Mind matters—the AIs and the n'AIs

The Emperor's New Mind. Roger Penrose. Oxford University Press, New York, 1989.

The nature of mind and matter, and the relationship between these two aspects of reality, is a perennially intriguing philosophical puzzle. The laws of classical physics imply that the principles governing the material aspects of the world are essentially mechanistic. This realization led to Cartesian dualism—the hypothesis that mind and, particularly, the quintessential yet elusive attribute of mind that we call 'consciousness' belong to a different kind of reality, operating on principles quite different from those of physics; consciousness is a kind of 'ghost in the machine'. Cartesian dualism was attacked by de la Mettrie, in a controversial book *L'Homme Machine*, published in 1748. (It is interesting to note that its publication date belongs to the period during which the art of constructing automata—clockwork mechanisms that mimic the appearance and actions of human beings—reached a high level of sophistication.) De la Mettrie's beliefs were a forerunner of the philosophical position known as materialistic monism—the proposition that consciousness is simply an artifact of mechanistic processes in matter at its most subtle and intricate level of organization. This viewpoint has gained increasing acceptance in recent years, as a result of rapid developments in computer technology and neurophysiology.

Machines already exist that are capable of activities that resemble mental activities. The likelihood of further developments in this kind of artificial intelligence (AI), together with the recognition that the brains of human beings and other animals seem to be, like computers, information-processing devices, points to the conclusion known as the 'strong-AI' hypothesis; that brains *are* computers and, therefore, that it should be possible to develop machines that would be aware of their own existence, and of what they are doing, just as we are. (Remember the character HAL in '2001'?) The realization of this possibility, according to the proponents of strong AI, awaits only further developments in programming technique.

The proponents of strong AI cling to their belief with enthusiasm and conviction. One can detect an almost evangelical fervour in the way they present their case. Roger Penrose's book *The Emperor's New Mind* is a substantial counter-attack, which makes the opposition's arguments appear somewhat simplistic. It is clearly the result of years of thoughtful attention to the many issues involved, from a phenomenally original and creative thinker.

Two major themes run through the book. The first addresses itself to the question: what is it that computers do? The answer is, of course, that they implement algorithms. The implementation of an algorithm is a deterministic process in the sense that the outcome is an inevitable consequence of the input. Note, in particular, that blindly implementing an algorithm is quite different from the act of understanding the meaning encoded in it. (This important point is illustrated by John Searle's famous 'Chinese room' thought experiment.) Having gone into this answer in considerable depth, Penrose then examines the ways in which human thought processes are able to transcend algorithmic behaviour.

The second theme takes us into the realm of modern physics. The strong-AI viewpoint is a position reached by extrapolation from the present state of scientific knowledge. Penrose's contention is that the present state of physical science simply does not warrant this extrapolation. Modern physics contains conceptual inconsistencies at its very foundations; it is very different from the classical (essentially mechanistic) physics that gave rise to materialistic monism. The supporters of strong AI would, presumably, argue that the conceptual difficulties encountered in present-day physics arise in quantum physics and in attempts to reconcile quantum physics with gravitational theory, that quantum physics deals with events at a scale that is *too small*, and that gravitation deals with forces that are *too weak* for such considerations to have any relevance to questions of mental activity. Penrose refutes this by identifying some of the characteristics that a future theory would need to have in order to resolve present conceptual inconsistencies, and then demonstrating that such a unified theory could quite well turn out to have crucial relevance to the clarification of the role of consciousness in mental processes.

In support of these two major themes, Penrose has presented us with magnificently lucid expositions of a great variety of scientific topics, some of which at first sight appear to be only tenuously related to each other and to the central question of the nature of mind. We are led on a fascinating exploration of Turing machines, Gödel's theorem, the nature of mathematical thinking and mathematical discovery, classical physics, quantum physics, relativity (special and general), cosmology, crystallography, neurophysiology and experimental psychology. We are brought to the frontiers of scientific knowledge (and, in the more speculative portions of the book, a bit beyond). All these topics are interlinked and related to the major themes. Penrose has achieved all this, in a delightfully entertaining way, without presupposing any prior specialized knowledge on the part of the reader! The very existence of such a book is itself abundant evidence of its claim that the human brain is something more than an algorithmic machine.

Eric A. Lord

Department of Mathematics
Indian Institute of Science
Bangalore 560 012