



Book Review

Consciousness: A Red Herring?

The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics

Roger Penrose (Oxford University Press, Cary, N.C., 1989, 480 pp., \$24.95, hardcover)

Astonishingly wide in its scope, this volume is really three books packaged into one. Half is devoted to physics, and the other half is equally divided between computability theory and mind.

"Mind" is the principal *raison d'être* for Penrose's book, in which he seeks answers to the following kinds of questions: Could a mechanical device — a computer, for example — ever be said to think, experience feelings, or have a mind? Can minds exist independent of the biological machinery (brains) with which we find them associated in nature? Are minds subjected to the laws of physics? What, for that matter, are the laws of physics?

We can see why half the book is concerned with elucidating the laws of physics. Computability theory enters the discussion for two reasons: first, to establish that algorithmic processes will not deliver all the mathematical truths (Godel's theorem); and second, to argue the possibility that the "true" physical laws may be deterministic but not computable (thereby accommodating concepts such as free will). Concerning the physical laws, Penrose maintains that a "correct quantum gravity" theory — yet to be discovered — would unify present quantum theory with general relativity theory and, in so doing, would provide a handle for understanding what "consciousness" is and how it arises and functions, which Penrose considers the central problems of mind.

Penrose's analysis of mind begins with a view that (following Searle) he calls "strong AI." According to this view, Penrose writes (on page 17) that "mental activity is simply the carrying out of some [algorithm] ... Most importantly, all mental qualities — thinking, feeling, intelligence, understanding, consciousness — are to be regarded according to this view ... features merely of the algorithms carried out by the brain."

**In a sense,
thermostats are "aware"
of room temperature.
Would we want
to attribute consciousness,
therefore,
to thermostats?**

Searle tried to counter this view by putting forward his "Chinese room" argument. But Penrose does not find this counter argument radical enough and observes (on page 23), "The belief seems to be widespread that, indeed, 'everything is a digital computer.' It is my intention, in this book, to try to show why, and perhaps how, this need not be the case."

To Penrose, the brain is the organ of consciousness. He assumes that consciousness is a scientifically describable "thing," and that this thing actually does "something." What is this thing, then? And what does it do? To answer the second question, Penrose would examine why evolution by natural selection has given rise to consciousness: He asks (on page 405), "What selective advantage does consciousness confer on those who actually possess it?"

Penrose feels that it would not serve any useful purpose, in our current state of ignorance, to explicitly and precisely define what consciousness is. Instead (he writes on page 406), "we can rely to a good measure on our subjective impressions and intuitive common sense as to what the term means and when this property of consciousness is likely to be present." Penrose's attitude seems to imply that "consciousness" is something we can all recognize intuitively — an illusory perception arising because language allows us to make "things" out of "attributes." Being conscious is a behavioral attribute

of an agent. What qualities must behavior have, and what criteria must it satisfy, before we would agree that an agent exhibiting that behavior is conscious? Stated thus, it is not at all clear that we can formulate an unambiguous set of criteria. Clearly, such criteria are likely to be situation dependent. It is even plausible that they would be culture dependent.

Some cultures might consider trees and plants to be conscious beings. Penrose writes (on page 407), "I take the word 'consciousness' to be essentially synonymous with 'awareness.'" This does not help us much, because "awareness" is not any single thing either. In a sense, thermostats are aware of room temperature. Would we want to attribute consciousness, therefore, to thermostats?

Penrose goes on, "There is also the question of what one means by the term 'intelligence' ... I do not believe that true intelligence could actually be present unless accompanied by consciousness ..." Like consciousness, intelligence is a behavioral attribute that is multidimensional. Being intelligent is a task-related attribute, just as being conscious is a situation-related attribute.

Putting aside the definition of consciousness, can we at least characterize what consciousness does? What selective advantage does it confer on those who possess it? Penrose dismisses the possibility that consciousness could be connected with the ability to model the world. He says this seems to be linked to the notion that a system would be "aware" of something if it had a model of that thing within itself, and that it becomes "self-aware" when it has a model of itself within itself. The fallacy of this view, he claims (on page 410), should be obvious from the following fact: "A video-camera has no awareness of the scenes it is recording; nor does a video-camera aimed at a mirror

possess self-awareness." That a physicist should make this statement is astonishing. If you stand before a mirror, what you see in the mirror is not a model of yourself but an image of yourself. A photograph of terrain is not a model (but is an image) of that terrain. On the other hand, a contour map is a model of the terrain. By definition, a model is a theory-based abstraction. The theory tells us how to manipulate the abstraction to derive implications from it. Having a model of something (be it an object, event, or agent) is prerequisite to being able to deal with it rationally. This, after all, is the reason for doing science.

Dismissing the idea of modeling, Penrose argues (on page 411) that "somehow consciousness is needed in order to handle situations where we have to form new judgments, where rules have not been laid down beforehand." For the most part, Penrose's elucidation of "judgment-forming" is specialized to doing mathematics, or to problem-solving contexts. The gist of the argument seems to be that during problem-solving, when we arrive at an impasse, we can generate insightful decisions on the correct move to make. This is an outcome of the act of consciousness, and is an essentially non-algorithmic process. Penrose points out (on page 412), "The judgment-forming that I am claiming is the hallmark of consciousness is itself something that AI people would have no concept of how to program on a computer."

The view that generating new ideas, designs, and solutions is not an algorithmic process is commonplace in AI these days, although Penrose does not seem to be aware of it. The concept of an algorithm in the Turing-machine sense is highly circumscribed. In this sense, an algorithm is a completely deterministic model of a computer program.

One of AI's objectives is to evolve computational formalisms that allow for the controlled nondeterministic behavior of programs. Heuristics, search, and learning are some of the computational techniques that have been evolved to make computer programs behave in ways other than what is possible for fixed Turing machines. Connectionism may be considered an attempt to evolve new programming methodologies that will come to grips with the acquisition and use of tacit knowledge. If we realize that "being creative" is a many-sided attribute of behavior, it is not clear that at least some of its aspects cannot be captured by these evolving AI techniques. It is even less clear that understanding "consciousness" is prerequisite to succeeding in this effort.

Most of the characteristics that Penrose attributes to "consciousness" can really be attributed to the language modality of behavior. Penrose is not unaware of this link's plausibility, but he dismisses it because he thinks language is not needed for thinking. Penrose's arguments in this context are based on an essential confusion in equating "language" with verbalization and the textual mode of representation in terms of words and sentences. For example, Einstein claims (and Penrose quotes him with approval) that in his case "the psychical entities which seem to serve as elements of thought are certain signs and more or less clear images which can be 'voluntarily' reproduced and combined." It is quite unclear whether or not his ability to think in this manner is actually underpinned by his language skills.

A preoccupation with "consciousness" is unlikely to lead to any useful results in understanding "mind." It is more likely that a useful understanding of the brain and mind would arise from viewing the brain as preeminently an organ supporting an organism's integrated behavior, and then studying systematically how the brain is able to achieve this. Human language behavior plays a crucial role in this context. The really deep research issues concern themselves with elucidating this role in computationally tractable ways.

R. Narasimhan
Tata Institute
Bombay, India

