output of a machine whose program we could *know*, then we are saying that something about us ("consciousness") is "scientifically inexplicable", but this is not a reasonable worry. That our norms in this or any other area cannot be reduced to a computer program is hardly a problem for physics.

However, since he believes that it is a problem for physics, Penrose turns to physics; and the second part of the book is a wonderful introduction to modern quantum mechanics and the particular ways in which Penrose thinks it may have to change, as well as to some very interesting speculations about neuronal processes. However, how all this might someday lead to the description of a physically possible brain which could carry out "noncomputational processes" is something that Penrose himself admits the author cannot tell us.

HILARY PUTNAM

HARVARD UNIVERSITY

E-mail address: hputnam@husc.harvard.edu