Lucas revived? An undefended flank

Jeremy Butterfield
Philosophy Faculty, Cambridge University, Cambridge CB3 9DA, England

What a marvellous book! I discern three main ingredients.

1. The best kind of popular science: not just detailed and clear, but also forthcoming about unresolved issues. Setting aside the better-known issues in the foundations of quantum theory, examples include: the distinction between "good" and "bad" uses of Cantor's diagonal argument (p. 111); the recursiveness of the Mandelbrot set (p. 125); self-energy in classical electromagnetism (p. 189); determinism in general relativity (p. 215); complexity theory and quantum computers (pp. 145, 402).

2. Various controversial arguments, mostly against strong AI ("the mind is a digital computer"). The main argument here is based on the nonalgorithmic nature of mathematical insight, allegedly shown by Gödel's theorem (especially pp. 108–12; 417–18).

3. An overarching speculation that two disparate problems – the reconciliation of quantum theory with relativity, and the relation of mind to body – are relevant to one another. This is filled out in various ways. The most striking is by a happy analogy with Penrose's work on tiling and quasicrystals: A thought that surfaces in consciousness is both one of many previously unresolved alternatives (cf. the reduction of the state-vector, and quantum computers), and the solution to a problem, involving global interactions of a characteristically quantum kind, as the growth of a quasicrystal might be (pp. 434–39; 446).

For such a *tour de force,* criticism of specific points is bound to seem niggling. But better that than panegyric. And better that than just scepticism about the speculations: That would be no news to Penrose, who always expresses them cautiously. So I take up two major, and then two minor, points.

First, I am not convinced that Penrose's "Gödel" argument against strong AI avoids the objections against his precursor, John Lucas (1961). (Penrose cites some: I would urge adding Lewis 1969; 1979.) Conscious of these objections, Penrose makes a final attack (pp. 417–18). Transposing the argument to Lucas's terms, it is: If Lucas's arithmetical output is that of a Turing machine, then the machine table must be so complex that Lucas cannot survey it to check that it delivers only truths. (For if he could, then he could "defeat" his own table by constructing its Gödel proposition.) But this is incompatible with the fact that in mathematics "we do not bow down to the authority of some obscure rules that we can never hope to understand. We must see that each step . . . can be reduced to something simple and obvious." (p. 418) Contraposing, Penrose denies that Lucas's arithmetical output is that of a Turing machine. I reply: The "but" is a non sequitur. Unsurveyable complexity of the machine table is, of course, compatible with mathematics' rigourous standards of proof.

Penrose's second argument against strong AI is based on the phenomenon of having "in a flash" a complex thought (pp. 418–23); and his speculation that this is connected to state-vector reduction and quasicrystals. Penrose is mainly concerned with mathematical thoughts. Indeed, he eventually says that he takes the essence of consciousness to be the "seeing" of such a necessary truth as logic and mathematics provide (p. 445). This use of "consciousness," though unusual, would be harmless were it not for the fact that Penrose briefly argues that other phenomena more usually associated with "consciousness" are a threat to strong AI, namely, qualia (pp. 14, 447), personal identity, and indexicality (pp. 27, 409, 448). The brief treatment of these threats engenders two problems. (1) You can get the impression that Penrose's argument involves a unitary notion of consciousness tying all these phenomena together. Not so: As far as I can see, Penrose says nothing against the "divide and rule" idea that "consciousness" is an umbrella term, all these phenomena being logically, indeed nomically, independent. That is, a being could have mathematical thoughts having neither qualia nor indexical thoughts, and so on. (2) Since these threats are much debated in the philosophical literature, Penrose has an undefended flank: Might not the materialist philosophers rebut his argument from the phenomenology of mathematical insight, in much the way they rebut the argument from qualia (e.g., Lewis 1990)?

Two minor points. (1) Whatever consciousness is, it is a non sequitur to infer (p. 408) from its having evolved to its having a selective advantage, and so an active role. It might be a necessary or nomic concomitant of something with such advantage, that imposes no or such little enough disadvantage as the weight of a polar bear's warm coat (cf. Jackson 1982). (2) It is a non sequitur to infer from the timelessness of mathematical truth to there being no threat of causal paradox in the transmission of mathematical beliefs, backward in time (p. 446). Even if the truths are timeless, beliefs in them (and if distinct: their physical correlates in brains) are in time. So such transmission threatens paradox, as backward causation usually does. [See Libet: "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action" *BBS* 8(4) 1985.]

Computing the thinkable

David J. Chalmers
Center for Research on Concepts and Cognition, Indiana University, Bloomington, IN 47405
Electronic mail: *dave@cogsci.indiana.edu*

The main thesis of Penrose's book is that mental processes might be nonalgorithmic. There appear to be three different arguments for this conclusion, which I will present in stripped-down form.

1. The argument from introspection. (1) Some mental processes are not algorithmic at a conscious level, therefore: (2) Some mental processes are not algorithmic.

If this statement of the argument seems a little bald, it is difficult to imagine what else might be meant by the numerous appeals to "intuition" and "judgment" (pp. 411–15; 418–23). It is clear that a premise is missing here. Penrose wishes to exclude from the start the possibility of conscious mental processes that are algorithmic at a level too low to be apparent to conscious introspection. This is a dangerous assumption, as the recent proliferation of connectionist models demonstrates. These models have made familiar the notion that the level at which a system is algorithmic might fall well below the level at which the system carries semantic interpretation (Smolensky 1988). It is not a huge leap to image that in many systems, including the human brain, the computational level might fall below the conscious level.

Connectionist models are not explicitly considered in the book under review, but on the face of it they would seem to fall into the class of "computational" models that Penrose would like to dismiss. It would be interesting to see Penrose declare an explicit position vis-à-vis these models. If he exempts them from his criticisms, then the force of his critique of algorithmic models is considerably weakened; if he wishes to dismiss these models, too, his arguments will need to be considerably strengthened.

It must be conceded that the connectionist approach has not yet had much success in modelling the kind of temporally extended processing, such as mathematical thought, that Penrose considers. Nevertheless, other work within the "subsymbolic paradigm" has made some progress on these matters. In particular, Mitchell and Hofstadter (1990) have produced an interesting model of perception and analogical thought in an abstract domain. In this model, high-level processes emerge from the interaction of a number of small, low-level agents.

Under the influence of various pressures, the model is able to come up with "insights" that are similar in kind to those of a mathematician. The high-level behavior of the model appears in no sense algorithmic, yet it emerges from a completely computational substrate.

2. The argument from Gödel's theorem. (1) Humans can "see" the truth of certain mathematical statements that lie outside the bounds of any given formal system, therefore: (2) Human mathematical thought is not constrained by any given formal system.

This is an interesting variant on the argument of Lucas (1961). Instead of focusing on the formal systems that specify a particular *machine*, Penrose (pp. 416–18) focuses on the formal systems that might specify our mathematical thought. Because we have the ability to "see" that the Gödel sentence for a given system is true, the argument runs, we are using processes outside the system. On page 418, Penrose states: "When we convince ourselves of the validity of Gödel's theorem we not only 'see' it, but by so doing we reveal the very nonalgorithmic nature of the 'seeing' process itself."

This seems fallacious. We do not have to invoke any mystical processes to explain this step; we do not even have to invoke consciousness, as Penrose suggests. The reason we can "see" that Gödel sentences are true is simply that we have a built-in faith that our mathematical systems are *consistent*. It would not be a difficult matter, in principle, to build such faith into an algorithmic machine. (And if Penrose would wish to argue that, unlike machines, humans can repeat the "Gödelization" process *ad infinitum, ad transfinitum*, the reply is that in practice the Church-Kleene result on enumerating constructive ordinals puts as many limitations on humans as it does on machines. We are finite creatures, and we cannot continue to the ultimate Omega.)

To gain his *reductio* of the notion of algorithmic thought, Penrose postulates a single algorithm for determining mathematical truth, shared by the mathematical community. Even to one who believes that mind is algorithmic, this seems a little strange. If we stay within the usual bounds of number theory, analysis and the like, such an idea is perhaps plausible. As soon as we move beyond these into more abstract strata of set theory and logic, disagreement about "truth" becomes rife. Some mathematicians "see" that the axiom of choice is true; others "see" that it is false. Moving further out, the continuum hypothesis and the axiom of constructibility are still more controversial. If such a "universal" algorithm exists, it is a fuzzy thing indeed; it becomes less and less universal the further we travel from the commonplace. This fuzziness alone is enough to defeat Penrose's argument: A fuzzy algorithm cannot be Gödelized!

3. The argument from physical processes. (1) At the lowest level, physical processes might not be algorithmically specifiable. (2) Mental processes are dependent upon physical processes, therefore: (3) Mental processes may be nonalgorithmic.

This is an ambitious argument, but one which must hold if Penrose's other conclusions are to be sustained. It is nothing but an attempt to subvert the force of Church's thesis about the universality of algorithms. There are two clear weak spots. First, even if (1) holds, it would still be far from clear that such microscopic nonalgorithmicity should make any difference on a macroscopic level. It seems plausible to hold that even if *electrons* don't behave algorithmically, *neurons* still might. Penrose acknowledges this gap, but does little to bridge it. Second and more serious, it seems to me that Penrose has in fact provided very little evidence for (1). He gives an impressive demonstration of the nonclassical, nonintuitive nature of microscopic physical phenomena, but he gives no clear justification of why these things should have any bearing on their *algorithmicity*. For example, physical processes may well be nonlocal, but algorithms were never committed to *locality* in the first place. Algorithmic specifications have many degrees of freedom. Although the final verdict will be determined empirically, I doubt that Church's thesis will give in easily.

The idea of algorithmic processing lies at the core of modern cognitive science for good reason. Anyone who succeeds in overthrowing this idea will have effected a deep conceptual revolution in the way we think about the human mind. Penrose has given it his best, and has written a fascinating book along the way, but his arguments are a little thin for the weight they have to bear.

Is mathematical insight algorithmic?

Martin Davis

Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

Electronic mail: *davism@csd11.nyu.edu*

Roger Penrose replies, "No," and bases much of his case on Gödel's incompleteness theorem: It is *insight* that enables us to *see* that the Gödel sentence, undecidable in a give formal system is actually true; how could this *insight* possibly be the result of an algorithm? This seemingly persuasive argument is deeply flawed. To see why will require looking at Gödel's theorem at a somewhat more microscopic level than Penrose permits himself.

It will be helpful (though not essential to our argument) to place the discussion in terms of what is usually called *first order logic*. This is just the formal system that embodies the elementary classical logic of *and, or, not, implies, all, there exists*. In a precise formulation of first order logic, it is necessary to explain when some particular formula F is to be taken to be a *logical consequence* of a set of formulas ("premises") Γ. This can be done in two essentially different ways: *semantically* and *syntactically*. In the semantic version, F is a logical consequence of Γ if F is *true* no matter how the extra-logical symbols appearing in F and Γ are interpreted, so long as all the formulas in Γ are true under that same interpretation. (Metaphorically: F is true in every Platonic world in which the formulas of Γ are true.) In the syntactic version, "rules of proof" involving the straightforward manipulation of symbols are specified, and F is said to be a logical consequence of Γ if F can be obtained from Γ by some finite number of applications of those rules (Penrose, p. 104 gives some samples of such rules). In Gödel's 1929 doctoral dissertation, he establishes his famous *completeness* theorem, which states that the semantic and the syntactic versions are equivalent. Moreover, this equivalence is largely independent of the detailed manner in which rules of proof are specified.

Gödel's completeness theorem answered a question Hilbert had posed in his address at the Bologna mathematical congress of 1928. Hilbert's Entscheidungsproblem for first order logic was also raised in 1928 (in the famous textbook by Hilbert & Ackermann (1928), not at the Bologna conference as Penrose asserts), and called "the fundamental problem of mathematical logic." The problem was to give an algorithm for deciding whether a given formula was a logical consequence (in the semantic sense) of a given (finite) set of premises. Hilbert singled out first order logic for this attention presumably because it seemed clear that all mathematical reasoning could *in principle* be carried out in this formalism.[1] For the premises one takes an *appropriate* set of mathematical axioms; a mathematical theorem is then simply a logical consequence in first order logic of those axioms. Since an argument based on the rules of proof of first order logic can be checked in a completely algorithmic way, we have no trouble understanding why mathematicians should agree about proofs (p. 417) so long as they agree about the axioms (and so long as these axioms are finite in number or at least are specified by an algorithm).

In this context, Gödel's incompleteness theorem (in a strengthened form based on work of J. B. Rosser as well as the solution of Hilbert's tenth problem) may be stated as follows:

There is an algorithm that, given any *consistent* set of axioms,