

Shadows of the Mind: A Search for the Missing Science of Consciousness

William Faris

Shadows of the Mind

by Roger Penrose
457 pages, hardcover
Oxford University Press
\$UK 16.99

There was a time when cultured Englishmen would embark on a Grand Tour of Europe, visiting important cities and inspecting monuments and vistas. The book of Penrose is a Grand Tour of Science. The ports of call include the Gödel Incompleteness Theorem, Quantum Theory, Gravitation, Artificial Intelligence, and the Brain. The ultimate destination is the mysterious Isle of Consciousness. According to Penrose these ideas may be linked in a way that points to a grand synthesis.

The first half of the book centers on the Gödel Incompleteness Theorem, which says that for every sufficiently strong formal system there are true sentences that cannot be proved. The fact that mathematicians can understand the implications of this theorem is to be taken as evidence that conscious awareness cannot be computationally simulated. Since Penrose makes much of his case depend on this point, it may be well to review the theorem and its proof. If you, the reader, understand the theorem, then (accord-

ing to Penrose) that differentiates you from any mere algorithm, formal system, computer, or robot. A few minutes of reading and pondering is a small price to pay for this distinction.

Penrose presents a version of the theorem in which the true but unprovable sentence is an assertion that a certain computer program runs on forever. The technical formulation is in terms of an ideal computer known as a Turing machine. A Turing machine is a computer equipped with unlimited memory and with a program for computing some function. Assign a numerical value to an input variable n and start the machine. The machine may eventually give an answer; we say in this case that it halts. It also may just go on computing forever; computer programmers will recognize this as a real possibility.

There is a special kind of Turing machine that is universal; this is the equivalent of a general purpose computer that can run any program. The universal Turing machine has two input variables q and n . Given any Turing machine with one input variable n , there is a value for q so that the universal machine simulates the given machine.

Consider the universal Turing machine in which the input variables have the same values n and n . Let $c(n)$ be the assertion that this machine does not halt. This is a definite mathematical assertion about each natural number n . However, it may be difficult to verify such an assertion—one can run the machine, but what if it runs for a long time without halting? Can one conclude that it will never halt? The difficulty of this problem is the key to this version

William Faris is professor of mathematics at the University of Arizona and Fulbright lecturer at the Mathematics College, Independent University of Moscow. His e-mail address is faris@math.arizona.edu.

[1] of the Gödel Incompleteness Theorem. It says that for every axiomatic system satisfying certain requirements there is a value of k so that the sentence that asserts $c(k)$ is true but not provable.

This axiomatic system is supposed to be expressed in a formalized language. Certain sentences of the language will play a special role; for each number n there is such a sentence $A(n)$. In order for the axioms to be of any use, we need to be able to check whether such a sentence is provable in some systematic way (such as enumerating all possible proofs). This is supposed to be accomplished, at least in principle, by a Turing machine. The machine tries to find a proof. If it succeeds, it halts. If it does not succeed for any reason, it is left to run on forever or it is deliberately thrown into an endless loop. The first requirement on our axiomatic system is that there is such a Turing machine.

Requirement 1: There is a Turing machine such that $A(n)$ provable is equivalent to the halting of the machine with input n .

The role of the sentence $A(n)$ is to attempt to express the assertion $c(n)$. This leads to the second requirement.

Requirement 2: If $A(n)$ is provable, then $c(n)$ is true; that is, the universal Turing machine with inputs n and n does not halt.

This preparation allows us to state the Gödel Incompleteness Theorem. Consider an axiomatized formal system with statements $A(n)$ for which the two requirements are satisfied. There exists a number k such that $c(k)$ is true, yet there is no proof of the corresponding sentence $A(k)$.

The demonstration of this form of the incompleteness theorem is not difficult, and it runs like this. Since the Turing machine is universal, by Requirement 1 there is a number k such that for each n , $A(n)$ having a proof is equivalent to the halting of the computation of the universal Turing machine with inputs k and n . In particular $A(k)$ having a proof is equivalent to the halting of the universal Turing machine with inputs k and k , that is, to the negation of $c(k)$. Suppose $A(k)$ has a proof. Then by Requirement 2, $c(k)$ is true. Hence $A(k)$ has no proof. This is a contradiction. Thus $A(k)$ has no proof, and so also $c(k)$ is true. This concludes the demonstration of the incompleteness theorem.

The theorem says that the axiomatic theory under discussion cannot give a proof of the sentence intended to express $c(k)$. Does the Gödel Incompleteness Theorem itself prove that $c(k)$ is true? Yes, if one can verify Requirements 1 and 2 for the axiomatic system. However, Requirement 2 implies that the system is consistent, that is, that axioms do not lead to a contradiction. So in effect one must verify that the system is con-

sistent. This is not so easy, and that is the content of Gödel's second theorem. This says that if the system is sufficiently strong and is consistent, then it is impossible for the system to produce a proof that it is consistent. Otherwise it would be possible to verify Requirements 1 and 2 within the system, and hence the system would contain a proof of $c(k)$. By the first theorem there is no such proof.

The Penrose argument (which runs over a hundred pages) is intended first to establish that: "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truths." (It also applies to robots: "*no knowable computationally safeguarded mechanisms can encapsulate correct mathematical reasoning.*") It runs something like this. Suppose we were using a knowably sound algorithm. Then there would be a number k corresponding to that algorithm such that $c(k)$ is true but unprovable. However since we know that the algorithm is sound, we can carry out the Gödel argument and prove $c(k)$ after all. This is a contradiction.

Penrose further claims that we do know our mathematical reasoning to be sound, so we cannot be using an algorithm. In his words, "working mathematicians are right in their opinion that they are not merely responding to an unknown (and unknowable) algorithm—nor to an algorithm that they do not firmly believe in." In particular he maintains, "I cannot really see that it is plausible that mathematicians are *really* using an *unsound* formal system F as the basis of their mathematical understanding and beliefs. I hope the reader will indeed agree with me that whether or not such a consideration is *possible*, it is certainly not at all *plausible*."

Some might say that we are using an algorithm, but not the kind of algorithm that is specified by axioms in a formal system. This line is taken by proponents of artificial intelligence that is based on learning from experience. However Penrose wants his argument to place limits on what can be accomplished by all forms of artificial intelligence. He maintains that there is something special about human understanding, perhaps related to the special nature of consciousness. This is a lot to get from the incompleteness theorem. Is there a flaw in Penrose's argument?

One might ask: Why is it supposed to be so plausible that our mathematical reasoning is sound? After all, Gödel's second theorem suggests that we could be reasoning with a complicated algorithm that we only hope is sound. The Turing machine that is supposed never to halt could print out its mocking answer tomorrow—that would shake our complacency. Indeed Penrose does contemplate the possibility

that such an algorithm contains a dubious element R , but he rejects it as unlikely. He doubts that human mathematicians would “see that the truth of [the Gödel sentence] depends precisely upon the soundness of the dubious procedure R which seems miraculously to be able to generate all the [arithmetical] sentences that *can* be unassailably humanly perceived.”

The assumption is that there is a faculty of unassailable perception. Penrose asserts: “In some Platonic sense, the natural numbers seem to be things that have an absolute conceptual existence independent of ourselves. Moreover, the specific *infinite* character of the totality of natural numbers is something that somehow we are able to perceive directly.” He thus reveals himself as a Platonist in the philosophy of mathematics, that is, as one who believes that there is an ideal world of perfect forms distinct from the physical world. He further believes that we can have direct access to this Platonic realm through an “awareness” of mathematical forms. For Penrose this Platonic world includes at least the infinite natural number system. He is not worried, for instance, by the existence of nonstandard models. “The fact is, however, that we actually *know* what the actual natural numbers are and ourselves have no problem about distinguishing them from some strange kind of supernatural number. The natural numbers are the ordinary things that we normally denote by the symbols $0, 1, 2, 3, 4, 5, 6, \dots$. Somehow we find we *know* what a natural number is, once we have been just roughly steered in the right direction!”

One can believe in the objectivity of mathematics without accepting unassailable perception. In any case, there are other views on the philosophy of mathematics. For instance, Lavine [2] develops the theory that our experience with infinite sets is an abstraction of our experience with large finite sets. Nelson [3] takes an extreme formalist position; he does not believe that counting by ones can ever produce a number of the form 2 to a large power. Thus the Platonic position is not self-evident; it requires more justification than Penrose provides.

The remainder of Penrose’s book is devoted to proposals about how the nonalgorithmic awareness might come about. He suggests that

*He suggests
that
appropriate
physical
action of the
brain evokes
awareness,
but the form
of this action
is related to
quantum
physics.*

appropriate physical action of the brain evokes awareness, but the form of this action is related to quantum physics. This opens the way to a discussion of this most peculiar subject.

Quantum theory has been criticized on the ground that “God does not play dice with the universe,” but its status is much worse than that. Probability gives a clear picture of the universe as an ensemble of different possibilities. Quantum theory is much harder to interpret. One can master the abstract mathematics of operators in Hilbert space and still have a very poor idea of how this describes the world. Physicists do learn how it describes the world, but they do this in the course of a long apprenticeship.

Let us review the relation between probability and quantum theory. We begin with a probability example. Consider two mutually exclusive states of some system, say fixed and broken. We may form a *mixture* of these two states by a process of randomization. Let p and q be two probabilities that add up to one. In the mixed state the system is fixed with probability p and is broken with probability q . We have a perfectly clear picture of what this means: there are many cases where the system is fixed and many cases where it is broken, and these occur in the proportions p and q . The possible mixtures can be described by the possible values of p ; they correspond geometrically to a line segment.

Contrast this with the situation in quantum theory. The corresponding notion is that of *superposition*. The possible superpositions of two mutually exclusive states lie on a two-dimensional spherical surface, with the two states at opposite poles. Each pair of states at opposite poles on the sphere corresponds to a pair of incompatible characteristics such as fixed-broken or pretty-ugly or alive-dead. At any time the actual state of the system is represented by some one point on the sphere.

Consider some pair of opposite points, say the one corresponding to fixed and broken, and temporarily think of these as a north pole and a south pole. Let θ be the angle on the sphere from the north pole to the actual state. The probabilities of fixed and broken are then given by $\cos^2(\theta/2)$ and $\sin^2(\theta/2)$. These play the role of p and q . Say instead that one decides to measure whether the system is pretty or ugly. This

indicates a new choice of north and south poles, and there will be corresponding probabilities of pretty and ugly. Notice that the longitude of the actual state with respect to the fixed-broken poles does not affect the probabilities of fixed and broken, but it may well affect the probabilities of pretty and ugly. This extra longitudinal dimension is called the *phase* of the actual state (with respect to the particular choice of north and south pole). The fact that the phase has to be taken into account in making probability predictions is the reason why a superposition is nothing like a mixture.

In physics the characteristics that are measured are more typically something like spin. In the simplest case the spin of a particle along each arbitrarily chosen north-south axis has the two possible values $\pm \frac{1}{2}\hbar$, where Planck's constant \hbar is the unit of angular momentum in quantum mechanics. The spin at the north pole defined by the present state of the system is sure to have the positive sign. Another north-south pair is defined by the choice of measuring apparatus, typically a magnetic field that deflects the particle one way or the other according to the spin value along this axis. Since the actual spin state is a superposition of the spin states at the two poles defined by the measuring apparatus, it gives one deflection or the other, with the appropriate probabilities.

In saying that a superposition is nothing like a mixture, I am cheating slightly. The states of a quantum system form a geometrical space (complex projective space) in which each two points determine a unique sphere (projective line). The simplest possible system is where the entire system is described by just one sphere, as in the case of spin variables with values $\pm \frac{1}{2}\hbar$ along each axis. For this system one can construct a single underlying probability model in which the variables have joint distributions. This is the only such case. See the appendix to [4] for an elementary discussion of this point.

Yet another subtle issue: there is an implicit assumption in the quantum mechanical framework that the probabilities are intrinsic to the system. Without this requirement it is possible after all to construct probability models that

duplicate the results of quantum theory. One needs a family of probability models that depends on the quantity to be measured, that is, on the different possible ways of introducing the measuring apparatus. Such dependence is certainly in accord with Bohr's admonition to take account of the measuring apparatus, but it is not customary to think of quantum mechanics in this way.

In the usual framework there are two ways that the states of a quantum system can change. The change in an isolated system is according to a deterministic mapping of a type that Penrose calls **U** (for "unitary"). This sends states to states in a way that preserves the angles on the spheres. The change in a measured system is according to a very different (and much less understood) random process that he calls **R** (for "reduction"). Some pair of opposing characteristics is singled out to be measured. This corresponds to a particular choice of north and south poles. If the actual state is at an angle θ from the north pole, it is sent to the north pole with probability $\cos^2(\theta/2)$ and to the south pole with probability $\sin^2(\theta/2)$. Thus one characteristic is selected to become actual physical fact. For instance, one could decide to measure along the fixed-broken axis, and in a particular measurement the **R** operation will produce a definite result—say, broken. Or one could decide to measure along the pretty-ugly axis, and the result of this could also be ascertained. One cannot perform the two measurements simultaneously;

their only link is this mysterious relationship of superposition.

The meaning of a quantum superposition is less than intuitive, and the notion of finding our everyday affairs in a superposition of two states seems bizarre indeed. But quantum theory is the fundamental theory of the world. Should not these superpositions also be part of our familiar experience? Quantum theory seems to give us a way out. Consider a system made out of two subsystems. There is a special kind of superposition state in which the behavior of the two subsystems is correlated. This might arise as the result of a **U** interaction with a de-

...even if the universe has random elements, even if it is in some sense an open system, its effect on a human or robot can still be simulated, at least in principle... algorithm is everywhere, except in the human mind.

vice that is supposed to prepare the system for a future measurement. It turns out that if the system is in such a state and the measurement is made on only one of the subsystems, then it cannot distinguish the superposition of the two states from a mixture of the two states. A typical situation is that in which the system is very large. In this case it may be difficult to perform a measurement that does not omit one of the subsystems, and the superposition will look like a mixture. This goes at least part way to explaining why the **R** operation can take place, at least in some practical sense. It does not completely explain **R**, because it gives no clue as to how a random mixture gives rise to a definite outcome in a particular measurement. Of course, the system is in principle not at all a mixture, and so even though the predictions are those of ordinary probability, the picture of the world remains quite obscure.

Quantum mechanics has other puzzling features, such as a kind of nonlocality (or “entanglement”). There is a famous example of a system of two particles, each with spin $\pm\frac{1}{2}\hbar$ along each axis and with total spin zero. If the two particles are widely separated in space and their spins in certain orientations are measured, then the probability of coincidences is larger than can be explained either by signaling or by prior preparation. (See again the appendix to [4].) Penrose presents a beautiful example in which the two particles with total spin zero each have possible spin values $\pm\frac{1}{2}\hbar$ or $\pm\frac{3}{2}\hbar$ along each axis, and the coincidences are perfect. All of this seems so strange that one is forced to remember that the successes of quantum mechanics include explanations of such familiar features of experience as color, chemical bonding, metallic conductivity, and so on.

David Wick’s recent book [4] makes the case that the position of the establishment in physics used to be that there is no problem with quantum theory, or at any rate that the solution has long been known. However, now there is a new establishment position: there is a problem, but it can be fixed. Some of the recent proposals for doing this, including “consistent histories” and “decoherence”, may be found in the book of Omnes [5]. Penrose’s position is that there is a problem and the solution has not been found, but gravitation might play a role. Quantum fluctuations in gravitation may change the structure of space-time, and this new physics may avoid the apparent paradoxes of quantum theory. Scrap the **R** reduction with its seemingly subjective reliance on the notion of measurement; replace it with **OR** (for “objective reduction”) based on the new physical principle.

A physicist always wants to know the magnitude of the effect. Here is Penrose’s guess for

the magnitude of gravitational effects in quantum mechanics. Consider a system of mass m and radius a . Its gravitational energy is

$$E = \frac{Gm^2}{a},$$

where $G = 6.67 \times 10^{-8}$ in cgs units. Now bring in Planck’s constant $\hbar = 1.05 \times 10^{-27}$ erg-seconds from quantum mechanics. This provides a correspondence between energy E and time T given by $T = \hbar/E$. So the gravitational time is

$$T = \frac{\hbar a}{Gm^2}.$$

Penrose interprets this as a “reduction time” for the spontaneous **OR** operation. Fix the density. Then the energy is proportional to a^5 , and the time is proportional to a^{-5} . Take the mass m divided by the volume of a ball of radius a to be one, the density of water. The energy is $E = 1.2 \times 10^{-6} a^5$ ergs, and the reduction time is $T = 9 \times 10^{-22} a^{-5}$ seconds. With $a = 10^{-4}$ centimeters the corresponding time is about a tenth of a second. For smaller a the times are huge; for larger a they are very short. Atomic systems keep their quantum behavior; macroscopic systems rapidly become classical.

How does this tie up with the rest of the story? Penrose would like this hypothetical **OR** to be noncomputable, so that it would help resolve the questions he raised about how humans can transcend computability. As Penrose says, “the complete theory of the putative **OR** process would have to be an *essentially non-computable* scheme.” Penrose wants to go further: “The unity of a single mind can arise, in such a description, only if there is some form of quantum coherence extending across at least an appreciable part of the brain.” And, “On the view that I am tentatively putting forward, consciousness would be some manifestation of this quantum-entangled cytoskeletal state and of its involvement in the interplay (**OR**) between quantum and classical levels of activity.”

Penrose sketches two mechanisms that might relate quantum gravity to noncomputability. One possibility is that the quantum-gravitational state might involve superpositions of all possible geometries at once, and this could mean that nature could solve the topological equivalence problem for four-dimensional manifolds. However, there is no algorithm for solving this problem. The other possibility is even more speculative. “Though it indeed seems reasonable to rule out space-time geometries with closed time-like lines as descriptions of the *classical* universe, a case can be made that they should not be ruled out as potential occurrences that could be involved in a *quantum superposition*.” Penrose attributes this idea to David Deutsch. He points

out that, curiously enough, space-time geometries with closed timelike lines were originally a proposal of Gödel. “If we now consider what it means to perform a quantum computation in such a situation, we apparently come to the conclusion that *noncomputable* operations can be performed. This arises from the fact that in the space-time geometries with closed timelike lines, a Turing-machine operation can feed on its own output, running round indefinitely, if necessary, so that the answer to the question ‘does that computation ever stop’ has an actual influence on the final result of the quantum computation.” Here the reader cries for more detail, but none is given.

What is the reader to make of all this? The book has some of the flavor of a popular exposition, but there is also a definite intent to make a case. In particular, the first half of the book, the part relating Gödel’s Incompleteness Theorem to nonalgorithmic human understanding, or even to consciousness, is relentless philosophical argument. The idea is that an algorithm about which there is doubt cannot be the basis of our mathematical beliefs, because we have at least some sure mathematical beliefs. Unfortunately, its force rests on the prior supposition of a perception of absolute mathematical truth. If one already believes in such perception, then one hardly needs the argument from Gödel’s theorem. While this idea of a Platonic universe of mathematical entities is appealing to many mathematicians, it has serious problems. How do we experience the actual infinite; where do we get our information about it? Mathematics is in some sense an empirical science, about counting and measuring and making marks on pieces of paper (or saving to disk, if you think of it in time). Perhaps human mathematical reasoning is algorithmic, though in this case it is very complicated, and both its functioning and its conclusions are open to doubt. Its validation rests in part on input from the external world. Whether this physical aspect is algorithmic is related to the question of whether the universe is mechanistic, and this we do not know. Penrose says on this question that “it would seem most unlikely that there is something in the non-human environment that more profoundly eludes computation than there is in a human being.” The claim is that even if the universe has random elements, even if it is in some sense an open system, its effect on a human or a robot can still be simulated, at least in principle. He wants to conclude that algorithm is everywhere, except in the human mind. This special role for humans is difficult to accept, at least for those of us who believe that humans are part of nature.

This leads to the second half of the book, which is more tentative. There is a fundamental

mystery in quantum mechanics, and a mechanism for objective reduction (Penrose’s **OR**) would be a stunning discovery. The concept that gravitation might do this is appealing, but it is not worked out in any definitive form, and a major task will be to find predictions that can be checked by experiment. Finally, there is the proposal of a link between quantum mechanics and consciousness. Penrose goes out of his way to admit that the evidence for this is practically nonexistent. As he says, his argument requires that “our brains have somehow contrived to harness the details of a physics that is yet unknown to human physicists.” So we must take it as a vision of a possible future for science. It is hard to see the appeal of this vision. We do not understand either quantum mechanics or the mind, but this does not suggest that one is the solution to the other; most likely each will bring its own surprise.

This book ranges over current issues in an impressive variety of fields and is written in excellent prose. It seems strange that such effort is spent on elaborating proposals with such fragile foundation.

References

- [1] STEPHEN COLE KLEENE, *Introduction to mathematics*, North Holland, Amsterdam, 1952.
- [2] SHAUGHAN LAVINE, *Understanding the infinite*, Harvard University Press, Cambridge, MA, 1994.
- [3] EDWARD NELSON, *Predicative arithmetic*, Princeton University Press, Princeton, NJ, 1986.
- [4] DAVID WICK, *The infamous boundary: Seven decades of controversy in quantum physics*, Birkhäuser, New York, 1995.
- [5] ROLAND OMNES, *The interpretation of quantum mechanics*, Princeton University Press, Princeton, NJ, 1994.