Roger Penrose
The Large, the Small and the Human Mind[*]
Cambridge: CUP, 1997, 185 pp., £14.95 / $19.95
ISBN 0-521-56330-5 (hbk.)

The Large, the Small and the Human Mind consists of Roger Penrose's 1995 Tanner Lectures, together with commentaries by Abner Shimony, Nancy Cartwright and Stephen Hawking. Large parts of it are contained in his previous books *The Emperor's New Mind* and *Shadows of the Mind*, and the ideas have been discussed in a number of articles in this journal, so they may be familiar to many readers, although some new material is included. Penrose's general claim is that there are a number of situations that pose problems for current science, whose resolution will demand the development of radically new theories. The kinds of issues he addresses himself to are major ones: for example, the degree of order of the cosmos, the relationship between mathematics and the physical world, whether human mentality can be duplicated by a computer, and whether there is a fundamental connection between quantum mechanics and the human mind.

In the first chapter he uses a formula of Beckenstein and Hawking concerning black holes to argue that the early universe ought to have been highly disordered, a result inconsistent with the observed uniformity of the cosmic background radiation. He rejects the standard explanation for this uniformity based on the inflationary model of the Big Bang claiming it to be invalid, and suggests that a more advanced and as yet unformulated physical theory may resolve the problem. His candidate for this role is quantum gravity, a theory that would combine the two highly successful theories of the twentieth century, Einstein's theory of gravitation and the quantum theory, which have so far defied attempts at integration, into a single whole. This unified theory might explain the observed uniformity by imposing constraints on space–time geometry. But the whole argument hinges on taking the Beckenstein-Hawking formula out of the context in which it was originally demonstrated, and it is far from obvious that it is valid to do this.

Quantum gravity is additionally invoked to explain away the 'Schrödinger's cat paradox'. Under standard quantum theory, paradoxical states are possible in which two seemingly inconsistent possibilities are actualized simultaneously (in Schrödinger's original exposition, a cat being fully alive and at the same time dead). The fact that we do not find such states in reality demands explanation. A number of explanations, or rather perhaps ways of talking around the problem, have been proposed, such as the many-worlds interpretation, the transactional interpretation, the decoherence point of view, and spontaneous collapse caused by additional terms in the Schrödinger equation: but none of these has gained universal assent. Other proposals have implicated consciousness or mind as the agent of wave function collapse, a suggestion due originally to Eugene Wigner, and taken up more recently by Henry Stapp (e.g. Stapp, 1996), which may have some affinities at any rate with the 'Platonic world' aspect of Penrose's exposition which I shall return to shortly, if not with the quantum gravity idea.

Penrose's exploitation of quantum gravity to dispose of possibilities such as a paradoxical 'dead–alive mixture' depends on a very tenuous line of argument. Nature, he supposes, abhors indefiniteness regarding space and time even more than it abhors superpositions of live and dead cats (space and time are more real to Penrose than cats, perhaps?). In Einsteinian gravitational theory, gravity is equivalent to a distortion of space and time, so it is logical to invoke gravitational influences as an agency that can prevent space and time getting unacceptably 'out of step'. The gravitational field of a cat turns out to be strong enough for the task, but unfortunately, as in the case of the earlier problem of accounting for the order of the cosmos, no proper mathematical theory is as yet on offer. Hameroff and Penrose have recently reviewed (Hameroff and Penrose, 1996) their 'Orchestrated Objective Reduction' (Orch OR) approach and made it clear again that space–time is the thing that they consider special, but still without giving reasons to justify this belief sufficient to satisfy people such as this reviewer.

Penrose also discusses the by now familiar hypothesis of Hameroff and himself that microtubules can become integrated into large scale quantum computational systems, perhaps connected with consciousness. Scepticism has been expressed as to the possible existence of such kinds of system (e.g. Scott, 1996), though the

[*] An earlier shorter version of this review appeared in the *Times Higher Education Supplement*.

existence of SQUID rings where the macroscopic and the quantum come into intimate contact as discussed by Srivastava and Widom (1987) — in this type of system, superimposed states of a macroscopic system involving only a few quanta are important in determining the behaviour of the system — perhaps makes such scepticism misplaced. We have here a perhaps familiar territorial pattern, where it would seem that the biologist rejects proposals implying that physics of a kind that he or she does not fully understand may be important. The large amount of research being carried out at the present time on quantum computation, some of which indicates (e.g. Chuang *et al.*, 1995) that this possibility is not necessarily damped out by thermal fluctuations, suggests this is not an idea to be dismissed too readily.

On a more cognitive level, *The Large, the Small and the Human Mind* elaborates an idea touched upon in previous writings, involving three worlds, physical, mental and Platonic, arranged in a triangle that depicts their mutual influence. The mental world, in accord with conventional wisdom, is presumed to be dependent on something physical such as the brain, while the Platonic world, concerned with universal truths such as mathematical truth, is one to which our minds have special access: we are supposed to be able to get at such truths because they are in some sense ‘out there’ for us to get at. In its turn, the Platonic world, in the light of the striking way in which the physical world conforms to mathematical description, is regarded as the source of the physical.

Non-locality (the idea that some influences act at a distance), and non-computability (that some physical processes cannot be computed by a finite computer program) enter into this equation as well. The predictions of quantum mechanics, as shown by John Bell, cannot be accounted for by models where there is no action at a distance while, correspondingly, it seems impossible to localise conscious experience in any one place; and so it is natural to imagine that the two might be connected. And, according to Penrose, mathematical thought cannot be reproduced by a computer program, and so must involve some special physics. Quantum gravity comes in as a candidate for such physics since one attempt to produce such a theory, the Geroch-Hartle scheme, involves classifications which are ‘non-computable’. As an added bonus, the same process which it is suggested might prevent mixtures of dead and alive cats from being realised in nature could prevent our conscious minds being overwhelmed by vast numbers of ideas at the same time.

It requires considerable optimism to speculate that in due course all these ideas will come together, resulting in a theory which will not only unite quantum theory and gravity but also resolve the problem of incompatible combinations of possibilities, as well as including the subtleties of the mind within its scope, although such is the pace of developments in fundamental physics that one cannot rule out such possibilities altogether. However, many scientists see Penrose’s problems as being non-problems, whilst others have proposed alternative ways of overcoming some of the difficulties. It is unclear in any case that one needs to go to quantum gravity to find non-computability. Non-computability in some sense already arises in what are known as chaotic systems, a point I shall return to later.

In any event, there appears to be widespread agreement among specialists in modelling the mind that Penrose’s arguments for the non-computability of mental processes, based on Gödel’s theorem, are misconceived. What these arguments actually disprove is one version of mind model, namely a piece of code that could be run on a computer to give the correct answer to any mathematical question. To treat the question in such terms is to ignore a distinction made many years ago by David Marr, between a theory of how a process is executed, and a corresponding computer simulation. These two entities may differ from each other in various ways, one being that in general a real computer program only approximates to the idealisation to which the theory refers. A theory of how we acquire mathematical skills might be based on the way networks of neurons can be trained to perform particular skills, and might be correct in the limit of an infinitely large neural network, but only approximate for any finite computer simulation. The Gödel-Turing type arguments that Penrose uses presume a system that has both a finite specification and perfect abilities, and thus simply cannot be applied to that kind of situation.

Neither is the attempt to defend Penrose’s ideas against such criticisms — in particular those of Grush and Churchland (1995) — made in a recent article by Penrose and Hameroff (1995) especially convincing. What follows their remark 'it may be helpful to clarify the issue' can be classified as clarification only with difficulty, since the argument can be followed only by looking concurrently at the Penrose and Hameroff defence, the Grush and Churchland article, and *Shadows of the Mind*. In the end, all that Penrose and Hameroff seem able to come up with is a statement asserting in effect that 'correct mathematical reasoning' (to be construed as ordinary mathematical proof, according to the clarification) cannot be encapsulated in any formal system that is 'acceptable to mathematicians as . . . reliable'. But supposing one adopts the point of view that the capacity to perform ordinary mathematical proof is the outcome of a process of familiarisation with particular ideas, where one's initial difficulties with the ideas evolve to familiarity with them, and then to confidence in the correctness in the arguments that others use, till eventually one uses them confidently oneself. With such a scenario in mind one would be in error to consider any formal system which captured the processes one had learnt to accept as valid as being totally reliable. But taking as valid this potentially erroneous step, regarding it as 'correct in principle' (see Penrose, 1994, p.104), seems to be just what Penrose and Hameroff are trying to force upon us.

But there may be another twist to the story. How actually do we come to understand abstract mathematical concepts such as continuity or infinity? Do we come across approximations to them, apply a process of abstraction to them as a result of some brain circuitry, and then learn rules which we become confident at applying, more or less the story suggested above? Or do the approximate realizations point us instead towards their ideal equivalents in a Platonic realm which then becomes the arena of our thinking, as Penrose or Gödel would probably claim? Penrose may be correct after all, but not on the basis of the arguments given in his writings. Would this postulate be just philosophy, or something more? I suggest it would, or could, be science. The Platonic realm would have its own characteristic laws which we could explore and systematically test, and concerning which we could make mathematical theories. Non-computability in the sense of Turing (as utilised by Penrose) may fit naturally into such a scheme, since if the Platonic realm is in some sense eternal then it may instantiate the outcome of 'infinite computations' rather then ordinary finite ones, and in virtue of this feature genuinely provide a source of mathematical truth, to the extent that it can be accessed by human beings.

The rational nature of mathematical thought may make it difficult to rule out explanations for it along conventional lines. Might the case be easier to make for the case of musical intuition? Some years ago I had an opportunity to explore this question in collaboration with an expert on the structure of music (Josephson and Carpenter, 1996). We examined the theories of musical perception proposed by cognitive scientists, and concluded that they addressed themselves only to the more superficial aspects of music, equivalent to the question of grammaticality in language, or that they addressed themselves to musical connotations acquired by association, a model that does not seem to apply to kinds of music that are listened to for themselves rather than being listened to in a particular context of activity. Such theories do not address the more aesthetic aspects of music, or elucidate those features which distinguish a potent 'musical idea' from more random patterns of sound. We argued that these have a significance beyond that which can be explained on cultural or genetic grounds. Perhaps, then, the understanding of these aspects of music is to be found in some kind of transpersonal or Platonic realm where music is a symbolism that is capable of evoking knowledge of an archetypal character (Josephson, 1995).

The non-computability concept discussed by Penrose could also be of value, perhaps in a modified form, even if the arguments presently used to support it are flawed. For example, chaos theory tells us that real systems may sometimes behave in a way that transcends analysis, because we cannot specify their states with sufficient accuracy to be able to do a definitive analysis. Biological systems might be instances of systems of this type. This idea dates back to Niels Bohr, who was later persuaded by Delbruck that it was ridiculous (the familiar territorial pattern again?). The idea is in fact not unreasonable and holds up under close examination (Josephson, 1988). Penrose's concept of a Platonic mind, with capabilities beyond those fitting into conventional models, can be viewed as a contribution to this tradition.

One might conclude from such considerations that Penrose may be right to emphasise creativity, non-computability, and the Platonic realm, but perhaps wrong to look for the integrative factor within his own discipline of quantum gravity. The most crucial element may be creativity. In the physical realm, creativity shows itself only in the minimalist guise of 'random fluctuations'. My collaborator Fotini Pallikari-Viras and I have argued (Josephson & Pallikari-Viras, 1991) that fluctuations have their systematic elements as well as random ones, and that biological organisms may evolve or develop to make creative use of them. It has proved difficult to put such ideas into mathematical form, but this is perhaps an area where significant new concepts can be developed.

Of the three concluding commentaries, that by Shimony is perhaps of the most interest. He argues for a juxtaposition of Whiteheadian philosophy, where mentality and potentiality play a fundamental role, and quantum physics. Physicists have developed, within the ontology of particles and fields, the framework of quantum mechanics which contains abstract concepts such as state, observable, superposition, and entanglement. His proposal is that similar concepts be applied to other kinds of ontology such as those of minds, or entities endowed with a 'protomentality'. This activity might lead to a 'quantum psychology' in which it could be the case that (in line with Stapp's proposals) a developed mentality might resolve the Schrödinger cat paradox. Nancy Cartwright also puts the case for going beyond physics in one's thinking, while Stephen Hawking, once a collaborator of Penrose's, in his commentary 'Objections of an Unashamed Reductionist', makes strong objections to a number of Penrose's claims.

Penrose's books are ones that most readers either considerably like or considerably dislike. Ideas are presented at a great rate, but are very speculative, and justified by tenuous and sometimes doubtful arguments. This book was transcribed from a recording of the original talks, with little attempt being made to improve the clarity of the arguments by rewriting, as a result of which following the arguments will prove a considerable challenge for the non-expert, and possibly problematic even for the expert. Nevertheless, it remains a very interesting and stimulating contribution.

Brian Josephson Cambridge

References

Chuang, I.L., LaFlamme, R., Shor, P.W. and Zurek, W.H. (1995), 'Quantum computers, factoring, and decoherence', *Science*, **270** (5242), pp. 1633–5.

Grush, R. and Churchland, P.S. (1995), 'Gaps in Penrose's toilings', *JCS*, **2** (1), pp. 10–29.

Hameroff, S. and Penrose, R. (1996), 'Conscious events as orchestrated space-time selections', *JCS*, **3** (1), pp. 36–53.

Josephson, B.D. (1988), 'Limits to the universality of quantum mechanics', *Foundations of Physics*, **18**, pp. 1195–204.*

Josephson, B.D and Pallikari-Viras, F. (1991), 'Biological utilisation of quantum nonlocality', *Foundations of Physics*, **21**, pp. 197–207.

Josephson, B.D. (1995), *A Trans-Human Source for Music?* Paper presented at conference on New Directions in Cognitive Science, Saariselka.*

Josephson, B.D. and Carpenter, T. (1996). 'What can music tell us about the nature of the mind? A Platonic Model', in *Toward a Science of Consciousness*, ed. S.R. Hameroff, A.W. Kaszniak and A.C. Scott, (Cambridge, MA: MIT Press).*

Penrose, R. (1994), *Shadows of the Mind* (Oxford: Oxford University Press).

Penrose, R. and Hameroff, S. (1995), 'What Gaps? Reply to Grush and Churchland', *JCS*, **2** (2), pp. 98–111.

Scott, A. (1996), 'On quantum theories of the mind', *JCS*, **3** (5–6), pp. 484–91.

Srivastava, Y. and Widom, A. (1987), 'Quantum electrodynamic processes in electrical engineering circuits', *Physics Reports*, **148** (1), pp. 1–65.

Stapp, H. (1996), 'The hard problem: a quantum approach', *JCS*, **3** (3), pp. 194–210.

* Also available on URL http://www.tcm.phy.cam.ac.uk/~bdj10/files/

Fred Dretske
Naturalizing the Mind

Cambridge, MA and London: The MIT Press, 1995
xvi + 208pp., £18.95, ISBN 0-262-04149-9
(hc/alk.paper)

If the ongoing turmoil in cognitive psychology and philosophy of mind were seen as a kind of philosophical Civil War, one of the most resourceful field marshalls for the forces of externalism would have to be Fred I. Dretske, chairman and professor in the philosophy department of Stanford University, and promoter of a naturalistic theory of mind under the banner of the Representational Thesis: 'All mental facts are representational facts, and all representational facts are facts about informational functions.'

In this volume, based on his 1994 Jean Nicod lectures, Dretske continues the campaign begun in *Explaining Behavior* (MIT Press, 1989), in