Murmurs in the Cathedral

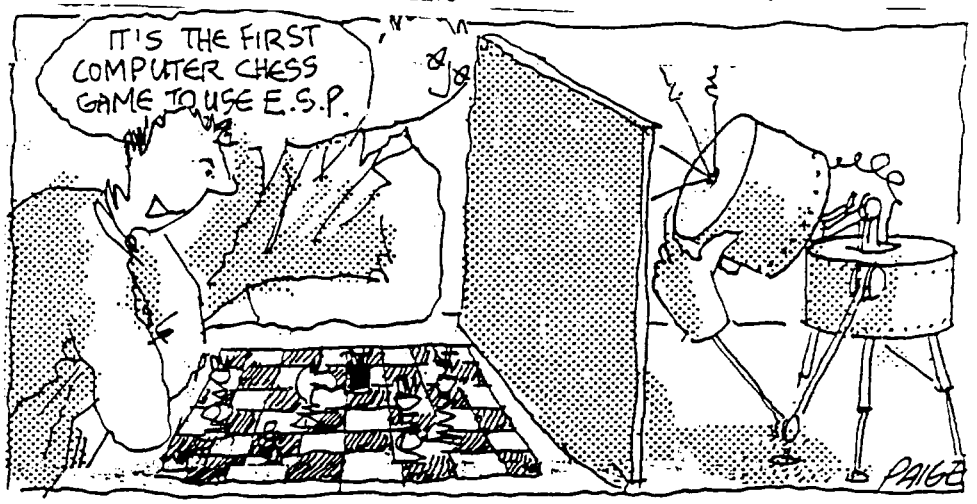
Daniel C. Dennett

ROGER PENROSE
The Emperor's New Mind: Concerning computers, minds, and the laws of physics
356pp. Oxford University Press. £20.
0 19 851973 7

The idea that a computer could be conscious – or equivalently, that human consciousness is the effect of some complex computation mechanically performed by our brains – strikes some scientists and philosophers as beautiful. They find it initially surprising and unsettling, as all beautiful ideas are, but the inevitable culmination of the scientific advances that have gradually demystified and unified the material world. The ideologues of artificial intelligence (AI) have been its most articulate supporters. To others, this idea is deeply repellent: philistine, reductionistic (in some bad sense), as incredible as it is offensive. John Searle's attack on "strong AI" is the best-known expression of this view, but others in the same camp would dearly love to see a principled, scientific argument showing that strong AI is impossible. Roger Penrose has set out to provide just such an argument.

It is a huge project. In order to build his case, Professor Penrose must lead the reader through detailed discussions of many topics in mathematics (Turing machines and computability theory, complex numbers, the Mandelbrot set, Gödel's Theorem, recursive function theory, complexity theory, Platonism versus intuitionism), classical Einsteinian physics and quantum physics, cosmology, and, of course, neuroscience. Most of these topics have been given excellent popular presentations in recent years – in Hofstadter's *Gödel Escher Bach* (1979), Hawking's *A Brief History of Time* (1988), Gleick's *Chaos: Making a new science* (1987) – but Penrose believes that he must go over this material again in his own way, digging deeper, explaining in more detail. The result is bracing reading, to say the least, and the topics for hundreds of pages apparently have nothing to do with the mind at all.

The inevitable first impression, then, is that the book is the ultimate academic shaggy-dog story, a tale whose fascinating digressions outweigh the punch-line by a large factor. Why does Penrose do it? Is there no swifter, more accessible route to his conclusion? No. Penrose sees that he has no hope of overthrowing the case for strong AI unless he can dislodge one of the most imperturbable objects in the intellectual universe: something I will call the Cathedral of Science. This is the highly structured, beautifully articulated amalgam of "what everyone should know" about science, crowned by the inscrutable but talismanic formula, "$e=mc^2$". Its façade, visible to the general public, is popular lore: familiar and decorative items of information and misinformation about the Galilean physics of everyday objects, cartoon-style renderings of black holes and language-using chimps, and pock-marked with such titbits as "you only use five per cent of your brain" and "no two snowflakes are alike". Items in this layer are easily replaced or swept away, but underneath it lies the scientists' (and philosophers') much denser version of the same material, created largely of the remembered oversimplifications of university-level textbooks, supplemented by *New Scientist* and *Scientific American* articles, and such other high-quality interdisciplinary communications as the books just mentioned. This material forms the communally shared understanding on which everyone relies while working on their more particular specialities. Aside from a few brilliant polymaths, the neuroanatomist, the biochemist, the experimental psychologist and the philosopher of mind have roughly the same workaday understanding of quantum mechanics, entropy and computability, for instance, and this understanding gives them sufficient reason to believe that they need not understand these topics any better in order to do their work. The Cathedral's architecture is the familiar hierarchy of mechanistic materialism: living bodies are self-preserving, self-replicating machines made out of cells made out of molecules made out of atoms – with some weird quantum physics isolated (one hopes) in the cellar. No Church has ever enjoyed a more entrenched or authoritative orthodoxy, an empire that expands with daily discoveries and protects itself from swift change by the distributed, mutual myopia of its adherents. Its heresies (ESP, creationism, vitalism) are readily identified and deplored in unison; its conservatism is hailed by almost all who participate in it, for good reason.



It is Penrose's immense task to restructure our vision of the Cathedral of Science, shaking our conviction that it is largely settled and safe and familiar (except, of course, for that baffling business about quantum physics). His task is made all the more intricate by his recognition that most of the Cathedral is sound. He is a revolutionary, but no bomb-throwing nihilist. Like Archimedes, he needs a place to stand if he is to move the world, so he introduces a new taxonomy of theories in science, SUPERB, USEFUL and TENTATIVE, to distinguish what is inviolable and must somehow survive any revolution, from what might be replaced or abandoned. Euclidean geometry, Galilean dynamics, Maxwell's equations, Einstein's special and general relativity theories, quantum physics and quantum electrodynamics are all SUPERB, but even these must be put into a new alignment if we follow Penrose.

Briefly, here is the path of Penrose's proposed revolution. If the brain is a computer, its powers are circumscribed by the limits on all computation uncovered by Turing and Gödel. Turing showed that each possible mechanical computation can be precisely specified by a recipe consisting of a sequence of dead-simple mechanical steps. Such a recipe is called an algorithm; all computer programs are algorithms. Gödel's Theorem showed that no algorithm for proving mathematical truths can prove them all. Doesn't Gödel's Theorem establish that there are tasks "we" (mathematicians, in any event) can perform that are beyond the capabilities of any machine? The idea that the human race can be saved from machinehood by riding on the coat-tails of those clever enough to understand Gödel's Theorem is well-explored territory, and the received wisdom is that all the previous arguments for this conclusion have been roundly defeated, so if Penrose is to get his needed premiss here, he must find a new wrinkle. The standard Cathedral vision is that Gödel's Theorem proves that there is just some single arcane truth of number theory (a machine's Gödel sentence) that is beyond all mechanical computation (by that machine), and Penrose's detailed exposition of the wealth of non-computable but knowable results replaces that vision with an appreciation of the depth and importance of the realm of non-computable mathematics, certainly a domain that is eminently accessible to human mathematicians relying on "insight".

Moreover, the results of complexity theory show that there are many officially computable results that are not *practically* computable – the algorithms that are guaranteed to yield the answer would take billions of years to run on the fastest conceivable computers. How, then, do "we" arrive at solutions to these problems? Penrose proposes that there is a "*theoretical* possibility that a quantum physical device may be able to improve on a Turing machine".

This leads then to a solid review of classical (non-quantum, but relativistic) physics, packed with novel perspectives and designed to impress on us that "we should not be too complacent that the pictures that we have formed at any one time are not to be overturned by some later and deeper view". With

our minds thus stretched open, we plunge into "quantum magic and quantum mystery", and are led yet one more time through the two-slit experiment, Heisenberg's Uncertainty Principle, the collapse of the wavefunction, the Einstein Podolsky Rosen paradox and Schrödinger's notorious cat – in more detail than I have ously encountered in a popular book. [illegible] the upshot is more radical: Penrose doubts whether the puzzles of quantum theory and its relation to classical theory will succumb to any tidy, local resolution, and, like Einstein, he resists the standard "anti-realist" interpretations favoured by most physicists. After a further chapter laying groundwork in cosmology on the flow of time and the curious status of the second law of thermodynamics, we are ready for the suggestion that if a unified theory is to be found, it will have to be a theory of "quantum gravity", requiring "a change in the very framework of the quantum theory". At this time Penrose can present only speculations about the "germ" of such a theory, which is not yet, in his own terms, even TENTATIVE, so he has to settle for some gestures in the direction he feels the revolution will take.

Finally, then, what does all this have to do with minds and brains? He returns to the topics of the early chapters, and resumes the argument that mathematical insight (in particular) is non-algorithmic. Here is where consciousness comes in. The function of consciousness, in Penrose's view, is to leapfrog the limits of (practical) computability by conjuring up appropriate judgments in circumstances in which "enough information is in principle [illegible] able for the relevant judgment to be [illegible], but the process of formulating the appropriate judgment, by extracting what is needed from the morass of data, may be something for which no clear algorithmic process exists – or even where there is one, it may not be a practical one". The way a "quantum computer" would achieve this apparent magic would be by being a sort of super-parallel computer, using superposition of computational states to perform a near-instantaneous global search through an otherwise untraversable space of possibilities, with the solution being output by the collapse of the wavefunction. This would not be old-fashioned Cartesian dualism, but radically new-fashioned (revolutionized) materialism. Several features of what is currently known or believed about connectivity of neural nets in the human brain suggest to him that the brain could in principle be such a quantum computer.

One of the defining doctrines of strong AI is the possibility in principle of teleportation – transporting a *person* from A to B by transmitting a complete, atom-for-atom *description* of the person's body (and brain) and using the description to construct a duplicate at the destination. Is teleportation murder-and-artifice or a means of transportation? Popular science [illegible]n has for years softened us up for the [illegible]r vision, but Penrose is among those who find this idea simply incredible, and one of the cardinal virtues of the quantum computer idea, in his eyes, is that it would rule out perfect duplication of the total quantum state of a brain on what he argues are relatively secure quantum-physical principles.

Where, though, would Penrose have quantum physics draw the line? In principle, could a geranium in a pot be teleported? (When as a child I first heard of "sending flowers by wire" I assumed that they were teleported, and was deeply disappointed to learn the truth.) We already teleport documents (by FAX) and CADCAM (computer-aided-design / computer-aided-manufacture) would readily permit us to teleport automobile parts. Is it all living things, or only all complicated living things, or only all human brains that quantum mechanics would prevent from being thus teleported? Are we the only things in the universe that require quantum computers for their persisting identity?

For those who share Penrose's suspicion about human teleportation, this is one of the comforting implications of his theory, but the price to be paid (in terms of revision of the Cathedral of Science) is high. Among the likely casualties, according to a tentative and impressionistic argument of Penrose, will be the standard neo-Darwinian theory of evolution by natural selection. He does not see how algorithms for mathematical judgment could evolve, and "to my way of thinking, there is still something mysterious about evolution, with its apparent 'groping' towards some future purpose. Things at least *seem* to organize themselves somewhat better than they 'ought' to, just on the basis of blind-chance evolution and natural selection." Creationists are not alone in harbouring deep misgivings about the standard view; there are biologists who dare to wonder about the one leap of faith still required by the standard view: has there really been enough time for evolution to do all the designing by blind methods? Penrose shares those doubts, but provides no new argument to support them.

Might Penrose be right about all this? I suppose he might; he is an extraordinarily inventive and undogmatic thinker with an awesome mastery of many fields. If anyone could make such a discovery, it would have to be someone like Penrose. But whether he is right or not, his strenuous efforts to combat strong AI by unsettling the Cathedral of Science show, more clearly than any of the manifestos for the other side, that strong AI is a straightforward implication of orthodoxy. We cannot simply add a new transept to the Cathedral, enshrining an alternative theory of the mind; if strong AI is mistaken, the whole structure of science must be rebuilt from the ground up. This will inevitably lead some readers to reason as follows: if an opponent as brilliant and dedicated as Penrose discovers he has to go to such lengths to build a presentable case against strong AI, and can come up with nothing stronger than a speculative suggestion about quantum gravity, strong AI must be more secure than I had thought.

The argument Penrose unfolds has more facets than my summary can report, and it is unlikely that such an enterprise would succumb to a single, crashing oversight on the part of its creator – that the argument could be "refuted" by any simple objection. So I am reluctant to credit my observation that Penrose seems to make a fairly elementary error right at the beginning, and at any rate fails to notice or rebut what seems to me to be an obvious objection. Recall that the burden of the first part of the book is to establish that minds are not "algorithmic" – that there is something special that minds can do that cannot be done by algorithm (ie, computer program in the standard, Turing-machine sense). What minds can do, Penrose claims, is to see or judge that certain mathematical propositions are true by "insight" rather than mechanical proof. And Penrose then goes to some length to argue that there could be no algorithm, or at any rate no practical algorithm, for insight.

But this ignores a possibility – an independently plausible possibility – that can be made obvious by a parallel argument. Chess is a finite game (since there are rules for terminating go-nowhere games as draws), so in principle there is an algorithm for either checkmate or a draw, one that follows the brute force procedure of tracing out the immense but finite decision tree for all possible games. This is surely not a practical algorithm, since the tree's branches outnumber the atoms in the universe. Probably there is no practical algorithm for checkmate. And yet programs – algorithms – that achieve checkmate with very impressive reliability in very short periods of time are abundant. The best of them will achieve checkmate almost always against almost any opponent, and the "almost" is sinking fast. You could safely bet your life, for instance, that the best of these programs would *always* beat me. But still there is no *logical* guarantee that the program will achieve checkmate, for it is not an algorithm *for* checkmate, but only an algorithm *for* playing legal chess – one of the many varieties of legal chess that does well in the most demanding environments. The following argument, then, is simply fallacious:

(1) X is superbly capable of achieving checkmate.

(2) There is no (practical) algorithm guaranteed to achieve checkmate.

therefore

(3) X does not owe its power to achieve checkmate to an algorithm.

So even if mathematicians are superb recognizers of mathematical truth, and even if there is no algorithm, practical or otherwise, for recognizing mathematical truth, it does not follow that the power of mathematicians to recognize mathematical truth is not entirely explicable in terms of their brains executing an algorithm. Not an algorithm *for* intuiting mathematical truth – we can suppose that Penrose has proved that there could be no such thing. What would the algorithm be for, then? Most plausibly it would be an algorithm – one of very many – for *trying to stay alive*, an algorithm that, by an extraordinarily convoluted and indirect generation of byproducts, "happened" to be a superb (but not foolproof) recognizer of friends, enemies, food, shelter, harbingers of spring, good arguments – and mathematical truths.

Chess programs, like all "heuristic" algorithms, are designed to take chances, to consider less than all the possibilities, and therein lies their vulnerability-in-principle. There are

many ways of taking chances, utilizing randomness (or just chaos or pseudo-randomness), and the process can be vastly speeded up by looking at many possibilities (and taking many chances) at once, "in parallel". What are the limits on the robustness of vulnerable-in-principle probabilistic algorithms running on a highly parallel architecture such as the human brain? Penrose neglects to provide any argument to show what those limits are, and this is surprising, since this is where most of the attention is focused in artificial intelligence today. Note that it is *not* a question of what the in-principle limits of algorithms are; those are simply irrelevant in a biological setting. To put it provocatively, an algorithm may "happen" to achieve something it cannot be advertised as achieving, and it may "happen" to achieve this 999 times out of 1,000, in jig time. This prowess would fall outside its official limits (since you cannot prove, mathematically, that it will not run for ever without an answer), but it would be prowess you could bet your life on. Mother Nature's creatures do it every day.

I may well have missed a crucial ingredient in Penrose's argument that somehow obviates this criticism, but it is disconcerting that he does not even address the issue, and often writes as if an algorithm could have only the powers it could be proven mathematically to have in the worst case. It will be interesting to see how he would repair this omission. In the meantime I would say that whether or not the Penrose revolution in physics is coming, he has not yet shown the need for the revolution in order to explain facts of human cognitive competence.

I have left no doubt about the difficulty of this book, and I must balance that impression by noting that it is nevertheless a pedagogical *tour de force*, with some dazzling new ways of illuminating the central themes of science. I was struck as never before by the gleeful staircase of human artifices – diagrams, mappings, formalisms – piled one on top of the other over the years, permitting our species so much as to *entertain* such audacious hypotheses about the world we live in. His discussion of phase spaces, for instance, and his development of the rationale for the second law of thermodynamics, are particularly refreshing. His exemplary candour, particularly in the chapters on cosmology and quantum physics, provides the uninitiated reader with a vivid experience of the way gut intuitions and aesthetic reactions call the tune in science *until* someone figures out a conversation-stopping proof, mathematical or experimental.

And along the way he makes important points that have been overlooked by the believers in strong AI, even if they can be incorporated into it. For instance, he closes the book with a speculation about time which I believe is exactly right:

> I suggest that we may actually be going badly wrong when we apply the usual physical rules for *time* when we consider consciousness! My guess is that there is something illusory here . . . and the time of our perceptions does not "really" flow in quite the linear forward-moving way that we perceive it to flow (whatever that might mean!). The temporal ordering that we "appear" to perceive is, I am claiming, something that we impose upon our perceptions in order to make sense of them in relation to the uniform forward time-progression of an external physical reality.

This is, in my opinion, the key to removing the last, harmful vestiges of Cartesian thinking from our standard vision of how consciousness relates to the brain, but you do not need quantum magic or quantum gravity to get there. A clear statement of the point has been given by Douglas Snyder in "On the Time of a Conscious Peripheral Sensation", *Journal of Theoretical Biology*, 1988, 130, 253-254, and I myself have more recently been developing the case for this claim from an entirely conservative – indeed an "engineering" – base, as the best way for Mother Nature to handle the synchronization problems that arise in a brain that must cope with events that sometimes occur on a time scale faster than its own internal transmissions.

A philosophy professor once said to his class, "I want you to believe the things I tell you, but not because you *believe me*; I want you, rather, to believe them because you yourself see that they must be true." This is Penrose's ideal, and indeed it should be every teacher's ideal, but we all fall short; the semester (or life) is too short, and at some point we fall back on "Take it from me: that idea just doesn't work". Penrose is positively heroic in his attempts to live by his standard. The reader is warned, *after* weathering over 200 pages on the lambda calculus, the class of NP-complete problems, Maxwell's equations, the Lorentz equation, special and general relativity, and much more, that in the next chapter, on quantum mechanics, things are going to get "a bit technical. In my descriptions I have tried not to cheat, and we shall have to work a little harder than otherwise." But although matters do then get still more technical, uninitiated readers cannot "work harder" – because we simply do not know the rules. If we are to "see for ourselves" the truths of quantum physics, we must be active and sceptical, but the world of quantum physics is so alien that we can no longer trust our untutored judgments of what counts as a telling objection and what is merely a misapplied maxim or analogy drawn from more familiar territory. I suspect that nothing short of extended immersion in the actual use of the mathematics to solve particular problems can give one a confident sense of how this game is played, and why the rules are what they are. We are assured by Penrose that various hard-to-swallow options make sense while others are just not on, and we have to take his word for it. His brilliant exposition up to this point gives us ample reason to respect his *obiter dicta* once they start to flow, but, contrary to his best intentions, his readers at this point must cease being participants and start being spectators.

This raises a perplexity about Penrose's intentions in writing this book. He repeatedly acknowledges that his colleagues, who already understand the difficult materials he is teaching us much better than we ever will, do not yet accept his idiosyncratic vision. But if he can't convince *them*, by pulling out all the stops, what good will it do if he convinces *us* with a relatively elementary version? What then? Are we supposed to join in a Children's Crusade to persuade his colleagues to get in step? This cannot be his intention.

I suspect he has a more subtle strategy in mind. When experts talk to experts, they are careful to err on the side of *under*-explaining the fundamentals. One risks insulting a fellow-expert if one spells out very basic facts. There is really no socially acceptable way for Penrose to sit his colleagues down and lecture to them about their oversimplified and complacent attitudes about fundamentals. So perhaps educated laypeople are only the presumptive audience for this book, hostages to whom he can seem to be addressing his remarks, so that the experts – his real target audience – can listen in, from the side, without risk of embarrassment. I think this is a wonderful strategy, perhaps the only way of getting experts who are talking past each other to refresh their mutual understanding of the fundamentals. (It is especially valuable in philosophy.) It may leave the non-experts in the role of spectators, but at least it gives them ringside seats.