

The Atheism of the Gaps

Stephen M. Barr

Copyright (c) 1995 *First Things* 57 (November 1995): 50-53.

Shadows of the Mind: A Search for the Missing Science of Consciousness. By Roger Penrose. *Oxford University Press*. 457 pp. \$25.

Roger Penrose, the Rouse Ball Professor of Mathematics at Oxford, has gone a long way toward burying materialism, which is remarkable since Penrose is apparently a materialist himself. He is also a formidable mathematician and mathematical physicist who is perhaps best known for the Hawking-Penrose singularity theorems in General Relativity.

My impression is that the prevailing view in the scientific community is a thoroughgoing materialism which claims that all reality can be understood as the behavior of matter governed by the laws of physics. This is understandable. For several centuries science has cast its net ever wider to bring within its purview everything from the forging of the elements to the processes of life, from the explosions of stars to the visual system of the housefly. Just about everything you see around you is pretty well understood by science-and many things you cannot see. Is it any wonder, then, that given any object of curiosity many scientists view it as simply a matter of time before the physical explanation of it is found? It is this confidence in science, born of long and uninterrupted success, that underlies the claims of the new science of Artificial Intelligence. The human mind is considered by this science to be nothing other than the operation of the physical system called the brain, and the brain to be nothing other than a computer, albeit a very powerful and sophisticated one.

In *Shadows of the Mind*, however, Penrose says, in effect, "Not so fast." The human mind cannot be a computer, if by "computer" one means the computer as currently conceived. There are, he argues, mathematically demonstrable limitations to the kinds of things computers can do, and these limitations are not shared by the human mind. Penrose, as a good materialist, has faith nonetheless that the human mind is nothing other than a machine. But it must be, he endeavors to show, a machine of a radically new and as yet inconceivable kind.

There are two parts, then, to Penrose's argument, and two parts to his book. He argues negatively about what the mind cannot be. And he argues speculatively about what the mind might be.

The first part of Penrose's argument was originally developed, as he admits, in essentially the same form over thirty years ago by another Oxford professor, John Lucas. The Lucas-Penrose argument rests, in turn, on a profound and epoch-making theorem in mathematical logic proven by Kurt Godel in 1930. It is generally recognized that this theorem is as fundamental and revolutionary in its implications as the discoveries of Newton, Einstein, and Heisenberg.

Godel's work in mathematics begins with David Hilbert, who asked in 1900 whether mathematics could be completely "formalized." In such a formalization an exact and unambiguous language would be constructed in which every mathematical proposition would be expressible as a string of symbols, like " $2 + 2 = 4$." Legitimate ways of deriving one proposition from another would be codified in a set of precise rules for manipulating these strings of symbols. To take a simple example, one such rule would allow the addition of the same expression to both sides of an equation, so that if " $a = b$ " is a true proposition, then so is " $a + c = b + c$." If all mathematical reasoning could be so "formalized," then any proof, however deep or intricate, could be reduced to a form in which it could be checked mechanically (literally by a machine).

If this were possible, then computers could "think mathematically" just as human beings do, for mathematics would then be nothing but computation. (Of course Hilbert did not frame the issue in this way, as digital computers had not yet been invented.) What Godel showed, however, and rocked the mathematical world by showing, was that mathematics could *not* be so mechanized. In particular, he demonstrated that if one is given any consistent formal mathematical system rich enough to include ordinary arithmetic, then there exist propositions (called "Godel propositions") that (a) can be properly stated or formulated in the symbolic language of that system, (b) cannot be proven using the mechanical symbolic manipulations of that system, and yet (c) can nevertheless be proven to be true-by going outside the system. Because the human mind can grasp the structure of the formal system and the *meaning* of its symbols, it is able to reason about them in ways that are not codified within that system's rules. As Penrose says:

One might imagine that it would be possible to list all possible obvious steps of reasoning once and for all, so that from then on everything could be reduced to computation-i.e., the mere mechanical manipulation of these obvious steps. What Godel's argument shows is that this is not possible. There is no way of eliminating the need for *new* "obvious" understandings. Thus, mathematical reasoning cannot be reduced to blind computation.

In proving his theorem, Godel showed explicitly how to construct a "Godel proposition" for any particular formal system.

The relevance of all this to computers is that all computers involve- indeed *are*-systems for the mechanical manipulation of strings of symbols (or "bits") carried out according to mechanical recipes called "programs" or "algorithms." Now suppose that there could be a computer program that could perform all the mental feats of which a man is capable. (In fact, such a program must be possible if each of us is in fact a computer.) Given sufficient time to study the structure of that program, a human mathematician (or group of mathematicians) could construct a "Godel proposition" for it, namely a proposition that could not be proven by the program but that was nevertheless true, and-here is the crux of the matter-which could be seen to be true by the human mathematician using a form of reasoning not allowed for in the program. But this is a contradiction, since this hypothetical program was supposed to be able to do anything that the human mind can do.

What follows from all this is that our minds are *not* just computer programs. The Lucas-Penrose argument is much more involved than the bare outline I have just given would suggest, and many people have raised a variety of objections to it. But Lucas and Penrose have had little difficulty in showing the insubstantiality of these objections, and I think it is fair to say that their argument has not been dented. And yet, the argument Lucas and Penrose have made is so disconcerting to certain habits of thought that the reflexive response of many people is to say that it *must* be wrong. Science has conditioned us to expect the breakthrough, the revolution in thought, the astonishing new possibility. To say that machines will never think is as foolish as it was to have said that man would never fly. But science has shown us not only possibilities but limitations. We now know that we will not be able to break the light barrier as we have broken the sound barrier, or chemically transmute lead into gold.

Godel himself seems to have understood quite clearly the implications of his remarkable theorem. (He said the idea that the mind is purely material is a "prejudice of our time.") But he did not draw out these implications in a clear philosophical argument. Lucas did so, but the fact that he was a professed Christian perhaps made it easier for the materialists to dismiss his arguments as the product of wishful thinking. Penrose, however, is that rare thing-a materialist who is not afraid to follow to the very end the implications of his materialism. And those implications are strong and specific.

Penrose establishes with admirable rigor that no machine that works "computationally" can think as we do. He then argues (convincingly) that all machines constructed using the known laws of physics will work computationally. And having assumed that the human mind is nonetheless entirely explicable by the laws of physics, he is forced to conclude that there must be new laws of physics involving processes that are intrinsically non-computational (which is not to say that they are not described by deterministic

mathematical laws). These new laws must be relevant to how the brain functions, and when these laws are understood they will in some way explain how human beings are not only able to compute but to *understand*. This is a very tall order, as Penrose is fully aware. Indeed, at one point, after he has proven that it is an extremely strong kind of "non-computability" that must be involved in these new physical laws, he is driven to exclaim "no doubt there are readers who believe that the last vestige of credibility of my argument has disappeared at this stage." That there are as-yet- undiscovered laws of physics is a near certainty. But scientists generally believe that the laws of physics relevant to describing what is going on at the level of the brain, of neurons, of molecules and atoms, and indeed down to distance scales much smaller than the nuclei of atoms are well understood. And all the laws that we do know are, as Penrose himself argues, essentially "computational." This is not to say his speculations in the second half of his book on how the brain might work are without value. But his materialist assumptions have painted him into a very tight corner.

Something very significant is going on here. Many atheists imagine that religion is based on ignorance. Religion supplies irrational explanations where rational ones are lacking; as lightning, for example, is still thought by primitive people to be the raging of the gods. In this view, religion has been fighting a long rear-guard action against the advance of knowledge, taking refuge in the unknown and the obscure by positing a "God of the gaps," and, as the gaps in our rational explanation of the universe disappear, God will be driven out. This is indeed one of the main motivations for a certain kind of scientist who supposes that when the job of Science is done there will be no room left for the "superstition" of religious belief.

Penrose shows that materialism itself is now the faith of the "gaps." It is in the gaps of undiscovered and unprecedented "non-computational" laws of physics and of uninvented and so-far unimaginable non-computational thinking machines that the "missing science of consciousness" is forced to lurk. But what will happen if the gaps in our knowledge of physics are closed? What will happen if the laws of physics are known in their entirety and turn out not to have the characteristics that Penrose shows they must if they are to explain the mind of man? Then indeed will superstition be overthrown, the superstition of materialism.

Penrose is all the more effective in overthrowing materialism because that is not his aim. He obviously does not take seriously what he calls "mentalism," the view that there is something about mind not reducible to physical description or explanation. In Penrose's understanding, the nonmaterialist holds a view that "regards the mind as something that is entirely inexplicable in scientific terms." He calls it the "viewpoint of the mystic," for it involves a "negation of scientific criteria for the furtherance of knowledge." Moreover, he asks, "if mentality is something quite separate from physicality, then why do our mental selves seem to need our physical brains at all?" "It is quite clear," he goes on, "that differences in mental states can come about from changes in the physical states of the associated brains."

There are several misconceptions here. Materialism does not follow from accepting the scientific method; that something can be studied using the scientific method implies nothing a priori about how it is constituted. We can study both whales and neutrinos using the scientific method, but this implies neither that whales are made up of neutrinos nor neutrinos of whales. That we can study both matter and mind by scientific methods does not imply that the mind is entirely material. And neither do all "mentalists" aver that the mind is "entirely inexplicable in scientific terms." Rather, they say that it is "not entirely explicable in scientific terms"-not the same thing. Moreover, the issue is not whether the brain is *necessary* to our "mental selves," but whether it is *sufficient*. Finally, not all mentalists regard mentality as "quite separate from physicality." "The unity of soul and body is so profound," the *Catechism of the Catholic Church* declares, "that one has to consider the soul to be the 'form' of the body. . . . Spirit and matter, in man, are not two natures united, but rather their union forms a single nature."

What underlies this objection to "mysticism" is the view that religion is a rejection of rational explanations of reality. Mentalism is, Penrose thinks, "unscientific." One of the hallmarks of scientific explanation is that it leads to testable predictions. And some implications of mentalism are not only testable, but have to some extent already been borne out in spite of the fact that they strike the materialist as incredible. A mentalist prediction is that digital computers will never and can never reproduce human intellect, and Penrose claims (correctly, I think) to have proven this very point.

A second prediction of mentalism involves the mysterious question of how the mind and brain are related. Clearly, what is immaterial in the human mind can influence the physical world, or our acts of will and understanding would be without effect. If our will is free these physical effects are not wholly predictable. And yet we know that physical systems are subject to laws that allow their behavior to be predicted. A conclusion of mentalism, then, is that physical systems must undergo change *in two ways*: a not wholly predictable way when minds are involved, and a predictable way governed by physical law. Compare this to Nobel laureate Eugene Wigner's account of what quantum theory implies:

The . . . states of a physical system . . . change in *two ways*[:] . . . continuously, according to Schrodinger's equation . . . and discontinuously according to probability laws if a measurement is carried out on the system [by an observer]. [emphasis added]

In other word, when an "observer"-I would prefer to say "rational mind" or "subject of rational knowledge"-is involved there can be discontinuous and not wholly predictable changes in a physical system, whereas otherwise the system behaves in a way that is completely determined by Schrodinger's equation.

Wigner concludes that while "solipsism may be logically consistent with present quantum mechanics, monism in the sense of materialism is not." Another great physicist, Sir Rudolph Peierls, asserts as an implication of quantum theory that "the premise that you can describe in terms of physics the whole function of a human being . . . including its knowledge and its consciousness is untenable. There is still something missing."

It is true that the traditional interpretation of quantum mechanics espoused by von Neumann, Wigner, and Peierls is rivaled by other views. Nevertheless, this second prediction of mentalism, which had no possibility of fulfillment within classical (i.e., pre-quantum) physics and still strikes many as absurd, is at least arguably a conclusion to which quantum theory leads.

On the other side, it is a prediction of materialism that human beings can have no free will. The behavior of purely physical systems as it has ever been described by science involves only two possibilities: the regularity of deterministic laws or the randomness of stochastic laws. There is no room for the *tertium quid* that is free will. But this prediction of materialism is falsified by the data of our own experience: we actually exercise free will, and thereby know it to be other than either determined or random. As Dr. Johnson observed, while "all theory is against the freedom of the will, all experience is for it." Our experience of exercising free will is as much a fact as any facts about the brain or nervous system. Indeed, it is a more primary fact, for the existence of the brain and its connection with thoughts and mental acts is known only by a long chain of inference (valid though it undoubtedly is), whereas thoughts and mental acts themselves are known directly.

One of the curious things about Penrose is that he is at once a materialist (at least operationally) and a "mathematical Platonist." Like many mathematicians (including Godel), he believes that mathematical truths have a timeless existence apart from human understanding of them. This mathematical Platonism is not necessarily incongruous with Penrose's materialism, since modern physics understands the physical universe in mathematical terms. Nevertheless, he is open to the charge of an interesting inconsistency. For one of his cardinal objections to mentalism is that we have no direct experience of minds subsisting without brains. But it is equally true that we have no direct experience of Ideas (mathematical or otherwise) subsisting without minds. And the Mind that could grasp the full contents of that timeless Platonic world of mathematical truths could certainly not be associated with the timebound operations of any physical brain.

The Penrose argument tells us that mere computers lack something that the human mind possesses. The mentalist position is that that something is spiritual. But granting all this it may still be asked whether computers do not exhibit a certain kind of intelligence, even if they cannot entirely duplicate human intelligence. Is there not evidence of this in the impressive increase in the strength of chess-playing programs like "Deep Thought," for example?

I would distinguish two senses of the word "intelligence." It can mean the ability to solve problems and perform certain kinds of tasks. Or it can mean the capacity for understanding and insight. It is the latter kind of intelligence that alone properly deserves to be called intellect, which humans possess and which computers utterly lack. Figuring out how to solve a new problem or perform a new task generally requires some degree of insight. But once the solution has been routinized it can be carried out in a purely mechanical, mindless way. A small child, for example, might be praised for his intellect if he figures out for himself how to make change correctly, but not so a vending machine. Of course, designing a vending machine requires intellect, but a vending machine itself has none. The same is true of chess-playing programs.

Chess is regarded as the intellectual game par excellence. But it is important to note that in principle chess could be played perfectly without any use of intellect at all. Since chess has a finite though vast number of possible positions, the game can in theory be "solved" by working backwards from the set of all final positions. A complete listing of all positions with the best move (or moves) in each could thus in principle be compiled. At that point an idiot of a machine could play perfect chess by simply looking up the correct moves. Alternatively, because there is an upper limit to how many moves a game of chess can last, a computer of vast though finite memory could calculate to the very end all possible variations from a given position, and thus play perfectly without any use of understanding. I hasten to add that this is not how actual chess-playing programs work. They can only calculate so many moves ahead and then must stop and evaluate the resulting positions. Such "positional judgments" are necessary based upon understanding of the game—the understanding of the programmers and their chess-playing consultants. The computer program itself has no more understanding of what it is doing than does a vending machine. In chess, essentially because of its finitude, computational brawn can substitute for insight. What Godel showed is that in mathematics the same does not hold. As some wit once observed, "A horse that can count to ten is a remarkable horse, but it is not a remarkable mathematician." That is just as true of a computer that can count to a trillion.

Penrose's book is an exceedingly important one, but it will not be an easy one for those not technically inclined to follow. Much more accessible, though probably still difficult for most readers, is the brilliant essay of John Lucas that appeared in *The Modeling of Mind: Computers and Intelligence*. An even more satisfactory treatment, which puts these arguments into the context of a wider attack on materialism and defense of free will, is Lucas' book *The Freedom of the Will* (though this otherwise admirable book is sadly marred by a bizarrely heterodox chapter on God's foreknowledge). The argument made by these two Oxford scholars, so very different in their views of the world, is one that ought to be much more widely known.

It must be admitted that while we have faith that the human mind is understandable, we do not in fact understand it. It is indeed a very profound mystery how spirit and matter are integrated into a single nature in man in such a way as to respect the accuracy and consistency of physical law. But a mystery is not something incomprehensible in itself. It is something uncomprehended by us. Doubtless, further research on the brain will much enlighten us about these issues. Whether it will succeed in dispelling the mystery entirely, only time will tell.

Stephen M. Barr is Associate Professor of Physics at the Bartol Research Institute, University of Delaware