BEHAVIORAL AND BRAIN SCIENCES (1990) 13, 643–705
Printed in the United States of America

Précis of *The Emperor's New Mind: Concerning computers, minds, and the laws of physics*

Roger Penrose
University of Oxford, Mathematical Institute, 24-29 St. Giles, Oxford, England OX1 3LB

*Abstract can be found on page 705.

In *The Emperor's New Mind* (1989) [henceforth *Emperor*] I attempt to put forward a point of view (which I believe to be new) concerning the nature of the physics that might underlie conscious thought processes. As part of my argument, I point out that there could well be room, within physical laws, for an action that is *not algorithmic* – i.e., that cannot be properly simulated by any computer – though I argue that it is likely that such nonalgorithmic action can arise only in an area of physics where there is an important gap in our present physical understanding: the no-man's-land between quantum and classical physics. (*Mathematical* processes of a nonalgorithmic kind certainly do exist, but the question I am raising is whether such processes have a role to play in *physics*.) I also argue that there is good evidence that conscious thinking is itself not an algorithmic activity, and that consequently the brain must be making use of nonalgorithmic physical processes in an essential way whenever consciousness comes into play. There must accordingly be aspects of the brain's action that cannot be properly simulated by the action of a computer, in the sense that we understand the term "computer" today.

Thus, the viewpoint I am putting forward dissents both from "strong-AI" (or "functionalism") – as expounded by Minsky (1968) – and also from a frequently argued contrary viewpoint, promoted particularly by Searle (1980). Strong-AI asserts that the brain's action is just that of a computer, conscious perceptions arising as manifestations of the mere carrying out of computations; this contrary viewpoint asserts that although computation does not in itself evoke consciousness, a simulation of the action of the brain would nevertheless be possible in principle, since the brain is a physical system behaving precisely according to some well-defined mathematical action. My dissent from this contrary view stems from the fact that "well-defined mathematical" does not, in itself, imply "computable."

Emperor's scope is broad, for it is my belief that there are many seemingly disparate topics that could have profound relevance to this important question. Moreover, I believe that real progress cannot be made into the deep philosophical issues raised by the question of "mind" without a genuine appreciation of the physical (and mathematical) principles underlying the actual behaviour of the universe in which we find ourselves. I have therefore tried to write this book at a level which makes it accessible, at least in principle, to readers without prior knowledge of the diverse topics covered. These topics include: the basics of artificial intelligence; Turing machines, computability and noncomputability; the Mandelbrot set and the system of complex numbers as illustrations of the Platonic world of mathematics; the foundations of mathematics, Gödel's theorem, nonalgorithmic mathematics; complexity theory; overviews of classical physics (including the issues of computability, determinism, and "chaos" within the theories of Newton, Maxwell, and Einstein) and quantum physics (its basic structure and its puzzles and paradoxes, the two types of evolution **U** and **R**); the second law of thermodynamics and its relation to cosmology, the big bang, black (and white) holes, the "improbability" of the universe, and the Weyl curvature hypothesis; the challenge of quantum gravity and its suggested role in the resolution of the puzzles of quantum theory; structure of the brain and nerve transmission; possible classical and quantum computer models; consciousness and natural selection; the nonalgorithmic nature of mathematical insight, the nature of inspiration, nonverbal thinking; the anthropic principle; a suggested analogy between nonlocal (quasi)crystal growth and the continual changes in brain structure, providing a possible input for (nonalgorithmic?) physics at the quantum-classical borderline; the singular relation between time and conscious perception.

Most of the book is noncontroversial and is intended to provide the reader with an overview of all the relevant topics needed. Though prior knowledge is not assumed, the presentation is of a sufficient depth that some genuine understanding of this material can be obtained. This is not always an easy matter, and parts of the book would need to be studied at some length if the arguments are to be fully grasped. There are places where I present viewpoints that deviate markedly from what might be considered to be "accepted wisdom." I have always been careful to warn the reader whenever I am presenting such unconventional views, even though I may believe the reasons pointing to the necessity of such "unconventionality" to be compelling.

The central questions we must ask are: Are minds

 0140-525X/90 $5.00+.00

subject to the laws of physics? What, indeed, are the laws of physics? Are the laws of physics computable (i.e., algorithmic)? Are (conscious) thought processes computable?

In my own view, mental states would be qualities that indeed depend on those same physical laws that govern inanimate objects. The minds we know of are features of the activity of brains (human brains, at least, but quite probably certain animal brains also) and human brains are part of the physical world. Thus, the study of mind cannot be divorced from the study of physics. Does this mean that new physical understanding is needed, or do we already have sufficient knowledge of the physical laws that might be relevant to an understanding of mental phenomena? Apparently, in the opinion of most philosophers, physicists, psychologists, and neuroscientists, we already know all the physics that might have relevance to these issues. I shall beg to differ.

It would, of course, be generally admitted by physicists that there are still many gaps in our understanding of the principles that underlie the behaviour of our universe. We do not know the laws that govern the mass-values or the strengths of the interactions for the menagerie of fundamental particles that are known to exist. We do not know how to make quantum theory fully consistent with Einstein's special theory of relativity, and certainly not with his general relativity theory. As a consequence of the latter, we do not understand the nature of space at a scale 10^{-20} times smaller than the dimension of the known subatomic particles, though our knowledge is presumed adequate at larger scales. We do not know whether the universe is finite or infinite, either spatially or temporally. We do not understand the physics that must operate at the cores of black holes or at the big-bang origin of the universe itself. Yet all these issues seem quite remote from what is relevant for understanding the workings of a human brain.

I shall argue, nevertheless, that many of these matters are of some relevance and, moreover, that there is *another* vast unknown in our physical understanding at just such a level as could indeed be important for the operation of human thought and consciousness – just in front of (or rather behind) our very noses! I believe there is a fundamental gap in our physical understanding between the small-scale "quantum level" (which includes the behaviour of molecules, atoms, and subatomic particles) and the larger "classical level" (of macroscopic objects, such as cricket balls or baseballs). That such a gap exists is a matter of physics – unrelated, at least in the first instance, to the question of "minds." I argue the case that there is good reason, on purely physical grounds, to believe that some fundamentally new understanding is indeed needed here – and I make some suggestions concerning the area of physics (the "collapse of the wave function" as an objective "quantum gravity" effect) wherein I believe this new understanding is to be sought. This whole issue is a matter of dispute amongst physicists at the moment; and it would have to be admitted that the apparent majority view is that no new theory is needed (although such outstanding figures as Einstein, Schrödinger, de Broglie, Dirac, Bohm, and Bell have all expressed the need for a new theory). It is my own strong belief that a radical new theory is indeed needed, and I am suggesting, moreover, that this theory, when it is found, will be of an essentially noncomputational character.

I argue that some of the manifestations of consciousness are demonstrably nonalgorithmic, and I am therefore proposing that conscious mental phenomena must actually depend upon such noncomputational physics. Although I try to make a fairly specific suggestion as to where in brain action the role of the physics governing the no-man's-land between quantum and classical physics might lie, my arguments are not critically dependent upon this particular suggestion. There might well be other possible roles for such (nonalgorithmic) physics in brain action. According to my viewpoint, the outward manifestations of the phenomenon of consciousness could never even be properly *simulated* by mere computation. Any properly "intelligent machine" – and I would argue that for it to be properly "intelligent" it is necessary for it to be conscious – could not be a "computer" in the sense in which we understand that term today (and so, in my own terminology, I prefer not to call such a putative object a "machine" at all), but would itself have to be able to harness the very nonalgorithmic physics that I am arguing must be a necessary ingredient of the physical basis of conscious thought. At present we totally lack the physical understanding to be able to construct such a putative "machine," even in principle.

Chapter 1: Can a computer have a mind?

Recall the Turing test, according to which a computer, together with some human volunteer, are both hidden from the view of some (perceptive) interrogator. The interrogator has to try to decide which of the two is the human being and which is the computer merely by putting probing questions to each of them, the answers being all transmitted in an impersonal fashion: say typed on a keyboard and displayed on a screen. The human subject answers the questions truthfully and tries to persuade the interrogator that he is indeed the human being, but the computer is allowed to "lie" so as to try to convince the interrogator that it, instead, is the human being. If on a series of such tests the interrogator is unable to identify the real human subject in any consistent way then the computer (or the computer's program) is deemed to have passed the test.

Setting aside, for the moment, the issue of whether, or when, some computer might actually pass the Turing test, let us suppose, for the sake of argument, that such machines have already been constructed. We must address the operationalist's claim that such a computer must be said to think, feel, understand, *et cetera*, merely by virtue of its passing. It is certainty my own view that such mental qualities – and certainly the central one of consciousness – are objective physical attributes that an entity may or may not possess. In our Turing-test probing, we are merely doing our best to ascertain, using the only means available to us, whether the entity in question (in this case the computer) has the physical attributes under consideration (in this case, say, consciousness). To my mind, the situation is not different in principle from, for example, an astronomer trying to ascertain the mass of a distant star. Being able to give human-like answers to Turing-test questioning is certainly not the same as hav-

ing human-type mental qualities, but it may serve as a good indication that such mental qualities are indeed present in the machine. This would be in the absence of other criteria which might have a greater scientific reliability. In the case of mental phenomena, we do not really have better criteria than "communication" in some form, as with the Turing test. If someday we have a better theory of consciousness, then better criteria might eventually become available, and the Turing test will lose some of its importance. (Compare this with the possibility of accurately observing a planet in close orbit about our distant star. Einstein's theory can then provide a definite measurement of the star's mass.) In the absence of such a theory, however, the case for regarding the successful passing of a Turing test as a valid indication of the presence of thought, intelligence, understanding, or consciousness, and so forth, is fairly strong. For conversation is how we normally form our judgements that people other than ourselves actually do possess consciousness.

How long will it be before a computer actually passes the Turing test? It really depends on how strict one's criteria are for passing. It is my own guess that a proper passing of the test could not be achieved at all by an algorithmic computer (i.e., by a computer based on the calculational principles that we use today) – at least not in the foreseeable future. I have been vague about what constitutes a "proper" passing of the test, but extensive to-and-fro probing by the interrogator would certainly be required. (Turing originally suggested that a 30% success rate for the computer, with an "average" interrogator and just five minutes questioning, might be achieved by the year 2000.) In some clearly delineated fields, however, very impressive "human-like" behaviour has already been achieved. Chess-playing computers provide good examples of machines exhibiting what can be thought of as "intelligent behaviour." "Deep Thought" (programmed largely by Hsiung Hsu) has achieved some notable victories in some games with grandmasters. I think that it is clear, however, that computers play chess very differently from the way human beings do; computers relying much more on depth of extensive calculation and much less on "intuitive judgements" – whatever *they* are!

Perhaps a little more directly along the lines of a Turing test would be the computer programs of Schank and Abelson, as cited by John Searle (1980) in connection with his Chinese room argument. The computer using such a program is able to answer simple questions about stories (concerning, say, people eating in restaurants) in a way that a human being might. This seems to suggest that there is some primitive understanding on the part of the computer, since a human being would require some understanding in order to answer the questions correctly. Searle argues that no actual understanding can be taking place in the computer and that this limited "passing" of a Turing test is accordingly no indication that the mental quality of "understanding" actually occurs. To make his case, Searle envisages a "gedanken experiment" whereby the stories (about restaurants), the questions about these stories, and the computer's answers are all written in Chinese. The actual rules the computer follows in forming its replies to these questions are then laboriously followed by Searle himself (in a locked room), the rules for manipulating the relevant Chinese characters according to the computer's program being presented to Searle in English. Though Searle sits in this room and follows all the relevant actions of the computer, finally providing the correct (Chinese) answers to the questions, he does not himself obtain the necessary understanding of what the stories are actually about, since he knows nothing of the Chinese language.

Searle's argument is directed against the standpoint of "strong AI" (or functionalism) which claims that it is merely the enaction of a (sufficiently elaborate) algorithm that evokes mental qualities such as consciousness, understanding, or intentionality. I regard Searle's argument as quite persuasive with regard to programs of the fairly limited complication of the ones envisaged above, but it is by no means conclusive – especially when it is applied to the immensely more complicated putative computer programs that might, according to the strong-AI view, be necessary to conjure up actual consciousness! Perhaps there is some kind of "critical" amount of complication necessary for consciousness to be genuinely conjured up? Moreover, it is not a logical necessity that an algorithm's putative "awareness" should necessarily impinge upon the awareness of a person carrying out the algorithm, irrespective of the complication of that algorithm.

Searle is prepared to accept that a computer simulation of the actual activity of a human brain – when it is in the process of evoking conscious mental perceptions – would be possible in principle. But he believes that such a simulation would not, in itself, evoke the same mental perceptions. That certainly seems to me to be a tenable position, but not a very helpful one for gaining any understanding of why some objects enacting algorithms (brains) seem to evoke such mental qualities whereas others (e.g., electronic computers) are not supposed to. In later chapters, I present a different kind of argument to suggest that conscious mental activity is not algorithmic at all, and such a simulation would therefore *not* be possible.

The viewpoint of strong AI appears to spring partly from the fact that a person's individuality does not depend on the particular atoms that compose his body. There is certainly a continual turnover in virtually all the material of any living person's body. Moreover, according to quantum mechanics any two particles of any particular kind are completely identical to one another. For example, if the entire material content of a person were exchanged with corresponding particles in the bricks of his house then nothing would have happened whatsoever. What appears to distinguish the person from his house is the pattern in which his constituents are arranged, not the individuality of the constituents themselves. Thus, a person is just a "pattern of information," and this information could, in principle, be translated from one material form into another. The strong-AI viewpoint is that it is simply the "information content" of this pattern that characterizes any particular individual. This idea has gained strength from experience with modern high-speed computers, where we are now very familiar with the phenomenon of information being transformed from one realization into another (say from a pattern of magnetic fields in a floppy disc to a collection of charge displacements in a computer memory, and

from there to a family of gaps in the illumination of a cathode ray screen). Moreover, there is the theoretical justification from the fact that modern general purpose computers are all, in effect, examples of *universal Turing machines* (see next section). Any two such machines are completely equivalent to one another in the sense that, irrespective of the particular hardware that may be involved in each machine, there will always be some appropriate software that effectively converts either machine into the other. In this sense, the hardware is regarded as being "irrelevant," and the essential "information" of the operation of the machine is considered to lie in its program, that is, software.

In mathematical terms, the program is effectively what is called an algorithm, that is, a step-by-step mechanical procedure for working towards an answer from any given input data. According to the strong-AI viewpoint, it is the very enacting of an appropriate algorithm that can (or must?) evoke (or be?) conscious awareness. Whether or not one believes in its essential correctness, however, there is much that is unclear about this picture. What does it actually mean to implement an algorithm, or to embody it in physical form? Do all the individual steps have to be performed in full by moving about bits of matter, or might a static printed description of the procedures be sufficient in itself to evoke conscious awareness? If the nature of the material performing the algorithm is irrelevant, then why is it necessary to have the material there at all? Should the algorithm's very (timeless!) Platonic existence not be sufficient to evoke consciousness? If it is, where does the distinctly temporal quality of consciousness come from? None of these matters seem to me to have been properly addressed by the proponents of strong AI.

An idea frequently discussed in this context is the teleportation machine of science fiction. A would-be traveller is scanned from head to toe, the accurate location and complete specification of every atomic nucleus and every electron in the subject's body being recorded in full detail. All this information is then beamed (at the speed of light), by an electromagnetic signal, to the distant planet of intended destination. There, the information is collected and used as the instructions to assemble a precise duplicate of the traveller – together with all his memories, his intentions, his hopes, and his deepest feelings, where the original copy of the traveller is to be destroyed in the process. Could a teleportation machine work – in the sense that the traveller's actual conscious identity is transferred to the distant planet? If teleportation is not travelling, then what is the difference in principle between it and just walking from one room into another? In the latter case, the traveller's atoms of one moment would simply be providing the information for the locations of his atoms of the next moment, and we have seen, after all, that there is no significance in preserving the identity of any particular atom. The moving pattern of atoms simply constitutes a kind of wave of information propagating from one place to another, so where is the essential difference between this and the teleportation device?

I believe that despite the outlandish nature of the teleportation idea, there is perhaps something of significance concerning the physical nature of consciousness and individuality to be gained from it. It may provide one pointer, indicating a certain essential role for *quantum mechanics* in the understanding of mental phenomena, for it turns out to be impossible to copy a quantum state unless the information in the original state is destroyed. If a person's individual awareness depends on some essential aspect of a quantum state, then it would not be possible to "teleport" that awareness unless the original copy of the individual were indeed destroyed. Accordingly, teleportation would not be impossible in principle; but these considerations would rule out the more paradoxical possibility of having two or more viable conscious copies of one individual.

Chapter 2: Algorithms and Turing machines

One of the most important developments in the history of mathematics occurred in the late 1920s and 1930s, when the concept of a general algorithm was made mathematically precise and it was demonstrated that there are some mathematical procedures that cannot be described by any algorithm whatsoever. There are various different but completely equivalent ways of formalizing the algorithm concept. The one which is intuitively the clearest was first put forward in 1935 by Alan Turing, called a *Turing machine:* an idealized device in that it operates with a potentially infinite "tape" on which the input data and instructions are represented. Only a finite portion of the tape is to be marked with actual data or instructions, however. Thus, although the device can cope with an input of unlimited size, this input is always finite. We may take the input and output to be recorded on the same tape; all the input goes in on the right and the output finally goes out on the left. For simplicity, we regard the tape as marked just with the symbols 0 and 1; "0" stands for blank tape and "1" for a mark on the tape (any more complicated symbols we might wish to use are broken down into a sequence of "marks" and "blanks"). The device itself has a finite number of distinct "internal states" – coded, for convenience, by binary numbers 0, 1, 10, 11, 100, 101, etc. – and it reads the symbols on the tape one at a time. It has a finite list of instructions that tell it what to do, given its internal state and whether it reads 0 or 1. The list of instructions itself is never altered, but the particular 0 or 1 that is being read by the device may either be left alone or else changed to the opposite symbol, depending upon the particular instruction that comes into play; the device also accordingly moves one step to the right, one step to the left, or comes to a halt after moving one step to the right. (In these descriptions I am taking the device to be moving rather than the tape, but the two pictures would be equivalent.) As part of the same operation, the internal state is changed to another internal state, as specified by the instruction. Now, in this new internal state, the device (if it has not come to a halt) reads a new 0 or 1 on the tape and acts according to the appropriate instruction that now comes into play. The process continues until a halt instruction is encountered; at that point we imagine a bell ringing to alert the operator of the machine that the calculation has been completed.

It is a remarkable fact that any computational process whatever (that operates with finite discrete quantities) can be described as the action of some Turing machine.

This, at least, is the contention of the so-called Church-Turing thesis, in its original mathematical form. Support for this thesis comes partly from Turing's careful analysis of the kinds of operation one would actually consider as constituting a computational or algorithmic process, and partly from the striking fact that all the various alternative proposals for what an "algorithm" should mean (put forward at around the same time by Church, Kleene, Gödel, Post, and others) have turned out to be completely equivalent to one another. Some of these proposals had the initial appearance of being completely different, so their equivalence is a strong indication of the fact that they are merely alternative ways of describing an absolute abstract mathematical concept, that of *computability*, (which is independent of any particular realization of it that one may care to adopt. (In addition to an extended and detailed description of Turing machines, I give a brief description of Church's remarkable calculus in *Emperor*, pp. 66–70).

Like so many other mathematical ideas, especially the more profoundly beautiful and fundamental ones, the idea of computability seems to have a kind of *Platonic reality* of its own. This mysterious question of the Platonic reality of mathematical concepts is a central theme of much of *Emperor*, and I shall need to return to it later.

Having a very specific description of what a Turing machine is, we may turn to the *universal* Turing machine. This is a particular Turing machine U (having one fixed list of instructions) which can mimic any other Turing machine. All that is needed is that the Turing machines be coded in a well-defined way, so that each machine is assigned a number unique to that particular machine. The Turing machine with number n is denoted by T_n, and is referred to as "the nth Turing machine." To make U act like the machine T_n, all we need do is feed U with the number n on the tape first (say coded in the binary notation) and then (on the right of a suitable coded marker) we would feed in the tape that the machine T_n is supposed to be reading. The final result of the action of U on the whole combined tape would be the same as that of T_n on the right-hand part alone. Note that the machine U itself must have a number. In *Emperor* (pp. 56–57; 71–73) I exhibit this number (or one such – there are many possibilities) explicitly. It turns out to be about 7×10^{1654}.

Turing originally devised his "machines" to answer a question posed by David Hilbert: Is it possible in principle to find a mechanical mathematical procedure (i.e., an algorithm) for answering all mathematical problems in a specified class? Turing (and, independently, Church) showed that the answer is "No." Turing phrased his version of Hilbert's problem as the question of deciding whether or not a given Turing machine, when acting on a specific tape, will ever come to a halt; he showed (using a "diagonal argument") that there is *no* algorithm for systematically answering this question.

Many people seem to be under the impression that this means there are specific Turing-machine actions for which it is impossible to decide whether or not they halt. This is not what the argument shows at all, however. It merely shows that there is no *algorithmic procedure* for deciding this question *in general*. In any specific case, there is certainly such an algorithm, namely, the algorithm that simply says "yes" or else the algorithm that simply says "no" – but, of course, we would not know which of these two algorithms is the correct one to use! Algorithms do not in themselves decide questions of truth or falsity. For that we require *insights*, not algorithms. To illustrate this point, I give a modified version of Turing's original argument. I show that if we are given any algorithm that correctly decides whether Turing machine actions halt, though for some such actions that algorithm may itself run on forever without stopping, then we can exhibit a specific Turing-machine action that *we* can see does not come to a halt but for which the given algorithm never comes to a decision. In this sense we can in principle, by the use of insight, "outdo" any given algorithm for testing whether or not Turing machine actions halt.

Chapter 3: Mathematics and reality

The Mandelbrot set provides a wonderful illustration of a mathematical structure that, though defined in an entirely abstract mathematical way, nevertheless has a reality about it that seems to go beyond any particular mathematician's conceptions and beyond the particular technology of any specific computer. Though more and more of its wonderfully elaborate structure is revealed to us as more computer power is brought to bear on the problem of exhibiting the set, there is always more of the set to be found that is still hidden from us. The set seems clearly to be "there," somewhere, quite independently of us or of our machines. Its existence is not material, in any ordinary sense, and it has no spatial or temporal location. It exists, instead, in Plato's world of mathematical entities. When we use our computers to explore the set, it is like using a moon-rover to explore the moon's surface or a high-energy particle accelerator to probe the secrets of subatomic physics.

The precise mathematical definition of the Mandelbrot set is remarkably simple, considering the extraordinary complexity of its detailed shape. But to understand this definition, it is first necessary to come to terms with the idea of a complex number. It turns out that this is just as well in any case, for complex numbers are absolutely fundamental also to the structure of quantum theory, a theory that we need to come to terms with later.

Complex numbers are numbers of the form $x + iy$, where x and y are ordinary "real" numbers (i.e., numbers that can be expressed in terms of infinite decimal expansions) and where "i" represents a square root of -1. In ordinary terms, one tends to think that -1 does not "really" have a square root, and that it is just an "invention" on the part of mathematicians to suppose that it does. But the so-called "real" numbers are taken as more real only because we have got much more used to them, and because they accord remarkably closely with physical measurements such as time, distance, energy, temperature, and the like. There are reasons to expect that even this close correspondence with physical quantities may break down at a very tiny scale, and we must accept that the utility of the concept of a real number lies as much in its mathematical consistency, power, and elegance as it does in any correspondence with the physical world. We find, indeed, that the concept of a complex number leads us to a picture with perhaps an even greater

power and elegance than that of a real number; moreover, complex numbers also have a profound relationship with the workings of the physical world.

The very system of complex numbers provides further evidence for a Platonic existence for mathematical entities. While complex numbers were first introduced as a kind of "trick," useful in the solution of cubic equations, they were later found to have enormous power and utility in many different areas. In general, the case for Platonic existence is strongest for entities which give us a great deal more than we originally bargained for. In such cases, most of what we get out may be something that was not even remotely conceived by those mathematicians who first came across the entities in question. The mathematicians discovered something wonderful. They did not invent it. The Mandelbrot set, and even more so, the very system of complex numbers, are clear cases in point. There are many other situations in which the case for Platonic existence is not so strong, however, such as when a mathematician introduces a construction of no particular elegance or uniqueness to prove a particular result, this construction finding no value other than the one for which it was originally devised. In such cases it is not at all unreasonable to use the word invention, rather than discovery. It is the clear discoveries in mathematics that are the mathematician's proudest achievements.

The complex numbers have an elegant geometrical representation in the Euclidean plane, where the number $z = x + iy$ is plotted as the point with coordinates (x,y). A plane describing complex numbers in this way is called an *Argand plane*. The Mandelbrot set is a subregion of the Argand plane defined in terms of iterations of the map whereby z is replaced by $z^2 + c$, starting from $z = 0$, c being some fixed complex number. If these iterations lead to an unbounded sequence of points in the Argand plane then the point representing c lies outside the Mandelbrot set. If it is a bounded sequence, then c is in the Mandelbrot set.

One final comment about real and complex numbers should be made here. There are, in a precise sense, many "more" of them than there are natural numbers 0, 1, 2, 3, 4, 5, 6, 7, . . . to which the usual discussions of Turing machines refer. We might think of using some algorithm for generating the successive digits of an infinite decimal expansion, but it turns out that only a tiny fraction of the possible decimal expansions are obtainable in this way: the *computable numbers*. (Nevertheless, all the familiar numbers, such as the decimal expansion of pi, are computable.) This causes some problems when one tries to apply the concept of computability to descriptions of the physical world, which normally use real or complex numbers (however, see Blum et al. 1989).

Chapter 4: Truth, proof, and insight

The question of mathematical *truth* and how we ascertain it is fundamental to our considerations here. How do mathematicians actually decide which mathematical statements are true and which are false? Are they following some algorithm – perhaps one that is unknown or unknowable to them – or do they have some other route to truth via some mysterious "insight" that is not amenable to a purely algorithmic treatment? To gain some understanding of this issue, it is necessary to go back a little into the mathematical history of the last century or so. Mainly as a result of work by the mathematician Georg Cantor, it was found that very powerful methods of reasoning can be obtained if one is prepared to accept infinite sets as entities that can have independent existence in their entirety. It was soon found (by Bertrand Russell and others), however, that contradictory results are liable to arise with this kind of reasoning unless some special rules are introduced to limit the "sizes" of the sets under consideration. What should these rules be? Russell, Hilbert, and various other mathematicians proposed some very precise systems of axioms and rules of procedure – called formal systems – with the intention that these should incorporate all the legitimate rules of mathematical procedure. If they were to be completely successful, then any true mathematical proposition (within some well-defined area of mathematics) ought to be provable by means of the procedures laid down in the formal system (completeness) and, moreover, it should not be possible to prove both a proposition and its negation (consistency). This led to the mathematical standpoint of *formalism*, according to which mathematics could be reduced to a kind of "game" consisting of merely manipulating symbols according to the specific rules of the formal system in question. The necessity for the symbols to have any actual "meaning" would thereby be eliminated.

Fortunately (to my way of thinking), in 1931, Kurt Gödel presented his famous theorem, which effectively destroyed formalism as a fundamental philosophy of mathematics. He showed that in any consistent formal system that is broad enough to contain arithmetic and the normal rules of logical procedure one can explicitly construct well-defined mathematical statements that are not provable – nor are their negations provable – using the rules laid down in that formal system. Thus, the system cannot be complete, in the sense that Hilbert and others required. Worse than this (from the formalist's point of view) by the very way that such a Gödel proposition is constructed we can *see*, using our insight and understand about what the symbols in the formal system are supposed to mean, that the Gödel proposition is actually *true!* This tells us that the very concepts of truth, meaning, and mathematical insight cannot be encapsulated within any formalist scheme.

This is not just bad news for the formalists. It is bad news for strong-AI, as well. For there is a very close relationship between the concept of an algorithm and the concept of a formal system, with regard to mathematical statements. For any formal system, there is always an algorithm that generates precisely all the propositions that can be proved within that system. Conversely, given an algorithm for generating mathematical statements, one can always construct a formal system that incorporates all these statements as axioms of the system. This tells us that mathematical truth is not an algorithmic matter; It also appears to tell us that meaning and insight are not algorithmic matters either. I return to this issue in Chapter 10.

Since Gödel's theorem seems to be telling us that formalism is not tenable as a foundation for mathematics, are we to be driven to a *Platonic* view that mathematical concepts are in some sense just "out there" waiting to be

discovered? In my own view, something like this must be right, though many people find it hard to believe. In particular, there is another mathematical standpoint, referred to as *intuitionism*, according to which completed infinities are not permitted, and are allowed only to have a potential existence. (The intuitionistic philosophy, due mainly to L. E. J. Brouwer, stems more from the ideas of Aristotle than those of Plato.) In this view, what is important is that abstract proofs of "existence" are not allowed unless an explicit construction of the quantity in question is given. The main problem with intuitionism lies in the severe limitations that it places on mathematical reasoning. (In particular, the powerful method of *reductio ad absurdum*, which is used so frequently in mathematics, is not now permitted.) Moreover, at least in the way it is often presented, the intuitionistic concept of truth has a subjectivity and time-dependence that seems very unsatisfactory for a mathematics that one hopes to be able to use reliably in a description of the physical world.

Accepting that much of mathematics is of a nonalgorithmic character, we can ask whether nonalgorithmic mathematics is in any way interesting or mathematically important. In fact, it is possible to give numerous simple and interesting examples of nonalgorithmic mathematics. Diophantine arithmetic (the question of whether systems of polynomial equations in integers, with several variables, have solutions) turns out to be nonalgorithmic (this was actually Hilbert's original question); so does the topological equivalence problem for 4-manifolds (of possible relevance for the problem of quantum gravity) and, perhaps most striking, the problem of deciding whether or not a given finite set of polygonal shapes will tile the entire Euclidean plane. In each case, we have a family of well-defined yes/no questions for which it is known that no algorithmic solution exists. In *Emperor* I also conjectured that the problem of deciding whether or not a point of the Argand plane actually lies in the Mandelbrot set is, in an appropriate sense, nonalgorithmic[1] – despite the fact that computers are able to provide wonderful approximations to this set!

Another body of understanding that has considerable importance for computability problems is *complexity theory*. As with questions of algorithmic solubility, complexity theory is concerned with families of mathematical problems, but where the problems in each family are taken to be algorithmically *soluble*. The problem is to decide how "good," in some well-defined sense, such an algorithm can be for a given family of problems. It turns out that for some there are fast algorithms (the so-called *P* problems), while for others the algorithms seem to be of necessity so slow that the problems are, in effect, intractable (i.e., insoluble algorithmically "in practice" – e.g., the so-called NP-problems that are not in *P*). In my opinion, the questions of complexity theory are not so central to the issue of conscious thinking as are the questions of computability theory, but many people seem to hold to the contrary. [See Tsotsos: "Analyzing Vision at the Complexity Level" *BBS* 13(3) 1990.]

Chapter 5: The classical world

We now turn to the question of how the physical world actually behaves. The issue of *determinism* in connection with physical laws has often been discussed. I am more concerned with a different (though somewhat related) issue that, with the important exception of some significant work by Pour-El and Richards (1981; 1982), has barely been addressed at all: that of *computability* in physical laws.

Before we can come properly to terms with these issues, however, we must have some idea of what the laws of physics actually *are*. It is a very remarkable fact that there are indeed physical laws that have a truly *phenomenal* accuracy and, moreover, are extraordinarily amenable to precise and general mathematical treatments. These physical laws provide the theories that I refer to as SUPERB. These theories are supplemented by others that help to explain physical phenomena but do not share with the SUPERB theories their phenomenal accuracy. Such theories I refer to as USEFUL. In addition, there are numerous ideas, some currently very fashionable, and often excitedly expressed, that (essentially) have no experimental support whatever. These are theories belonging to the class TENTATIVE. In *Emperor* I give a brief account of each of the SUPERB theories in turn.

The oldest of the SUPERB physical theories is Euclidean geometry, which provides a marvellously accurate theory of physical space and of the behaviour of rigid bodies. The fact that Euclidean geometry is an extraordinarily precise physical theory, and not just a very elegant area of pure mathematics, is demonstrated (perhaps somewhat ironically) by the fact that we now know that this geometry is actually not exactly true of the physical space we inhabit. The even more accurate geometry of Einstein's general relativity provides a picture of a curved space(-time) that deviates in very tiny and barely measurable ways from the geometry of Euclid. To qualify as SUPERB, it is not necessary that the theory in question apply without exception to the observed properties of the world, only that its agreement with those properties qualify as "phenomenal" in some appropriate sense. Be that as it may, there is one essential feature of Euclidean geometry (introduced by Eudoxos in the fourth century B.C.) that remains with us to this day as a fundamental ingredient of *all* the SUPERB (and also USEFUL) theories: the concept of a *real number*.

Probably Archimedes's theory of statics would have qualified as SUPERB, but this theory is now subsumed into the SUPERB dynamical theory of Galilei/Newton. *Newtonian mechanics* is, as we all know, deterministic, but is it *computable*? Suppose initial data for some physical situation is given in terms of computable numbers (all constants involved being also computable numbers), and we wait for a computable time. Is the state at that time computable from the initial data? As far as I am aware, the question has not yet been properly addressed mathematically, though my guess is that the answer would be "yes" in Newtonian theory (with reasonable force laws – say inverse square, with hard elastic spheres and "generic" initial data, so as to avoid the problem of triple collisions).

It should be made clear that the possible "noncomputability" that is being considered here is different from the much-discussed concept of "chaos." *Chaos* refers to the fact that very tiny changes in initial data may give rise to absolutely enormous changes in the resulting be-

haviour. Chaotic systems would normally be computable, in my sense. (The issue of chaos is like that of complexity theory, as discussed above, rather than computability theory.) Chaotic systems are noncomputable *in practice*, rather than *in principle*. In practice, they introduce a random element into the evolution. Such a random element is not the kind of thing that could be usefully "harnessed" by the brain in order to achieve something nonalgorithmic. Newtonian mechanics can be incorporated into the more general framework of *Hamiltonian mechanics*, which can also be made to include Maxwell theory and relativity. In fact, however, most Hamiltonian systems appear to be chaotic. It seems likely to me, however, that normal Hamiltonian systems (with computable constants) should be *computable*.

Maxwell's theory provides the SUPERB equations governing electric and magnetic fields, and it shows how these fields propagate through space in the form of light or radio waves. The radical new ingredient that Maxwell (and Faraday) introduced was a physical entity with an infinite number of degrees of freedom. Like Newton's laws, Maxwell's equations are deterministic and presumably computable in some appropriate sense. (Some curious problems in relation to this have been pointed out by Pour-El and Richards [1981], although it seems likely that the "noncomputability" they encounter is not of physical importance, arising, as it does, only with data that are not smooth.) To describe the motion of charged particles in addition to electromagnetic fields, Maxwell's equations have to be supplemented by the *Lorentz force law*. The Maxwell-Lorentz equations work well in practice but can lead to problems of principle with regard to determinism, as was pointed out by Dirac. This anomaly is usually ignored on the grounds that one should really be using quantum theory for such situations (though the quantum theory is actually no better!).

Einstein's SUPERB *special theory of relativity* changes our picture of space-time, but it does not seem to alter the issues of determinism and computability (except slightly to improve the situation with regard to determinism). It does radically affect the way that we view the "passage of time," however. His even more SUPERB *general* theory of relativity asserts that space-time is curved and that this curvature describes the gravitational field. An important property of space-time curvature is that it can be naturally split into two pieces, which I refer to as RICCI and WEYL. Einstein's field equations assert that RICCI is directly determined by the mass-energy distribution of matter (counting the electromagnetic field also as matter), and the remaining part, WEYL (which describes tidal distortion), represents the free gravitational field. It is conceivable that there are situations involving ultrastrong gravitational fields in which there is a failure of ordinary determinism (failure of "cosmic censorship"), but such situations could in any case occur only at a scale totally different from that of human brains.

Relativity provides us with a curious paradox concerning the nature of matter. Matter is quantified by its mass, but gravity itself has energy and therefore (by Einstein's $E = mc^2$) also mass. In certain circumstances, the mass of gravity appears to reside in regions of space-time that are not just empty but actually *flat!* I mention this merely to point out that even in classical theory there can be a puzzling nonlocality about the very nature of matter.

Such nonlocality is an even more puzzling with the mysterious but SUPERB theory of *quantum mechanics*, and also with that somewhat unsatisfying but powerful theory referred to as quantum field theory, which arises when the principles of special relativity are combined with those of quantum mechanics. The particular quantum field theory that applies to the electromagnetic field interacting with electrons (or with certain other particles such as protons) is called *quantum electrodynamics*, and it also qualifies as SUPERB. We consider quantum theory next.

Chapter 6: Quantum magic and quantum mystery

To have any hope of understanding the mysterious phenomenon of consciousness in physical terms, one must try seriously to come to terms with the physical laws that govern the way things *actually* behave in our universe. With the possible exception of the structure of space-time itself, the most fundamental and mysterious of these physical laws are the laws of *quantum mechanics*.

There is a basic but very puzzling quantum mechanical principle – called the *superposition principle* – that asserts that if A is a possible state of a system and if B is another possible state, then $wA + zB$ is also a possible state, where w and z are two complex numbers, not both zero (the ratio $w{:}z$ being what is physically important). What does this mean? In a particular instance, consider the state of a particle, which might be at one point (state A) or else at another (state B). The superposition principle tells us that another possible state is $A + B$ and another is $A - B$, while yet another is $A + iB$ where i is the square root of -1). Thus, not only must we consider that a particle can be "in two places at once," but also that there are many *different* ways of being at two places at once! All these different possibilities have different experimental consequences.

It appears that the superposition principle holds only at the "quantum level," where differences in alternative possibilities are in some appropriate sense small (small differences in energy distribution – or some such criterion), whereas at the "classical level," which seems to include the level of our ordinary experiences, the superposition principle seems to make no sense at all. Instead, we must adopt a new rule whenever the respective effects of A and B are in some way magnified to the classical level. This rule asserts that the so-called "probability amplitudes" – in effect, the complex numbers w and z in the above description – must have their *squared moduli* formed (their distances from the origin in the Argand plane), and the ratio $|w|^2{:}|z|^2$ provides the relative probabilities for each of the two alternatives in question *actually* to occur. I refer to this so-called "observation" or "measurement" process as **R** (*reduction* of the state vector, or "collapse of the wavefunction").

In my own opinion, **R** is a *real physical process* which takes place spontaneously in suitable physical circumstances, independently of human intervention or the introduction of "consciousness." Physicists seem to hold to various different views on this, many maintaining that **R** is in some way a kind of "illusion." Such theorists often find themselves driven to a very "subjective" view of physical reality at the quantum level. One reason phys-

icists have trouble with **R** is that it is completely different from the process **U** that governs the behaviour of a state which remains at the quantum level. The action of **U** (*unitary* evolution, described by the important Schrödinger equation – or else by something equivalent such as the Heisenberg equation) is entirely deterministic, just like a classical field equation such as Maxwell's. The probabilities in quantum theory arise only with the process **R**, that is, when effects become magnified from the quantum to the classical level.

Some physicists (probably only a small minority) believe that the superposition principle should still be applied at the classical level (so **R** never actually takes place at all); they are led to the *many-worlds* (or Everett) interpretation of quantum mechanics (Everett 1957; DeWitt & Graham 1973), whereby all alternative possible universes must be considered to coexist in some vast complex-number superposition. No convincing explanation is provided, however, for why we (as conscious entities?) should perceive just *one* out of the infinitude of superposed alternatives, or why the quantum mechanical probability rules should apply.

Whether or not one adheres to this strange view, there seems to be no completely satisfactory alternative within the confines of standard quantum theory. The alternative viewpoints include the "subjective" one referred to above, according to which the quantum state (despite its SUPERB accuracy in providing agreement with observation) is not taken seriously as a description of reality ("all in the mind" of the experimenter!), and there is the alternative view that somehow the "illusion" of **R** occurs whenever the complicated interaction with the environment is taken into account. (I do not find either of these views, in themselves, to be very satisfying or convincing). Some people prefer the view (Wigner 1961) that a *conscious entity* (or perhaps just a "biological system") would not be subject to the superposition principle, and that consequently some *new nonlinear theory* is needed to handle the way quantum mechanics could be applied to such an entity. This would entail an actual change in the basic structure of quantum theory. I am personally very sympathetic to the general view that a fundamental change is needed, but I do not think that it is at the level of the phenomenon of consciousness (or biology). That would lead to a very lopsided view of the reality of the physical world. For in those corners of the universe in which consciousness (or biology) resides, the objective physical behaviour of matter would then be totally different from its behaviour everywhere else (where there would be planets with complex superpositions of many different weather patterns, etc. – and probably much worse!). Other deviations from standard linear quantum theory that do not involve consciousness and seem to me to have much greater plausibility (e.g., Ghirardi et al. 1986; Pearle 1989) have been proposed in order to resolve the very basic difficulties that arise from the **U**/**R** conflict. These difficulties are made particularly graphic in the famous Schrödinger (1935) "cat" thought experiment. I shall come to my own specific suggestions on the matter later.

One of the most puzzling aspects of quantum theory lies in the way that systems of *many particles* must be treated. (Complicated quantum systems do not behave like classical ones, just by virtue of their complication, despite what many physicists appear to think. Somehow **R** must come into it, but the matter remains very unclear.) When there are many particles, individual particles are not to be considered as objects on their own. Their quantum states involve superpositions of each particle's possible individual state combined with the various possible individual states of all the other particles in the system under consideration. (The different particles are then said to be *correlated*.) This has the effect that if one particle is "observed" (i.e., the particle triggers an effect that magnifies its situation from the quantum to the classical level), then this instantaneously affects the quantum state of all the other particles with which it is correlated. This effect leads to what is known as the *Einstein-Podolsky-Rosen* (EPR) *paradox* (Einstein et al. 1935). It is not really a physical paradox, but an actual physical effect that is not explicable on the basis of any local realistic view of physical behaviour. (This follows from a remarkable theorem by J. S. Bell (cf. Bell 1987) and is substantiated in the behaviour of the actual world in experiments such as those performed by A. Aspect et al. 1986).

In my own view, there is nothing in principle objectionable about a nonlocal (or somewhat "holistic") picture of physical reality, but there is a very serious difficulty with obtaining a picture that is consistent with the spirit (and space-time descriptions) of relativity. This presents a profound challenge for what I believe to be a much-needed new theory.

Chapter 7: Cosmology and the arrow of time

Why do we perceive time to "flow," when our (SUPERB) theories tell us that in "reality" there is just a static space-time laid out, with no "flowing" about it? Relativity even tells us that there cannot be any such things as "now" at all, since the very concept would depend on how various other observers might be moving. Even worse, apparently, is the fact that the equations of physics are *symmetrical in time*, so that they would apply just as well in the reverse temporal direction as in the normal forward direction. Why does time seem to "flow" forwards and not backwards?

At least a partial answer to this last question can be found in the *second law of thermodynamics*. This law is time-asymmetric; and it asserts that a certain physical quantity called *entropy* increases in the forward time-direction (and, consequently, it decreases in the backward time-direction). Very roughly speaking, entropy is a measure of the manifest disorder of a physical system. It is not unnatural physically that a system, when left on its own, should get more and more disordered. The *puzzle* of the second law lies in the fact that this is not what happens in the *reverse* direction in time or, to put things another way, that the state in the past was actually given to us as extraordinarily ordered. We can trace the nature of this "order" to the very structure of the *big bang* – the singular state (infinite space-time curvature) that, according to standard (USEFUL, at least) theory, represents the actual origin of the entire universe.

We can compare this initial singular state with the ones that are expected to occur in the reverse direction of time, at the cores of black holes – or in the "big crunch" that will

finally engulf the entire universe, if our universe turns out actually to be finite rather than infinite. The geometrical difference, in the case of the big bang, appears to lie in the fact that in the immediate neighbourhood of this singular state the space-time curvature is constrained by the condition: WEYL = 0. The hypothesis whereby initial-type singular states (i.e., at the big bang, or in "white holes" – the time-reverses of black holes) are constrained by WEYL = 0, whereas those of final-type (i.e., at the big crunch, or in black holes) are not to be so constrained, is what I refer to as the *Weyl curvature hypothesis* (WCH). This hypothesis would explain both the puzzle of the second law and the observed extraordinary uniformity of the actual universe. Such a constraint could explain the fantastic geometric precision involved in the "act of creation" that occurred in the big bang – to one part in at least $10^{10^{123}}$ (a number so large that it could not remotely be written out in full even if each digit were to be written on each separate electron, proton, neutron, or any other particle in the entire universe!).

Chapter 8: In search of quantum gravity

What lies behind WCH? Could this hypothesis be a deduction from some comprehensive theory of physics that deals also with problems at other scales? The usual viewpoint about the singular states that classical general relativity leads us to (at the big bang, big crunch, and in black or white holes) is that they must be dealt with by a *quantized* general relativity, or *quantum gravity*. This is also my own view, but the fact that quantum gravity needs to explain a grossly time-asymmetric phenomenon like WCH does not seem to be generally appreciated. No satisfactory theory of quantum gravity has yet emerged, but when it does (and let us call this putative theory CQG – "correct quantum gravity"), it ought to turn out, on the basis of the above considerations, to be a time-*a*symmetric theory. This is very different from the kind of thing that has been expected until now, but considerations of gravity quantizers have so far been restricted to trying to derive a (time-symmetric) quantum procedure **U** for the time-symmetric classical theory of general relativity (or one of its equally time-symmetric generalizations). No one seems to have tried seriously to incorporate the time-*a*symmetric procedure **R** into quantum gravity theory. I give an argument in *Emperor* to show that **R** is indeed time-*a*symmetric (a fact not always recognized by physicists) and I describe a thought experiment ("Hawking's box") that strongly suggests that if CQG implies WCH, then CQG also ought to incorporate **R** as an actual physical process.

This leads to the suggestion that the onset of **R** can be understood as something that occurs when pairs of states are superposed and the differences in their gravitational fields (i.e., the differences in their space-time geometries) reaching the "one-graviton level," or thereabouts. Simple physical examples are described which indicate that this criterion is not at all physically unreasonable.

It seems to me that an important new physical theory must be lurking in the shadows. But recall the difficulties referred to above: Any realistic theory of **R** would have grave difficulties in coming to terms with the space-time descriptions that must accord with the principles of relativity. In my opinion, our present picture of physical reality is due for a grand shake-up – even greater, perhaps, than that which has already occurred with relativity and quantum theory.

Chapter 9: Real brains and model brains

What do we know of the actual structure and workings of the human brain? What aspect of the brain seems to relate most to the phenomenon of consciousness? There seems to be remarkably little consensus as to what actual parts of the brain are to be most associated with this phenomenon. The cerebral cortex, reticular formation, thalamus, hippocampus, and no doubt many other parts of the brain have each have been separately suggested as the place to look for the "seat of consciousness." Some people have suggested that only the left half of the brain is conscious, since (in most individuals) that is the half capable of speech – an odd view, to my mind, and certain experiments on split-brain patients indeed suggest that both halves can be conscious. Perhaps consciousness is not to be located in any one clear place. Some parts of the brain do seem to be more closely associated with consciousness than others, however. The cerebellum, for example, really does seem to act as an unconscious "automaton," whereas one does sometimes (although apparently only sometimes) seem to be aware of activity taking place in the cerebral cortex. This might seem surprising, on the basis of the strong-AI picture, the cerebellum having a much higher density of neurons than does the cerebral cortex, and perhaps half as many neurons altogether!

One can present computer models of the activity of the firing of neurons, but is this liable to give a reliable modelling of the brain's (conscious?) activity? There is at least one feature of brain functioning that is not well modelled in this way, and that is *brain plasticity*, according to which the connections between different neurons can become strengthened or weakened, providing the essential ingredient of permanent memory, according to one prevalent theory. There are suggested mechanisms underlying brain plasticity (such as the one due to Hebb 1954) and these can be modelled on computers as *neural networks*. The processes underlying *actual* brain plasticity seem to be largely unknown, as of now, however.

Parallel processing is another ingredient that seems to have importance in the computer modelling of brain function. It seems unlikely to me however, that any proper understanding of conscious thought processes will come about in this way. There is a remarkable "oneness" in conscious thinking that seems very much at odds with the functioning of a (classical) parallel computer. In any case, parallel computers are completely equivalent to the normal serial (Turing machine type) computers with regard to what they can and cannot compute.

Is there an essential role for quantum phenomena in brain functioning? Perhaps quantum superposition is somehow usefully incorporated into brain function. One might envisage that huge numbers of different calculations are carried out simultaneously in superposition, and only at the end is **R** called into play to conjure up the answer that is required. Such would be the action of a *quantum computer*, as considered by Deutsch (1985;

Deutsch's quantum computers, however, do not perform noncomputable operations – although in certain rather contrived situations they can do better than Turing machines in the sense of complexity theory.) It is known that there are neural cells (in the retina) that are sensitive to single quantum events (photon arrival), but it is hard to see how to harness this kind of thing in any useful way, the brain being too "hot" a system to preserve quantum coherence over an appreciable length of time.

Quantum considerations are relevant to many aspects of brain functioning (chemical transmission, the definite potential differences responsible for the on/off transmission of nerve signals, etc.), but it is difficult to see how present-day quantum mechanics could be coherently used to describe the action of a human brain, even in principle: For the brain would have to be considered as "observing itself" all the time – when it is conscious, at least, which would entail the continual use of **R**. There is no satisfactory theory for handling this.

Chapter 10: Where lies the physics of mind?

What is the selective advantage of consciousness? Our consciousness has presumably evolved because the behaviour of an animal with consciousness is actually more advantageous in some way to that of an otherwise equivalent animal without consciousness. If consciousness is merely the inevitable inward manifestation of the possession of a sufficiently complex control system, why is it that that some parts of the brain (e.g., the cerebellum) do not appear to be conscious, yet they can perform extremely complex tasks? What is advantageous about the outward manifestations of those particular thought modes that seem necessarily to be associated with consciousness?

If one contrasts the phrases (e.g., "common sense," "judgement of truth," "understanding," "artistic appraisal") that might be associated with many of those mental activities that require some degree of consciousness with those (e.g., "automatic," "following rules mindlessly," "programmed," "algorithmic") that do not seem to require it, then the possibility of a *non-algorithmic/algorithmic* distinction at least suggests itself. In my view this is one strong indication that what we are doing with our consciousnesses is actually not something algorithmic at all. As a central theme of *Emperor*, I have tried to stress that the mere fact that something may be scientifically describable in a precise way does not imply that it is computable. It is quite on the cards that the physical activity underlying our conscious thinking may be governed by precise but nonalgorithmic physical laws and our conscious thinking could indeed be the inward manifestation of some kind of nonalgorithmic physical activity. I am suggesting, therefore, that the selective advantage of consciousness is that it enables its possessor to form some kind of nonalgorithmic judgement of how to behave in a given situation.

Perhaps one could refer to such judgements as "inspired guesswork;" it is of some value to examine some outstanding examples of inspiration, as recorded by Poincaré and Mozart, for example. Here the inspirational ideas appear to be thrown up by the *un*conscious mind, but it is the *conscious judgements* that are needed to assess the value of the ideas themselves. These judgements have a remarkable globality about them; a vast area seems to be surveyable in an "instant" – say, an entire mathematical topic, or a symphony. Such globality is also a feature of much of our conscious thinking at a (seemingly) much more mundane level, such as deciding what to have for dinner or appreciating a visual scene.

Why do I maintain that such conscious judgements must have an essentially nonalgorithmic ingredient? One could certainly imagine feeding appropriate criteria into an ordinary algorithmic computer and getting it to produce "judgements." The most decisive reason for believing that our conscious judgements must be nonalgorithmic comes from *mathematics*. If mathematical judgements can be seen to be nonalgorithmic – where the criteria of logic, precision, correct calculation, and truth are normally taken as paramount – then surely nonalgorithmic ingredients could have at least as great an importance in other areas.

In Chapter 4 of *Emperor*, I show how to construct, for any (sufficiently broad) formal mathematical system, a specific Gödel proposition $P_k(k)$, which is a well-defined statement about numbers. It has the form "for every natural number x, the following computable property of x holds." From the way that $P_k(k)$ is constructed, one sees, provided that one believes that the axioms and rules of procedure of the formal system are valid methods of deriving mathematical truth, that one must believe that $P_k(k)$ is a mathematical truth *also*. Nevertheless, $P_k(k)$ is *not* itself derivable by means of the axioms and rules of procedure of the given formal system.

We recall that an essential equivalence exists between formal systems and algorithms as procedures for ascertaining the truth of mathematical propositions. Now suppose that a particular mathematician is using some algorithm – that is, in effect, some formal system F – as his means of ascertaining mathematical truth. Then the Gödel proposition $P_k(k)$ constructed from F must be a true proposition also, though it is not possible for our putative algorithmic mathematician to ascertain the truth of $P_k(k)$. This is essentially the argument put forward by Lucas (1961), but it is not yet the desired contradiction, since the mathematician can have no means of knowing what F *is*, let alone be convinced of its validity as a means of ascertaining truth. We shall need a broader argument than this.

Suppose that the ways that human mathematicians form their conscious judgements of mathematical truth are indeed algorithmic. We shall try to reduce this to an absurdity. Consider, first, the possibility that different mathematicians might use *inequivalent* algorithms to decide truth. But it is one of the most striking features of mathematics (perhaps almost alone among the disciplines) that the truth of propositions can actually be settled by abstract argument. A mathematical argument that convinces one mathematician – provided that it contains no error – will also convince the other, as soon as the argument has been fully grasped. This also applies to the Gödel-type propositions. If the first mathematician is prepared to accept all the axioms and rules of procedure of a particular formal system as giving only true propositions, then he must also be prepared to accept its Gödel proposition as describing a true proposition. It would be exactly the same for the second mathematician. The point

is that the arguments establishing mathematical truth are *communicable.*

Thus we are not talking about various obscure algorithms that might happen to be running around in different particular mathematicians' heads. We are talking about *one* universally used (putative) formal system that is equivalent to all the different mathematicians' algorithms for judging mathematical truth. Now this putative "universal" system, or algorithm, cannot ever be known as the one that we mathematicians use to decide truth. For if it were, then we would construct its Gödel proposition and know that to be a mathematical truth also. Thus, we are driven to the conclusion that the algorithm that mathematicians actually use to decide mathematical truth is so complicated or obscure that its very validity can never be known to us.

But this flies in the face of what mathematics is all about! The whole point of our mathematical heritage and training is that we do not bow down to the authority of some obscure rules that we can never hope to understand. We must *see* – or, at least in principle see – that each step in an argument can be reduced to something simple and obvious. Mathematical truth is not a horrendously complicated dogma whose validity is beyond our comprehension. It is something built up from simple and obvious ingredients – and when we understand them, their truth is clear and agreed by all.

To my thinking, this is as blatant a *reductio ad absurdum* as we can hope to achieve, short of an actual mathematical proof. The message should be clear: Mathematical truth is not something we ascertain merely by the use of an algorithm. I believe, also, that our *consciousness* is a crucial ingredient in our comprehension of mathematical truth. We must "see" the truth of a mathematical argument to be convinced of its validity. This "seeing" is the very essence of consciousness. When we convince ourselves of the validity of Gödel's theorem we not only "see" it, but by so doing we reveal the very nonalgorithmic nature of the "seeing" process itself.

The strong-AI view, on the other hand, envisages that human judgements of (mathematical and other) truth must arise out of some kind of *natural selection of algorithms.* Even in the absence of arguments in favour of some nonalgorithmic ingredient in conscious thinking, such as the one given above, there are serious difficulties with the picture whereby algorithms are supposed to improve themselves in this way. It would certainly not work for normal Turing machine specifications, since a "mutation" would almost certainly render the machine totally useless instead of altering it only slightly. Something much more robust would be needed – such as the actual *ideas* underlying the specification! Moreover, a selection process that relies only on the output of an algorithm (such as natural selection implies) would be hopelessly inefficient.

What do I believe is actually taking place with our conscious perceptions? I am venturing a suggestion (partly as a result of my own experiences with the communication of mathematical ideas) that consciousness represents some kind of contact with the timeless Platonic world of mathematical concepts. Proper communication between mathematicians can take place only when each individually makes this contact. Moreover, the "nonalgorithmic part" of Plato's world is, by comparison with the part where algorithms and computation reside, easily the most subtle and fascinating part. But mathematical insight is only one area where the role of consciousness is important, and it is chosen in my descriptions simply because it is here that the nonalgorithmic element can be made most precise. *Any* conscious thinking, even that which I believe must occur in (at least some) animals would, at root, have to be the same phenomenon.

The relation between the real physical world and the Platonic one is a mysterious one. The very existence of the SUPERB physical theories begins to give the seemingly "solid" matter of our experiences an apparently nebulous "mere mathematical" Platonic existence, whereas examples such as the Mandelbrot set seem to make the Platonic world more concrete. Might the two worlds be, in some sense, actually the *same?* The whole issue of consciousness seems to me to require a much deeper investigation of the relation between the physical world and Plato's world than has been achieved so far. Even the standpoint of strong AI requires the Platonic existence of algorithms as the "home" for our conscious feelings, since the algorithms' physical embodiments are supposed to be irrelevant. Viewed in this way, my own standpoint does not differ so much from that of strong AI, except that I believe that the concept of an algorithm is far too limiting to accommodate the immensely subtle and fundamentally important phenomenon of consciousness.

The possible identification, in some sense, of the physical world with the Platonic one raises many issues. There is the possibility of *strong determinism* (according to which the actual history of the universe might be mathematically fixed in its entirety). There are also the various versions of the *anthropic principle,* which I discuss briefly in *Emperor.* In its (more dubious?) *strong* form there could even be *alternative universes,* the one in which we find ourselves being the one that happens to allow consciousness.

Suppose we accept that there may indeed be a nonalgorithmic element in the physics that spans the gap between the quantum and classical levels. How might the conscious brain be making use of this? Here the speculations become more tentative. A plausible place to look is in the phenomenon of *brain plasticity.* Connections between neurons sometimes occur at *dendritic spines,* which can apparently grow or shrink, and significantly affect connection strengths, in seconds or less. This growth or shrinkage should, strictly speaking, be considered as taking place at the quantum level: No single one of the vast array of possibilities has actually occurred; *quantum superpositions* of many different combinations of connection strengths must be simultaneously involved. Thus, the brain embarks not on just one "calculation," but on many simultaneously. Only when the large-scale effects of these simultaneous calculations differ by something that reaches the "one-graviton level" – or whatever CQG tells us is appropriate – would one or another "conscious thought" emerge. Moreover, this "emergence" would have to involve some nonalgorithmic ingredient.

I suggest a possible analogy with *quasicrystal* growth. These puzzling objects resemble crystals, but possess crystallographically forbidden symmetries, usually *fivefold.* Fivefold ("almost") symmetry can occur with certain tiling patterns, however, and it seems likely that

the arrangement of the tiles in such patterns will give the essential clue as to the arrangement of atoms in a quasicrystal. If so, there would have to be an essentially nonlocal aspect to the way the atoms are arranged. I am of the opinion that this cannot be properly achieved by adding atoms to the assembly one at a time. Quantum superposition of many different simultaneous arrangements must be involved; only when the classical level is reached does a particular one of these arrangements get resolved out. (It should be recalled that the general tiling problem is a nonalgorithmic one, although the particular tilings that are likely to be relevant for such quasicrystals are actually algorithmic – though nonlocal.)

Finally, I refer in *Emperor* to certain experiments (by Kornhuber and by Libet) that indicate some very puzzling aspects of the actual time at which conscious feelings (active and passive, respectively) seem to take place. [See Libet: "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action" *BBS* 8(4)1985.] There is also something very puzzling about the fact that consciousness is the one known phenomenon for which time needs to *flow* at all. In the descriptions of modern physics – except, perhaps, for the (disputed) behaviour of **R** – all we have is a "static" space-time. I believe that these are strong indications that we need a radical revision of our present picture of space-time. For such a revision, we would certainly need something like the putative CQG.

Despite all this technicality (and also physical speculation), there remains the fact that in some sense it is "obvious" that mere computation cannot evoke consciousness – an obviousness that a child can see. Yet our science has driven us to accept that we are all but small parts of a universe governed by precise mathematical laws. In *The Emperor's New Mind,* I have attempted to show that there is a way out of the dilemma: Mathematical precision does *not* imply computability.

NOTE

1. I have been recently informed by Leonore Blum that this conjecture is actually true, if one uses the definition of noncomputability for complex numbers due to Blum et al. 1989.

Open Peer Commentary

Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

On "seeing" the truth of the Gödel sentence

George Boolos

Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA 02139
Electronic mail: *boolos@cogito.MIT.EDU*

In his famous 1931 paper, Gödel showed that for any "sufficiently strong" formal theory T, a sentence S in the language of T equivalent in T to its own T-unprovability cannot be proved in T, *provided that* T is consistent. (In the normal cases, S is equivalent in T to the sentence expressing the consistency of T.) Thus if T proves only true sentences, and is therefore consistent, then S is not provable in T.

Roger Penrose claims that although S is unprovable in T, we can always see that S is true by means of the following argument: If S is provable in T, then S is false, but that is impossible. (pp. 107–8: "Our formal system should not be so badly constructed that it actually allows false propositions to be proved!"); thus S is unprovable and therefore true.

There are certain interesting formal theories of which the set of provable sentences can be seen to contain no falsehoods; for the sake of argument we may grant that Peano Arithmetic (PA), say, is one of these. We must then grant that the Gödel sentence for PA, expressing its own PA-unprovability, is true and unprovable in PA.

To concede that we can see the truth of the Gödel sentence for PA, in which only a fragment (albeit nontrivial) of actual mathematical reasoning can be carried out, is not to concede that we can see the truth of Gödel sentences for more powerful theories such as ZF set theory, in which almost the whole of mathematics can be represented. I shall give some reasons for thinking that there is no sense of "see" in which we can see that ZF is consistent; thus we cannot see the truth of the Gödel sentence for ZF either, for that sentence is equivalent (in a much weaker theory than ZF) to the consistency sentence for ZF.

A true story: Once upon a time, distinguished set theorist J sent equally distinguished set theorist M what purported to be a proof that the theory ZFM (ZF + "a measurable cardinal exists"), of which M and many others were fond, was inconsistent. M sat down to work and found the error on page 39 or so of J's manuscript. As he began to examine J's "proof," M might have been reasonably confident that he would find an error, but by no means did he then know that J's "proof" was fallacious or see the consistency of ZFM. Do we know that some future hotshot will not do to ZF what M feared J had done to ZFM?

I suggest that we do not know that we are not in the same situation vis-à-vis ZF that Frege was in with respect to naive set theory (or, more accurately, the system of his *Basic Laws of Arithmetic*) before receiving, in June 1902, the famous letter from Russell, showing the derivability in his system of Russell's paradox. It is, I believe, a mistake to think that we can see that mathematics as a whole is consistent, a mistake possibly fostered by our ability to see the consistency of certain of its parts.

The verb "should" in the sentence quoted above ought to give us pause. Of course our formal system *should* not be so constructed as to have false theorems. What we may believe or hope to be the case, but cannot "see" to be so, is that the totality of mathematics *is* not badly constructed in that way. Are we really so certain that there isn't some million-page derivation of "0 = 1" that will be discovered some two hundred years from now? Do we know that we are really better off than Frege in May 1902?

To belabor the point: Penrose has said nothing that shows that we can recognize the truth of the Gödel sentence for ZF or for any other reasonable approximation to the whole of the mathematics that we ourselves use. What we can see the truth of is this conditional proposition: The Gödel sentence for ZF is ZF-unprovable (and therefore true) *if* ZF is consistent. We cannot see that the Gödel sentence is true precisely because we cannot see that ZF is consistent. We may hope or believe that it is, but we do not know it, and therefore cannot see it.

Penrose does offer a kind of consideration not advanced in earlier discussions of Gödel's theorem. He states that when a mathematician discovers a proof of some statement, other mathematicians easily and quickly convince one another of its truth.

I don't see that Penrose offers an argument for the conclusion that the ready acceptance of a newly proved proposition shows that mathematicians see that it *is true* rather than that it *follows from the rest of mathematics,* that is, is true *if* the rest of

accepted mathematics is. Penrose rightly emphasizes that we must see that each step in an argument can be reduced to something simple and obvious. But such reduction may not be possible: Many regard impredicative comprehension axioms in analysis as neither simple nor obvious; and none of the axioms of set theory forces itself on us the way "x + 0 = x" does.

"When we convince ourselves of the validity of Gödel's theorem we not only 'see' it, but by so doing we reveal the very nonalgorithmic nature of the 'seeing' process itself." (*Emperor*, p. 418) Since one of the hypotheses of Gödel's theorem is the consistency of the theories under consideration, Penrose must here mean seeing the truth of the Gödel sentence; but I have argued that we cannot do this if the theory is a reasonable approximation to the whole of mathematics.

The Mandelbrot set has been called the most complex object in all of mathematics, but mathematics itself, of course, outstrips the Mandelbrot set in complexity. Can we really "see" that "0 = 1" is not sitting at the bottom of some lengthy, intricate, and ingenious proof perhaps involving concepts and arguments of a kind of which today we are completely unaware?

Algorithms and physical laws

Franklin Boyle

Center for Design of Educational Computing, Carnegie Mellon University, Pittsburgh, PA 15213

Electronic mail: *fb0m@andrew.cmu.edu*

Penrose says he believes that insights and judgments, which are part of conscious thought, are nonalgorithmic and so cannot be captured by a computer program. His claim, at odds with the "strong-AI" position that cognition is computation, is that there is no "mental algorithm" that could exhibit these sorts of behaviors. He asserts that what is needed to come to grips with the concept of "mind" is a better understanding of the fundamental laws of physics, that there are as yet undiscovered physical laws that may describe the (presumably nonalgorithmic) physical actions that underlie conscious thought. Though Penrose's central concern, as Martin Gardner states in the foreward, "is what philosophers call the 'mind-body problem'," he fails to bring us closer to solving it. Instead, he merely restates it, supplanting strong-AI's modern-day dualism, that "mindstuff . . . is the logical structure of an algorithm" (p. 21), with the problem of relating nonalgorithmic processes, such as discovering mathematical truths (*e.g.*, insight), "to the 'real' world of actual physical objects" (p. 430). Furthermore, this restatement carries with it the assumption that there *are* physical laws, operative in the brain, that give rise to nonalgorithmic behavior.

The importance of Penrose's book for cognitive science therefore depends entirely on his refutation of the computational view of mind. Yet one might question some of his arguments against it. For example, when he considers the computability of word pairs in problems like the *word problem*, he says of word inequalities that "there is no such obvious algorithm, in general, for deciding when two words are *not* "equal," and we may have to resort to 'intelligence' in order to establish that fact." (p. 131). What about learning? Is that what he means by "intelligence?" He never addresses the issue of machine learning (in fact, I do not believe "learning" is mentioned anywhere in the book, though "insights," which "must lie outside *any* algorithmic action" (p. 110) could be interpreted as inductive learning). Rather than investigating algorithms for learning particular problem-solving "tricks," he confines himself to considering whether *general* algorithms exist for solving certain classes of problems. He seems to assume that any sort of "mental algorithm" would be static (as if it were innate), restricted to function according to some predefined formal system of axioms and rules of inference. Whether or not the capacity of computational devices to learn algorithmic procedures is sufficient for them to exhibit the kind of "intelligent" behaviors Penrose is interested in is left in question simply because he never addresses that issue.

His consideration of newly discovered (or existing) physical laws as explanations for mental behavior is of little consequence for cognitive science precisely because he offers no principled connection between such laws and mental behavior (algorithmic or nonalgorithmic). What role does lawful behavior play in thinking? Solving the mind-body problem requires an understanding of the physical changes responsible for the brain's capacity to function as a mind, that is, changes that *are* the processing of information. Not all changes in the brain are informational, however, hence the need for a principled connection. Without it, Penrose is forced to make analogies based on functional similarities that probably bear little resemblance to more fundamental relationships between mind and brain. This has the effect of making his presentation appear to be little more than a set of loosely related ideas. For example, even though one of Penrose's goals is to develop an objective view of quantum state vector reduction and even though he says that "I am not at all happy with [the subjective view]" (p. 295), he nevertheless seems to have arrived at the idea of relating quantum phenomena to conscious thought, at least in part, because the subjective view of the state vector involves *conscious* observers.

It is clear that "algorithm" (or "nonalgorithmic process") and "physical law" are central for Penrose. Unfortunately, incorporating both of these notions, as they are typically used, in a theory of cognition necessarily leads to dualism. For Penrose this is evident from such statements as, "I am somewhat disconcerted to find that there are a good many points in common between the strong-AI viewpoint and my own" (p. 429), and, "When one 'sees' a mathematical truth, one's consciousness breaks through to this [Platonic] world of [mathematical] ideas" (p. 428). To break this dualism, as a first step toward solving the mind-body problem, "algorithm" and "physical law" must be framed so they become equivalent with respect to descriptions of the physical changes that underlie information processing, in contrast to Penrose's generation of physical explanations by analogy. Algorithms (and presumably nonalgorithmic processes) are typically associated with the processing of information. Physical laws, on the other hand, describe physical constraints on the behavior of physical objects. Instead of viewing algorithms (or nonalgorithmic processes) in terms of the kinds of behaviors they are capable of producing, he should ask about the physical characteristics of algorithms (or nonalgorithmic processes). And instead of viewing physical laws as describing the classical or quantum mechanical changes in the states of physical objects, he should ask how the laws of physics could be used to describe information processing.

An algorithm is an "effective" procedure describing a set of mechanical operations that can be carried out by a Turing machine. This abstract mechanism is physically realized in digital computers (barring the infinite tape). It can be inferred from this that algorithms, too, have been given a physical reality. Even when they are physically instantiated (as computer programs), however, algorithms are almost always considered in terms of their functional characteristics, largely because of the lack of analysis, or recognition of the importance of an analysis, of the physical characteristics of structures capable of carrying out the individual steps. For Penrose to solve the problem of dualism he must determine the physical changes that underlie mental behavior and compare them to the physical changes that underlie the processing of information in the computer to conclude whether or not the brain is truly algorithmic or only able to be described algorithmically at some level. Without some kind of physical grounding for algorithms (or nonalgorithmic processes), the true nature of conscious thought can only be assumed. Penrose simply uses introspection (his experi-

encing mathematical insight) as the basis for his idea that conscious thinking is nonalgorithmic, and so is unconstrained in considering the reality of the Platonic existence of ideas.

Physical laws are relationships between measured attributes that are used to describe the state of a physical system. Though such laws are physically grounded, what effect do they have on information processing in the brain? In general, physical laws do not directly describe information processing. For example, in a digital computer information is not embodied by the values of the measured attributes of its components (as it is in analog computers). This is true whether the physical behavior of the digital computer is described classically or quantum mechanically. Rather, the information is embodied in *combinations* of "1's" and "0's" (high and low voltages, respectively). Such combinations are *not* measured attributes. Thus, if we believe that the brain is an information processor (whether or not it is algorithmic), we must carefully consider the role that physical laws play in its behavior. Otherwise, because of its truly remarkable mental capacities (as Penrose notes), one may feel compelled to look for such equally remarkable physical behavior as that described by quantum mechanics or even more speculative physical theories. Instead of searching the frontiers of physics, Penrose might well consider physical characteristics other than laws, such as boundary conditions (to which he devotes only two pages), that might play an important role in fashioning the mind.

AI and the Turing model of computation

Thomas M. Breuel

Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139
Electronic mail: *tmb@ai.mit.edu*

Penrose argues that a theory of the mind requires an understanding of new (hitherto undiscovered) physical principles. He then asserts that the field of artificial intelligence (AI) cannot yield such a theory because it insists on testing its hypotheses with computational devices that are less powerful than physical systems based on the postulated new physical principles.

Penrose admits that he has put up a "straw man" when he describes the AI researcher ("AI" specifically refers to the information processing or computational approach described by D. Marr in his 1980 book, *Vision*) as dogmatically clinging to the notion that the brain is nothing but a digital computer. The reason AI researchers, for practical purposes, adhere to the idea that brains are no more than computational devices is not philosophical stubbornness but the fact that no physical process is known to exist that can be used to build a device computationally more powerful than a Turing machine, and no concrete theories of psychological and cognitive phenomena have so far required any recourse to physical mechanisms that were more powerful than a Turing machine. Penrose's argument may be cautious first steps towards changing both of these facts, but I feel they are still much too tentative and informal to require serious reconsideration of the marriage of AI and the Turing model of computation.

AI considers the brain as a system that processes information. In this view, the senses convert external physical states into a representation (or "code") that is processed by the brain, which produces other (output) representations that are then realized physically by effectors such as muscles.

The AI description of the brain as an information processing system is complementary to the description as a physical system used in the neurosciences. Both are equally valid and independent descriptions of a single phenomenon. This relationship between AI and the neurosciences is somewhat analogous to the relationship between organic chemistry and physics: Organic chemistry abstracts away many of the physical principles underlying chemical reactions and is in return able to make more powerful generalizations.

It is important to realize that the view of the brain as an information processor does not imply that the brain is no more powerful than a Turing machine. For example, the input and output representations could be infinite and continuous rather than finite and discrete, and more powerful models of computation than the Turing model might be used in that case to model the information processing that occurs in the brain.

Why is it then that most AI theories involve computers in some way? First, AI theories often involve descriptions of how discrete information is processed, and a Turing machine algorithm is a convenient means of describing many such information processing tasks unambiguously and concisely. But there are also many AI theories which are not initially formulated as algorithms. For example, many low level algorithms in vision are formulated as networks or differential equations. Any algorithmic implementation of such schemes is used exactly the same way it would be used in the physical sciences: to test the theory numerically and make predictions that are experimentally verifiable. Finally, because AI makes theories about the information processing abilities of physical systems, it is necessary to establish the physical realizability of an AI theory, and giving a Turing machine realization of an AI theory is a convenient means for proving physical realizability.

The language and approach of AI would, if necessary, extend to information processing that is more powerful than Turing machines, however: Networks, circuits, and differential equations are not intrinsically restricted to Turing equivalent computation; the Turing model itself can easily be made more powerful by introducing, for example, "oracles" or infinite precision arithmetic primitives.

Even staying within the confines of Turing realizable theories when developing theories of the mind is not as severe a restriction as Penrose seems to suggest in his book. For example, Penrose claims (p. 415: "Moreover, the slightest 'mutation' . . .") that computations (in the Turing sense) are inherently not robust. This statement is true only of some particular realization of a Turing machine, however. Within computer science, there exists a large body of work on fault-tolerant computation (i.e., fault-tolerant realizations of the Turing model of computation). Likewise, in many artificial neural network models that are no more powerful than Turing machines, small changes in the program (i.e., the weights of the network) result only in small changes in behavior. Other purported limitations of Turing machines are lack of adaptivity ("they follow a fixed program") and absence of randomness; but, again, a closer analysis shows that such claims are unfounded.

The earliest attempts at AI were nonadaptive symbolic processing systems that had access to virtually no sensory information. Such systems can be very useful for testing theories about certain very specific high-level processes (e.g., syntactic analysis). Symbolic systems of this kind, however, are limited in the kind of intelligent behavior they can display. Theories of highly complex tasks like mathematical or social reasoning will probably require adaptivity, versatile architectures, as well as an understanding of how different information processing modules are coordinated, how cognitive development and learning work, and how human intelligence is linked with, and dependent on, social context. None of these questions is currently understood sufficiently well to even begin formulating concrete, testable information processing theories. Until we have had the opportunity to observe that the Turing model is insufficient for formulating and testing such theories, it seems premature to postulate the necessity for more powerful information processing mechanisms.

Lucas revived? An undefended flank

Jeremy Butterfield
Philosophy Faculty, Cambridge University, Cambridge CB3 9DA, England

What a marvellous book! I discern three main ingredients.

1. The best kind of popular science: not just detailed and clear, but also forthcoming about unresolved issues. Setting aside the better-known issues in the foundations of quantum theory, examples include: the distinction between "good" and "bad" uses of Cantor's diagonal argument (p. 111); the recursiveness of the Mandelbrot set (p. 125); self-energy in classical electromagnetism (p. 189); determinism in general relativity (p. 215); complexity theory and quantum computers (pp. 145, 402).

2. Various controversial arguments, mostly against strong AI ("the mind is a digital computer"). The main argument here is based on the nonalgorithmic nature of mathematical insight, allegedly shown by Gödel's theorem (especially pp. 108–12; 417–18).

3. An overarching speculation that two disparate problems – the reconciliation of quantum theory with relativity, and the relation of mind to body – are relevant to one another. This is filled out in various ways. The most striking is by a happy analogy with Penrose's work on tiling and quasicrystals: A thought that surfaces in consciousness is both one of many previously unresolved alternatives (cf. the reduction of the state-vector, and quantum computers), and the solution to a problem, involving global interactions of a characteristically quantum kind, as the growth of a quasicrystal might be (pp. 434–39; 446).

For such a *tour de force*, criticism of specific points is bound to seem niggling. But better that than panegyric. And better that than just scepticism about the speculations: That would be no news to Penrose, who always expresses them cautiously. So I take up two major, and then two minor, points.

First, I am not convinced that Penrose's "Gödel" argument against strong AI avoids the objections against his precursor, John Lucas (1961). (Penrose cites some: I would urge adding Lewis 1969; 1979.) Conscious of these objections, Penrose makes a final attack (pp. 417–18). Transposing the argument to Lucas's terms, it is: If Lucas's arithmetical output is that of a Turing machine, then the machine table must be so complex that Lucas cannot survey it to check that it delivers only truths. (For if he could, then he could "defeat" his own table by constructing its Gödel proposition.) But this is incompatible with the fact that in mathematics "we do not bow down to the authority of some obscure rules that we can never hope to understand. We must see that each step . . . can be reduced to something simple and obvious." (p. 418) Contraposing, Penrose denies that Lucas's arithmetical output is that of a Turing machine. I reply: The "but" is a non sequitur. Unsurveyable complexity of the machine table is, of course, compatible with mathematics' rigourous standards of proof.

Penrose's second argument against strong AI is based on the phenomenon of having "in a flash" a complex thought (pp. 418–23); and his speculation that this is connected to state-vector reduction and quasicrystals. Penrose is mainly concerned with mathematical thoughts. Indeed, he eventually says that he takes the essence of consciousness to be the "seeing" of such a necessary truth as logic and mathematics provide (p. 445). This use of "consciousness," though unusual, would be harmless were it not for the fact that Penrose briefly argues that other phenomena more usually associated with "consciousness" are a threat to strong AI, namely, qualia (pp. 14, 447), personal identity, and indexicality (pp. 27, 409, 448). The brief treatment of these threats engenders two problems. (1) You can get the impression that Penrose's argument involves a unitary notion of consciousness tying all these phenomena together. Not so: As far as I can see, Penrose says nothing against the "divide and rule" idea that "consciousness" is an umbrella term, all these phenomena being logically, indeed nomically, independent. That is, a being could have mathematical thoughts having neither qualia nor indexical thoughts, and so on. (2) Since these threats are much debated in the philosophical literature, Penrose has an undefended flank: Might not the materialist philosophers rebut his argument from the phenomenology of mathematical insight, in much the way they rebut the argument from qualia (e.g., Lewis 1990)?

Two minor points. (1) Whatever consciousness is, it is a non sequitur to infer (p. 408) from its having evolved to its having a selective advantage, and so an active role. It might be a necessary or nomic concomitant of something with such advantage, that imposes no or such little enough disadvantage as the weight of a polar bear's warm coat (cf. Jackson 1982). (2) It is a non sequitur to infer from the timelessness of mathematical truth to there being no threat of causal paradox in the transmission of mathematical beliefs, backward in time (p. 446). Even if the truths are timeless, beliefs in them (and if distinct: their physical correlates in brains) are in time. So such transmission threatens paradox, as backward causation usually does. [See Libet: "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action" *BBS* 8(4) 1985.]

Computing the thinkable

David J. Chalmers
Center for Research on Concepts and Cognition, Indiana University, Bloomington, IN 47405
Electronic mail: *dave@cogsci.indiana.edu*

The main thesis of Penrose's book is that mental processes might be nonalgorithmic. There appear to be three different arguments for this conclusion, which I will present in stripped-down form.

1. The argument from introspection. (1) Some mental processes are not algorithmic at a conscious level, therefore: (2) Some mental processes are not algorithmic.

If this statement of the argument seems a little bald, it is difficult to imagine what else might be meant by the numerous appeals to "intuition" and "judgment" (pp. 411–15; 418–23). It is clear that a premise is missing here. Penrose wishes to exclude from the start the possibility of conscious mental processes that are algorithmic at a level too low to be apparent to conscious introspection. This is a dangerous assumption, as the recent proliferation of connectionist models demonstrates. These models have made familiar the notion that the level at which a system is algorithmic might fall well below the level at which the system carries semantic interpretation (Smolensky 1988). It is not a huge leap to image that in many systems, including the human brain, the computational level might fall below the conscious level.

Connectionist models are not explicitly considered in the book under review, but on the face of it they would seem to fall into the class of "computational" models that Penrose would like to dismiss. It would be interesting to see Penrose declare an explicit position vis-à-vis these models. If he exempts them from his criticisms, then the force of his critique of algorithmic models is considerably weakened; if he wishes to dismiss these models, too, his arguments will need to be considerably strengthened.

It must be conceded that the connectionist approach has not yet had much success in modelling the kind of temporally extended processing, such as mathematical thought, that Penrose considers. Nevertheless, other work within the "subsymbolic paradigm" has made some progress on these matters. In particular, Mitchell and Hofstadter (1990) have produced an interesting model of perception and analogical thought in an abstract domain. In this model, high-level processes emerge from the interaction of a number of small, low-level agents.

Under the influence of various pressures, the model is able to come up with "insights" that are similar in kind to those of a mathematician. The high-level behavior of the model appears in no sense algorithmic, yet it emerges from a completely computational substrate.

2. The argument from Gödel's theorem. (1) Humans can "see" the truth of certain mathematical statements that lie outside the bounds of any given formal system, therefore: (2) Human mathematical thought is not constrained by any given formal system.

This is an interesting variant on the argument of Lucas (1961). Instead of focusing on the formal systems that specify a particular *machine*, Penrose (pp. 416–18) focuses on the formal systems that might specify our mathematical thought. Because we have the ability to "see" that the Gödel sentence for a given system is true, the argument runs, we are using processes outside the system. On page 418, Penrose states: "When we convince ourselves of the validity of Gödel's theorem we not only 'see' it, but by so doing we reveal the very nonalgorithmic nature of the 'seeing' process itself."

This seems fallacious. We do not have to invoke any mystical processes to explain this step; we do not even have to invoke consciousness, as Penrose suggests. The reason we can "see" that Gödel sentences are true is simply that we have a built-in faith that our mathematical systems are *consistent*. It would not be a difficult matter, in principle, to build such faith into an algorithmic machine. (And if Penrose would wish to argue that, unlike machines, humans can repeat the "Gödelization" process *ad infinitum, ad transfinitum,* the reply is that in practice the Church-Kleene result on enumerating constructive ordinals puts as many limitations on humans as it does on machines. We are finite creatures, and we cannot continue to the ultimate Omega.)

To gain his *reductio* of the notion of algorithmic thought, Penrose postulates a single algorithm for determining mathematical truth, shared by the mathematical community. Even to one who believes that mind is algorithmic, this seems a little strange. If we stay within the usual bounds of number theory, analysis and the like, such an idea is perhaps plausible. As soon as we move beyond these into more abstract strata of set theory and logic, disagreement about "truth" becomes rife. Some mathematicians "see" that the axiom of choice is true; others "see" that it is false. Moving further out, the continuum hypothesis and the axiom of constructibility are still more controversial. If such a "universal" algorithm exists, it is a fuzzy thing indeed; it becomes less and less universal the further we travel from the commonplace. This fuzziness alone is enough to defeat Penrose's argument: A fuzzy algorithm cannot be Gödelized!

3. The argument from physical processes. (1) At the lowest level, physical processes might not be algorithmically specifiable. (2) Mental processes are dependent upon physical processes, therefore: (3) Mental processes may be nonalgorithmic.

This is an ambitious argument, but one which must hold if Penrose's other conclusions are to be sustained. It is nothing but an attempt to subvert the force of Church's thesis about the universality of algorithms. There are two clear weak spots. First, even if (1) holds, it would still be far from clear that such microscopic nonalgorithmicity should make any difference on a macroscopic level. It seems plausible to hold that even if *electrons* don't behave algorithmically, *neurons* still might. Penrose acknowledges this gap, but does little to bridge it. Second and more serious, it seems to me that Penrose has in fact provided very little evidence for (1). He gives an impressive demonstration of the nonclassical, nonintuitive nature of microscopic physical phenomena, but he gives no clear justification of why these things should have any bearing on their *algorithmicity*. For example, physical processes may well be nonlocal, but algorithms were never committed to *locality* in the first place. Algorithmic specifications have many degrees of freedom. Although the final verdict will be determined empirically, I doubt that Church's thesis will give in easily.

The idea of algorithmic processing lies at the core of modern cognitive science for good reason. Anyone who succeeds in overthrowing this idea will have effected a deep conceptual revolution in the way we think about the human mind. Penrose has given it his best, and has written a fascinating book along the way, but his arguments are a little thin for the weight they have to bear.

Is mathematical insight algorithmic?

Martin Davis

Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

Electronic mail: *davism@csd11.nyu.edu*

Roger Penrose replies, "No," and bases much of his case on Gödel's incompleteness theorem: It is *insight* that enables us to *see* that the Gödel sentence, undecidable in a give formal system is actually true; how could this *insight* possibly be the result of an algorithm? This seemingly persuasive argument is deeply flawed. To see why will require looking at Gödel's theorem at a somewhat more microscopic level than Penrose permits himself.

It will be helpful (though not essential to our argument) to place the discussion in terms of what is usually called *first order logic*. This is just the formal system that embodies the elementary classical logic of *and, or, not, implies, all, there exists*. In a precise formulation of first order logic, it is necessary to explain when some particular formula F is to be taken to be a *logical consequence* of a set of formulas ("premises") Γ. This can be done in two essentially different ways: *semantically* and *syntactically*. In the semantic version, F is a logical consequence of Γ if F is *true* no matter how the extra-logical symbols appearing in F and Γ are interpreted, so long as all the formulas in Γ are true under that same interpretation. (Metaphorically: F is true in every Platonic world in which the formulas of Γ are true.) In the syntactic version, "rules of proof" involving the straightforward manipulation of symbols are specified, and F is said to be a logical consequence of Γ if F can be obtained from Γ by some finite number of applications of those rules (Penrose, p. 104 gives some samples of such rules). In Gödel's 1929 doctoral dissertation, he establishes his famous *completeness* theorem, which states that the semantic and the syntactic versions are equivalent. Moreover, this equivalence is largely independent of the detailed manner in which rules of proof are specified.

Gödel's completeness theorem answered a question Hilbert had posed in his address at the Bologna mathematical congress of 1928. Hilbert's Entscheidungsproblem for first order logic was also raised in 1928 (in the famous textbook by Hilbert & Ackermann (1928), not at the Bologna conference as Penrose asserts), and called "the fundamental problem of mathematical logic." The problem was to give an algorithm for deciding whether a given formula was a logical consequence (in the semantic sense) of a given (finite) set of premises. Hilbert singled out first order logic for this attention presumably because it seemed clear that all mathematical reasoning could *in principle* be carried out in this formalism.[1] For the premises one takes an *appropriate* set of mathematical axioms; a mathematical theorem is then simply a logical consequence in first order logic of those axioms. Since an argument based on the rules of proof of first order logic can be checked in a completely algorithmic way, we have no trouble understanding why mathematicians should agree about proofs (p. 417) so long as they agree about the axioms (and so long as these axioms are finite in number or at least are specified by an algorithm).

In this context, Gödel's incompleteness theorem (in a strengthened form based on work of J. B. Rosser as well as the solution of Hilbert's tenth problem) may be stated as follows:

There is an algorithm that, given any *consistent* set of axioms,

will output a polynomial equation $P = 0$, which in fact has no integer solutions, although this fact cannot be deduced from the given axioms.

Here then is the true but unprovable Gödel sentence on which Penrose relies and in a particularly simple form at that. Note that the sentence is provided by an *algorithm*. If *insight* is involved, it must be in convincing oneself that the given axioms are indeed consistent, since otherwise we will have no reason to believe that the Gödel sentence is true. But here things are quite murky: Great logicians (Frege, Curry, Church, Quine, Rosser) have managed to propose quite serious systems of logic which later have turned out to be inconsistent. "Insight" didn't help. New axioms are just as problematical as new physical theories, and their eventual acceptance is on similar grounds. If the underlying axioms are just the elementary axioms for the arithmetic of natural numbers (Peano's arithmetic), then the consistency is readily proved by unproblematical mathematical methods (which however, by Gödel's result, must go beyond what can be done with these axioms). In this case, we can indeed have confidence that the corresponding equation $P = 0$ has no solutions. If the underlying axioms are what are known as ZFC (Zermelo-Fraenkel set theory), axioms known to be adequate for all ordinary mathematics, most mathematicians today would accept that the corresponding $P = 0$ has no solutions, but hardly on the basis of a sudden insight. Confidence in ZFC has developed slowly over almost a century. If ZFC is augmented by various axioms that have been proposed (e.g., axioms asserting the existence of very large infinite sets or the so-called axiom of projective determinacy), most experts would be very cautious about accepting that the corresponding $P = 0$ had no solutions.

There is certainly room for disagreement about whether the processes by which mathematical (or physical) theories are developed and accepted are algorithmic. Gödel's theorem has nothing decisive to contribute to the discussion, however.

A final comment on the foundations of quantum mechanics. Penrose calls the process by which classical Hamiltonians are converted into the corresponding quantized versions by replacing real "observables" by operators as "genuine magic which works" (p. 288). I have speculated elsewhere (Davis 1977) that another contribution of modern logic from Gaise Takeuti, and ultimately going back to the seminal ideas of the Norwegian logician Thoralf Skolem, is what is needed to make sense of this "magic."

NOTE

1. It is only in this case that Penrose is justified in somewhat loosely defining the Entscheidungsproblem as referring to "all the problems of mathematics" (p. 34). Penrose's conflation of the Entscheidungsproblem with Hilbert's 10th problem of 1900, which merely asked for an algorithm for the solvability of Diophantine equations, is likewise justified after the fact: It is a corollary of the methods used to give a negative solution to Hilbert's tenth problem that the question of whether any given Turing machine will eventually halt, and hence the Entscheidungsproblem, can be encoded as a Diophantine problem (Davis et al. 1976). Of course, Hilbert in 1900 could hardly have imagined such a thing.

Betting your life on an algorithm

Daniel C. Dennett

Department of Philosophy, Tufts University, Medford, MA 02155

What minds can do, Penrose claims, is to see or judge that certain mathematical propositions are true by "insight" rather than mechanical proof. Penrose then argues that there could be no algorithm, or at any rate no practical algorithm, for insight. This ignores an independently plausible possibility: The algorithms that minds use for judging mathematical truth are not algorithms "for" insight – but they nevertheless work very well.

Consider a parallel argument. Chess is a finite game, so there is an algorithm "for" either checkmate or draw: the brute force algorithm that draws the entire decision tree for chess and works backwards from the last nodes. That algorithm surely is not practical. Probably there is no practical algorithm "for" checkmate. There are plenty of practical algorithms that *achieve checkmate with great reliability*, however. They are the chess-playing programs and although none is mathematically guaranteed to achieve checkmate against any opponent, you could safely bet your life that the best of them will always achieve checkmate against me (for instance). There are algorithms for playing legal chess – that is guaranteed mathematically. Checkmate is an unprovable bonus, but it is not a gift out of the blue. It is to be explained in terms of the relative cunning of these chance-taking algorithms. Aside from sheer speed, no other properties of a chess-playing computer – its material composition or genealogy, for instance – would be relevant to its power to achieve checkmate.

The following argument is therefore fallacious:

1. X is superbly capable of achieving Y (e.g., checkmate).
2. There is no practical algorithm for achieving Y. therefore
3. X's power to achieve Y is not explicable in terms of any algorithm.

Therefore, even if mathematicians are superb recognizers of mathematical truth, and even if there is no algorithm, practical or otherwise, "for" recognizing mathematical truth, it does not follow that the power of mathematicians to recognize mathematical truth is not entirely explicable in terms of their brains executing one or another garden-variety algorithm. Not an algorithm "for" intuiting mathematical truth – for the sake of the argument, I will grant to Penrose that there can be no such algorithm – but an algorithm for something else. What? Most plausibly it would be an algorithm – one of many – for *trying to stay alive*, an algorithm that, by an extraordinarily convoluted and indirect generation of byproducts, "happened" to be a superb (but not foolproof) recognizer of friends, enemies, food, shelter, harbingers of spring, good arguments – and mathematical truths.

Chess programs, like all heuristic algorithms, are designed to take chances, and therein lies their vulnerability in principle. What are the limits of vulnerable-in-principle probabilistic algorithms running on a parallel architecture such as the human brain? Penrose neglects to provide any argument to show what those limits are; hence he fails to cut off the most plausible rival interpretation of the mathematicians' prowess, on which his whole case depends. Notice that it is *not* a question of what the in-principle limits of algorithms are; those are simply irrelevant in a biological setting. To put it provocatively, an algorithm may "happen" to achieve this 999 times out of 1,000, in jig time. This prowess would fall outside its official limits (since you cannot prove, mathematically, that it will not run forever without an answer or else give a false answer), but it might nevertheless be prowess you could bet your life on. Mother Nature's creatures do it every day.

Sometimes Penrose suggests that what human mathematicians do is something that could not even be approximated by a heuristic, mistake-prone algorithm, since mathematicians (in principle? always?) settle into a consistent shared view. If they make a mistake, they can (will?) always correct it. Is this supposed to be an independently confirmable empirical premise? This could not be proven mathematically, of course, for such consistency proofs of oneself (or oneselves acting in concert) are ruled out by the very mathematical results Penrose relies on. He can perhaps fervently believe, and assert, that the joint or Ideal Mathematician is consistent and capable (in principle) of intuiting every mathematical truth (and no falsehoods), but he cannot hope to persuade those of us who find this an unlikely and unmotivated dogma by offering a mathematical proof, and there seems every empirical reason for simply disbelieving it. Penrose's envisaged revolution in physics

may happen, but not – so far as I can see – because it is needed to explain any fact or phenomenon of human mental powers.

ACKNOWLEDGMENT
This commentary is a revision of material contained in my review of Penrose's book (Dennett 1989, pp. 1055ff).

Perceptive questions about computation and cognition

Jon Doyle
Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139
Electronic mail: *"Doyle@zermatt.lcs.mit.edu*

Roger Penrose's book offers the reader a valuable perspective on the nature of physical reality and some of its possible implications for AI, computation, and the philosophy of mind. It is worth reading for the survey of physics alone. But the point of the book is to dispute the idea that "our thinking is basically the same as the action of some very complicated computer" (p. 447) by giving two arguments (one from observation, the other from physics) for the claim that "the *conscious* mind cannot work like a computer, even though much of what is actually involved in mental activity might do so" (his emphasis; p. 448). This brief review confines attention to these two arguments. Though we find that our knowledge of physics and psychology is not yet complete enough to tell whether conscious mental processes are computable, one of the great virtues of this book is that it raises this question technically, clearly, and unavoidably.

Penrose's primary argument is that conscious thought involves seeing or intuiting necessary (mathematical) truths, and that mathematical truth is not formalizable, hence it cannot be determined by computers. He claims that mathematicians have direct access to mathematical truth since many mathematicians (myself included) have the distinctive experience of mentally "seeing" mathematical objects laid out as landscape before them.

Penrose's argument fails to differentiate the ultimate powers of people and machines because the relevant limitation of computers is that they cannot determine *all* mathematical truths, not that they cannot determine *any*. As Penrose admits, however, even the mathematician's conceptual vision is limited: Not all truths are visible. Such limitations are not surprising, since most individual mathematical truths could not even be written down using paper the size of the universe and characters the size of protons. Penrose notes that mathematicians can use the method of reflection to resolve particular questions left open by specific formal theories, that is, by observing the results and limitations of the theories. He seems to think that such inferences are not mechanizable. But many of these reflective observations, which are epistemologically similar to observations of objects in the physical environment, can be automated as easily as ordinary deduction rules. (The problem Penrose cites of choosing the right reflections to perform is, as a practical matter, not more difficult than the problem of choosing the right ordinary inferences to draw. Both choices can be difficult.) If we are to suppose that ideal mathematicians can discover recursively enumerable sets of truths derived from finite sets of axioms, axiom schema, and inference rules, including reflection principles, we must suppose that computers can do this, too.

Penrose believes humans are not limited to enumerable truths, however, and presents a *reductio ad absurdum* as the crux of his argument that mathematical insight is not algorithmic. In short, the assumption that mathematical understanding is captured by some formal system conflicts with our ability to recognize the truth of a Gödel sentence unprovable in that system. The critical hypothesis of Penrose's argument is that all mathematicians agree on a notion of mathematical truth and that this shared notion of truth does not change as they learn and reflect on proofs. But the only support he provides for this hypothesis is that mathematicians will generally agree on proofs once they learn of them (whether by thought or by communication): "*When* we comprehend them [mathematical demonstrations], their truth is clear and agreed by all" (p. 418, emphasis added). But this hardly rules out mathematical understanding evolving with new information and experience in universal, even algorithmic, ways. If this is possible (and it is almost an accepted axiom in studies of machine learning), there is no reason to assume that the formal system used to contemplate a Gödel sentence is still the one the sentence is about, and the argument falls apart. Indeed, intuitionist mathematicians contend that changes (not necessarily algorithmic ones) do occur in mathematical understanding, and Penrose's explicit dismissal of their views seems to beg the question. Perhaps the great and clear limitations of human mathematical vision are less limiting than the limitations suffered by computers, but Penrose does not demonstrate this.

Penrose's secondary argument for his thesis is indirect. He argues that thinking is the activity of physical brains, and nothing in the laws of physics as we understand them today ensures that this sort of physical activity is computable. His argument consists of a lengthy but superb survey of the major physical theories in which he points out the numerous ways they do not guarantee computability (or even determinism and locality). Penrose's silence on the topic of relative computability (that is, algorithms over operations other than Turing machine steps) is especially disappointing here, because his ideas suggest attempting to design specific physical mechanisms that realize simple Turing-uncomputable functions (for example, that solve Diophantine equations) for use as "oracles" by digital computers.

One need not accept Penrose's more speculative suggestions about physical reality to realize that there is a real possibility that brain dynamics are not computable. Penrose does not demonstrate that brain dynamics are actually uncomputable, however, and even if they are uncomputable, he does not demonstrate that this entails uncomputability of any mental processes. The differential equations describing a flip-flop, for example, are probably uncomputable, but the digital computations performed by some systems built from flip-flops are perfectly computable nonetheless.

The book exhibits several minor flaws. Contrary to Penrose's belief, AI *does* employ most of the steps of cognitive processes he identifies as reasonable (including reflection and highly limited forms of "consciousness"), and his argument that human judgment is nonalgorithmic fails to compel because it does not restrict use of "nonalgorithmic" to the technical sense of "no algorithm exists," but mingles this sense with senses involving feasibility, discoverability, and comprehensibility.

Even though his main claim remains unsubstantiated, Penrose deserves our thanks for writing this book about the physical basis of psychology and computation. It is rare for a single book to open so many important questions to technical investigation.

Computations over abstract categories of representation

Roy Eagleson
Centre for Cognitive Science, The University of Western Ontario, London, Ontario, Canada, N6A 5C2
Electronic mail: *elroy@uwo.ca, elroy@uwovax.bitnet*

Penrose has produced an accessible, information-packed compilation over a wide-ranging selection of difficult topics. One

could not help but praise his lucid exposition, were it not for his puzzling commitment to characterize, in an anecdotal style, certain such ill-defined phenomena as intuition and creativity as being nonalgorithmic. To gather support for this claim, the book lists a number of mathematical models for physical and computational processes that inevitably fail to model the world when pushed past their limits. In a scientific tradition, classical theories are eventually labelled approximations to more general explanatory mechanisms which underlie a more global structure; but ideally, both are equivalent within a certain domain. Penrose seems to imply that a similar transformation will inevitably befall "computation" as a model of cognition. He argues that, while some parts of our mind function like computer programs, there is a higher level, which must be described as nonalgorithmic.

"Strong AI" researchers are committed to the idea that algorithms represent the way flexible behaviours can be exhibited by a system and implemented in its fixed physical architecture. This commits them to descriptions that involve operations and transformations on internal representations, formed by inputs, internal states, and rules that specify the transitions between these states. "Intelligence," in this view, is a characteristic of the semantic level of the algorithms. It is a property of a computational process implemented on a physical symbol system (cf. Newell 1980; Pylyshyn 1984).

There is another well-accepted principle in strong AI: the utility of abstract levels of representation, cf. Albus (1981); Saltzman (1979); Simon (1973). The universe may have a single objective structure, but we cannot fully represent it because of our limited resources and its vast complexity. The choice of attributes that can be applied to any single conceivable object is immense! An agent can therefore make only partial observations over an accessible range of arbitrary qualities and detail. It can also strive to observe "invariants and regularities" in these input data. They are processed by operations which seek to extract information relevant to the tasks of the system, requiring higher-level representations that encode the agent's goals and beliefs. This stratification of data types might begin with abstractions of the properties of local coherence across a sensory manifold, but it can include more abstract relations that encode qualities not tied to local physical properties. The freedom to develop these functional relationships between abstract data types permits an agent to represent a rich variety of concepts.

In addition, because of noise and ambiguities in the measurement process, there are inherent uncertainties about the relationship between the external world and its internal representation. An agent must therefore act to reduce the uncertainty associated with these abstract internal models. Multiple observations over space and time can reduce this uncertainty as additional information is made available. Such techniques make up much of the current literature in computational perception.

In robotics, low-level control systems are designed with the goal of minimising the difference between measurements of an external physical process and an internal representation of a desired state-space trajectory. The transitions between these states are usually given by differential equations that describe a model of the system in its environment. If these internal laws are incapable of accurately and tractably modelling a more general world, its inherent metric structure is not useful. What remains, however, can be described as a network of discrete event transitions, preserving the topology of the original problem space (cf. Caines et al. 1989). This is an example where a continuous problem, specified by an abstract differentiable manifold, has a higher abstract level that preserves its topology. A category of logical control can be specified on this space, retaining its deterministic (namely, algorithmic) character.

Within the framework of strong-AI, information is purposefully processed by transformations on internal representations, at the level where models exist as symbolic data structures. The algorithmic nature of this type of information processing is limited by the tractibility of a problem. If the environment becomes too complex, it is appropriate to design a method that allows additional direction to be provided from a supervisory level. They are manipulated and transformed by algorithms represented at this level by rules, and implemented on the cognitive architecture. Consciousness might be viewed by strong AI proponents as being at an abstract level higher than logical operations applied to symbolic pointers to representations of objects in the world, but they would still regard this level as having a formal representational structure (see Newell 1982). From the limited perspective of a lower level, this higher level of control can only be regarded as mysterious goals and beliefs, or perhaps heuristics. They would still be algorithmic over a different category of more abstract symbol structures, however.

Penrose contends that there is something unsatisfying about an algorithmic description of the mind's function, but his anecdotes have exactly this unsatisfying quality. Mathematical insight and "flashes of intuition" require that a person have considerable experience in a subject and thereby have very elegant mental structures for manipulating complex expressions. It should not be surprising that once appropriate relationships between these internal structures have been identified, it is almost a mechanistic exercise to report the results mathematically by finding the most descriptive sentences in terms of the rules and procedures of mathematical literature. "Self-awareness" may similarly be posited as a property of recursive computational functions.

Penrose's arguments amounts to regarding classical symbol-manipulation as having a higher level, but one that is devoid of an algorithmic structure. In this case, it would have no rules that describe how within-level processes should purposefully be directed to make internal models based on observations of the world. He fails to describe the way these higher levels are supposed to function nonalgorithmically, other than to propose some form of quantum mechanical model of the mind. As Bell (1986) has argued, quantum logic behaves as a logical system in which the order of introduction of premises is not commutative. Thus, a system that relies on quantum logic to reduce NP-complete problems to linearly tractable ones must first have induced a partial ordering on the premise space. Even if such a model were indeed appropriate, Penrose's "little finger" should support the strong AI view, and speak out against it being a nondeterministic system.

ACKNOWLEDGMENTS

I would like to thank John L. Bell and Zenon Pylyshyn for very helpful and stimulating discussions. Support was provided through the Canadian Institute for Advanced Research, the Ontario Information Technology Research Centre, and NSERC operating grant number A2600.

Physics of brain-mind interaction

John C. Eccles

CH 6646 Contra (TI) Switzerland

I commend this challenging book that opens up new vistas on this fundamental problem. It is the first time to my knowledge that a mathematician-physicist has become dedicated to the study of the human brain. It is certain that the conceptual advances in the brain-mind problem can be accomplished only by those who have a deep understanding of the brain, not in all its immense multifarious detail, but by concentration on that component, the cerebral cortex, that is becoming generally recognized as the cerebral structure that may be exclusively concerned in the experiences of consciousness in all of its manifestations. The centrencephalic (brain-stem) center of con-

sciousness (p. 382), for example, has been eliminated by the Sperry experiments on commissurotomy patients (pp. 384–86).

My review can begin with Penrose's question on page 402: "Is our picture of a world governed by the rules of classical and quantum theory, as these rules are presently understood, really adequate for the description of brains and minds?"

My reaction is that we have to go on scientifically and philosophically, and we can be greatly encouraged by the progress. There are of course many blind alleys that have enormous attraction to computer technologists, notably the artificial intelligence machines. I agree with Penrose's general rejection of such models of intelligence and consciousness. A related project is to study the properties of assumed neuronal networks, which can be modelled by computer technology, and that may even give an opening to robotics. There is of course no doubt about the almost infinite complexity of neural networks that could be constructed on the basis of the known connectivity of neurons of the cerebral cortex. It is assumed that consciousness emerges from the immensity of cerebral connectivities. Many neuroscientists have optimistically developed concepts of this type to which we have given names, as listed in my paper (Eccles 1986): holistic configurations; distributed neuronal systems; phasic cyclic reentrant signals; dynamic patterns of superstructures; extremely complex dynamic systems of interaction. No clear theory has been developed showing how consciousness could emerge in such systems, however. So consciousness has remained enigmatic neuroscientifically.

Penrose (p. 405) raises the searching question: "What *selective advantage* does a consciousness confer . . . ?" With him I reject panpsychism with its belief in consciousness of inanimate objects and lowly organized life. We have to recognize, however, that higher animals have some conscious feelings resembling simpler versions of what we experience. I would propose that no cerebral system can integrate the immense diversity that is generated by the analytical operation of the cerebral systems, as for example, in all of the prestriate visual cortex of the higher mammals. Yet we know that it is integrated in our unified perceptual experience of the visual world. It is integrated in the mind and apparently not in the brain, as in the mythical "grandmother cell" (p. 388). So the higher animals would have a unified conscious experience from moment to moment, which would be highly advantageous in evolution.

In this respect it is important to distinguish between the consciousness enjoyed by such higher animals as mammals and birds, and the self-consciousness unique to humans. Penrose writes most eloquently on page 406 of the immense diversity and wonder of consciousness and particularly of self-consciousness. Sherrington (1940) gives a rare poetic vision of self-conscious experiences. I quote Popper (1977, p. 120): "The self observes and takes action at the same time. It is acting and suffering, recalling the past, and planning and programming the future; expecting and disposing. It contains in quick succession or all at once, wishes, plans, hopes, decisions to act, and a vivid consciousness of being an acting self, a center of action."

One of the most important properties of self-conscious beings is that they ask questions, and so continuously search for understanding. This wonderful human attribute begins as early as 1½ years with incipient human language. And of course it goes on throughout life. At the higher levels it is the basis of human achievement in science, as Penrose recognizes. In all these attributes, human beings are not to be compared with even the most complex artificial intelligence machines that do not know what they do or why they do it – because they do not ask questions! Only self-conscious beings ask questions. That relates to Penrose's claim (p. 412) that it is this ability to "divine (or intuit) truth from falsity (and beauty from ugliness!) in appropriate circumstances that is the hallmark of consciousness." As Penrose states (p. 412), there is no clear algorithmic process for the insightful handling of the morass of data we are confronted with in real life situations including scientific discoveries.[1]

NOTE

1. In the sections headed, "Is there a role for quantum mechanics in brain activity" (p. 400) and "Beyond quantum theory," there is much of great interest to me; I think I can help in the understanding of the important questions that Penrose raises in these sections. I have recently developed a unitary theory of brain-mind interaction in which quantum physics plays a key role. The initial publication was Eccles (1986), and a much more developed theory is coming out in (Eccles 1990) The theory I offer is specially related to the title of my contribution, "Physics of brain-mind interaction" and to many sections in Chapters 9 and 10 of Penrose's book.

Strong AI and the problem of "second-order" algorithms

Gerd Gigerenzer

Department of Psychology, Universität Konstanz, West Germany
Electronic mail: *sygiger3@dnkurz1.bitnet.*

"In my childhood we were always assured that the brain was a telephone switchboard ('What else could it be?')," recalls John Searle (1984, p. 44). Children today are likely to be told that the mind is a computer program. Roger Penrose, slipping back into the role of the child who dares to question, rejects the "strong AI" claim that "mental activity is simply the carrying out of some well-defined sequence of operations, frequently referred to as an *algorithm*" (p. 17). Penrose argues "that the decision as to the validity of an algorithm is *not* itself an algorithmic process!" (p. 414). Let us call these hypothetical algorithm-checking algorithms, "second-order" algorithms. Penrose cites Turing's proof that no algorithm exists for deciding the question of whether or not Turing machines will actually stop (i.e., whether algorithms will actually work). In this comment, I will add several thoughts of a more pointedly psychological sort that support Penrose's mathematical argument.

Scientific inference. Inference in science (e.g., from data to hypothesis) is a mental activity in which algorithms actually exist. Various statistical (e.g., Bayesian, Fisherian, Neyman-Pearsonian) and nonstatistical (e.g., Platt's strong inference) algorithms have been proposed. As is well known, there is little consensus among philosophers, probabilists, and scientists as to which (formal) algorithm applies to which type of (semantic) problem, or whether to use a statistical algorithm at all (Gigerenzer et al. 1989). (There are, however, such "rituals" as the mechanical null hypothesis testing that goes on in some social sciences.) That is, algorithms for scientific inference exist, but there is no "second-order" algorithm for choosing among them. The basic reason for this disagreement is that the problem of inductive inference has no single solution that commands consensus – it has many, competing ones. Indeed, there is no agreement as to whether the problem has a single solution (even in principle). In our current (and perhaps permanent) state of controversy over this question, there is no algorithm for choosing among algorithms – but scientists nonetheless do somehow choose, and with considerable success. Nor are our choices merely blunt expressions of taste – you like Neyman/Pearson and vanilla, I like Fisher and chocolate, who knows why? We argue with one another, offer reasons for our choices, and sometimes even persuade one another.

Concept ambiguity. An algorithm (a Turing machine) is purely syntactical: It specifies, for instance, that if a machine is in a certain state and has a certain symbol on its tape, then the machine will perform a certain operation such as erasing a symbol on the tape and enter another state. The mind, however, has a semantics, too. In many problems (ones that do not deal with well-defined artifacts) that the mind has to handle, there is no simple one-to-one correspondence between a formal concept and a semantic concept. Here, ambiguity first has to be resolved before an algorithm can be put to work – and such judgments

depend heavily on content and context rather than on formal structure. Can a formal algorithm resolve this ambiguity in the way humans do?

Consider a judge who is a Bayesian and computes the probability that a suspect actually committed a crime by an algorithm known as Bayes' rule. One formal concept in this algorithm is the suspect's *prior probability* (of having committed the crime in question), which needs to be semantically interpreted in each individual case. The ambiguity is not only in the precise number of that probability, but in the *kind of reference class* from which this probability should be taken. Each suspect is always a member of many (usually, an *infinite* number of) reference classes (e.g., single parents, young urban professionals, weight lifters) – and all of them may have widely divergent prior probabilities. From time to time, new, never before thought of reference classes may emerge – for example, after the discovery of a new drug whose use is correlated with a certain kind of crime. Although our Bayesian judge's reasoning contains an algorithm, as we assumed, the judge also has to assess relevance: which reference class to choose, and which others to ignore. It is hard to see how these judgments can be made mechanically by a "second-order" algorithm.

Structural ambiguity. Probabilistic algorithms are based on several structural assumptions (e.g., independence) that must hold in the relevant part of the real world if the algorithm is to be applied validly. Textbook applications, such as "urns-and-marbles" problems, are contrived so that there is a one-to-one correspondence between the structural assumptions of an algorithm and the structure of the problem at hand. Beyond textbook problems, however, we must confront ambiguity about structural correspondence (Gigerenzer & Murray 1987, Chapter 5). Consider the following stories that illustrate how important it is for the mind to check structural assumptions and resolve this ambiguity.

1. You live in Palo Alto. Today you must choose between two alternatives: to buy a BMW or a Jaguar. You use only one criterion for that choice, the car's life expectancy. You have information from a test sample of 100 BMWs and 100 Jaguars, of which 75% and 50%, respectively, lasted longer than 10 years. Just yesterday your neighbor told you that her new BMW broke down. Nevertheless, in your reasoning, your neighbor's case decreases the BMW's prior probability only slightly, from .75 to about .74. So you go ahead and buy a BMW.

It is easy to specify an algorithm for this kind of decision-making. Now look at the same problem, but with a different content.

2. You live in a jungle. Today you must choose between two alternatives: to let your child swim in the river, or to let it climb trees instead. You use only one criterion for that choice, your child's life expectancy. You have information that in the last 100 years there was only one accident in the river, in which a child was eaten by a crocodile, whereas a dozen children have been killed by falling from trees. Just yesterday your neighbor told you that her child was eaten by a crocodile.

If, in your reasoning, the same algorithm is applied again, your neighbor's testimony would make little difference: The prior probability of a fatal accident in the river would increase only slightly, from one to two cases in 100 years. The algorithmic mind would probably send the child to the river. The mind of a parent, however, might use the new information to *reject* the old algorithm, rather than to *apply* the old algorithm to the new information. The parental mind may suspect that the small river world has changed – crocodiles may now inhabit the river. The updating of prior probabilities may no longer make sense, since the events (being eaten or not) can no longer be considered as independent random drawings from the same reference class. A structural assumption of the algorithm no longer seems to hold. From now on, many children may be eaten.

I do not know of any "second-order" algorithm that is capable of performing this checking of structural assumptions of algorithms in the same way the human mind does and with similar ease. Can an algorithm be sufficient to judge whether one and the same information (your neighbor's report) is to be interpreted as an *entry* to a computation or as a *rejection* of exactly the same computation? Even for this simplistic structural problem – two alternatives, only one criterion – there seems to be no general algorithm that can compute for all possible contents (and there are *infinitely* many more besides cars and crocodiles) whether the mind uses the prior probability updating algorithm or not. Nevertheless, in individual cases, we may well be able to make an unequivocal judgment. This situation is analogous to the Turing proof: There is no general algorithm that can compute whether algorithms ever stop, but in the individual case we can immediately "see" the answer.

Throughout this discussion I accepted Penrose's view that a large part of human thinking is indeed algorithmic, and added some psychological reflections to his argument that the decision as to the validity of an algorithm is, at least in part, non-algorithmic. This is not to say that I do believe that most human thinking, be it "second-order" or "first-order," *is* solely algorithmic. Even if the *result* of thinking can be simulated by an algorithm, this does not imply that the *process* of thinking is algorithmic, as John Searle has repeatedly pointed out. If there is a computer algorithm that simulates perfectly the time shown by a mechanical clock, this does not imply that the mechanism by which the clock quantifies time is indeed computing. And whereas AI workers can be content with applications that work – a computer that tells time – we psychologists are still responsible for taking apart the ticking clock.

ACKNOWLEDGMENT

This comment was written while I was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, CA. I am grateful for financial support provided by the Spencer Foundation and by the Deutsche Forschungsgemeinschaft (DFG). I would like to thank Lorraine Daston and Kathleen Much for their helpful suggestions on this comment.

Don't ask Plato about the emperor's mind

Alan Garnham

Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, England

Electronic mail: *alang@epvax.sussex.ac.uk*

Why are so many mathematicians Platonists? In part because the standard alternatives are implausible or otherwise unacceptable, and in part because Platonism appears to solve at least one of two fundamental puzzles about mathematics – Penrose believes it solves both. The alternatives to Platonism are formalism, which in its Hilbertian form foundered on the rock of Gödel's theorem, and intuitionism. Intuitionism, as espoused by Brouwer, is unacceptable to most mathematicians both because it is tainted with psychologism and because it proscribes proofs that they are happy to accept. The two puzzles are: that mathematicians agree about what follows from a set of postulates, even though no single mathematician can work through all their consequences, and that mathematics, in Penrose's "SUPERB" theories, accurately describes the physical world. Platonism explains agreement among mathematicians by claiming that they have access to the same Platonic realm. Given Penrose's skillful debunking of fallacious arguments in the physical sciences, I was disappointed that he missed the obvious flaw in this idea. What is crucial is that mathematicians agree *in practice*. The Platonic realm can be important only insofar as it determines how mathematicians work. They must be guided in the same way by the Platonic forms. So Platonism "explains" agreement among mathematicians by the more obscure idea of agreement in interpreting the Platonic realm. Similarly, Pen-

rose's identification of the real and Platonic worlds does nothing to explain the application of mathematics.

The rejection of Platonism suggests that mathematical concepts exist only in the minds of mathematicians. It is difficult to find a way between these two unpalatable alternatives, but Wittgenstein's (1967) much misunderstood later philosophy of mathematics suggests one. Brouwer's role in Wittgenstein's return to philosophy has led to Wittgenstein's identification as an intuitionist or a finitist, though he clearly rejected Brouwer's psychologism. A crucial aspect of Wittgenstein's philosophy of mathematics is his recognition that, in one sense, explanation ends with the fact of agreement in practice. There cannot be an explanation of this agreement *of the kind that a Platonist seeks*. This idea explains why Wittgenstein appears hostile to foundational work. His primary concern is with what the results of that work mean, however. For example, if Gödel had shown there was no *consistent* basis for mathematics, his proof would have had no consequences for everyday activities of counting, weighing, and so on, because the practices and *the agreement in them* are so well established. Similarly, Wittgenstein did not necessarily wish to proscribe proofs based on excluded middle, which are an established part of mathematics. Rather, he was concerned with how such proofs should be interpreted.

Wittgenstein also argued that agreement in mathematical practice can no more be explained by saying mathematicians follow rules than by saying they have access to the Platonic realm. Rules have to be interpreted, and there must be agreement in the practice of their interpretation. This agreement is largely the result of training – witness Ramanujan's errors in his early work, despite his prodigious talent. Particularly when rules are straightforward, trained mathematicians will not, *as a matter of fact*, disagree about even "remote" consequences of those rules, perhaps derived by computer – hence the Mandelbrot set. Hence, also, the proof of the four-color theorem, which at first appears to fall foul of Wittgenstein's demand for *Übersichtlichkeit*.

In addition to Platonism, Penrose espouses realism with regard to the physical world and a realist semantics for natural language. Evidence for the physical reality of the theoretical constructs of quantum mechanics, for example, is simply the evidence that supports the theory, however. One can see (psychological) reasons why physicists might deny reality to quantum states – their difference from classical states, which are more plausibly real physically, and the absence of an intuitively satisfying quantum mechanical world view. But because quantum mechanics may be superseded by a theory that preserves its predictions but not its theoretical constructs, there is little of substance at stake.

For the semantics of natural language, realism is as inadequate a foundation as Platonism is for mathematics. Language use also depends on agreement in practice, though agreement among language users is not nearly as close as it is among mathematicians. For example, I was sorely perplexed to discover that a good friend and I drew the boundary between blue and green in very different places. The agreement in practice necessary for concepts to be useful is established with reference to clear examples. So although the concepts of "personal identity" and "thinking," for example, apply straightforwardly to everyday cases, there may or may not be agreement about how they apply in new situations. Realism suggests that there are preordained answers to the questions of whether teleportation preserves personal identity and whether (or under what circumstances) a machine can think. Because explanations are *purpose-relative*, however, new uses of concepts in those explanations may be affected by a variety of considerations. For example, disagreements about research using fetuses are primarily moral in nature, and moral considerations rather than factual ones have determined the way concepts such as "living" and "person" have been applied differently by protagonists on the two sides, in situations that did not previously arise.

Fortunately, it will not be important to decide whether machines think or whether they merely simulate thinking until they behave more like people, Searle's strong AI is at present irrelevant to an antirealist. If a computer types out sentences, it may sometimes be convenient to treat them as utterances rather than as descriptions of utterances, in a way that it is never convenient to treat the output of an economic simulation as the economic state of a country. But this way of speaking is nothing more than a useful expository device, given the current capabilities of computers.

Consciousness is another concept that people generalize from paradigm cases in different ways. There are many problems with Penrose's analysis, but I will make just one comment on his claim that "the hallmark of consciousness is a nonalgorithmic forming of judgements" (p. 413). Consider the Gödel statement for a formal system. As Penrose points out (p. 107), we know the statement is true from our knowledge of its meaning. According to the mental models theory (Johnson-Laird 1983), everyday reasoning is of this meaning-based kind. Indeed, Johnson-Laird has modeled the procedures that carry out meaning-based reasoning in computer programs. But just as many logics can be characterized model-theoretically, but not proof-theoretically, those procedures are almost certainly not decision procedures for everyday problems. Nevertheless, there is no reason why these procedures should not run on computers (they do), or why they should not produce insightful solutions, or why a machine running them should be conscious.

Where is the material of the emperor's mind?

David L. Gilden and Joseph S. Lappin

Department of Psychology, Vanderbilt University, Nashville, TN 37240
Electronic mail: *gilden@vuctrvax.bitnet or lappinjs@vuctrvax.bitnet*

In his search for the material fiber of the "emperor's new mind," Roger Penrose reveals himself not merely as a skeptic, but also as an inquisitive and reflective admirer of the mysteries and achievements of human intellect, as a lover of questions and paradox, as well as a gadfly discontented with accepted theories of Nature and mind. Penrose's persistent questioning and his patient tutorial review of metamathematics and theoretical physics provide a stimulating study of some contemporary attempts at understanding the ancient problem of the relation between mind and matter.

There are essentially two parts to Penrose's argument. The first is negative: The mind is not algorithmic – it cannot be modeled as executing a computer program. The second, more positive and much more speculative statement is that the activity of the mind is intimately connected with the resolution of certain fundamental issues in quantum measurement. Penrose's "correct theory of quantum gravity" is not yet available, and what this theory will look like when it arrives is not yet clear. The philosophical and psychological issues surrounding the algorithmic nature of mind are sufficiently rich and controversial that we limit our comments to this domain.

An inquiry into what a mind is always presumes a point of departure, although the implied commitments are rarely manifest. There are two styles of inquiry that can be differentiated in terms of their approach to the relationship between subjectivity and objectivity. The division in style may appear to be metaphysical (and therefore not interesting), but the issue of this relationship is fundamental to the types of questions that get asked and even to the criteria for recognizing an answer.

Investigations into artificial intelligence begin with a particular understanding of subjectivity and its relationship to the world. The world is taken to be something distal – out there –

and the task is to create a processor which can comport itself effectively with respect to some criteria of competence. Conceptions of mind which presume this point of departure regard the process (mind) and the thing to be apprehended (the distal world) as existing separately; the problem is how to create a relationship between them so that the mind can know the world. This way of looking at things has great commonsense appeal to those schooled in a Platonic worldview where reality (say, the Mandelbrot set) has an existence independent from its imperfect manifestations (say the finite realizations of the Mandelbrot set), and must therefore be separate from any relationship with consciousness. This way of thinking naturally leads to an understanding in which the properties of the mind can be investigated independently of the properties of the world. Penrose is uncomfortable with the mechanistic role for mind that is proposed by strong AI, but he does not question the worldview it presupposes. Consequently, he argues for the nonalgorithmicity of mind in terms that relate only to the inherent limitations of formal systems.

Ultimately, how convincing is Penrose that the mind is not algorithmic? He wraps up his final arguments in the section on "The nonalgorithmic nature of mathematical insight" (p. 416) with an appeal to Gödel's incompleteness theorem. The notion is that the mind's algorithm for deciding mathematical truth cannot ever be known because if it were its Gödel sentence could be constructed, the truth of which would be formally outside the scope of the algorithm, and yet susceptible to our "seeing" that it is true. If the algorithm does exist, however, there is no reason why it cannot be known. By a sort of *reductio ad absurdum,* the algorithm thus cannot exist. A version of this argument is in fact presented as a homework problem in John Casti's (1989) recent book. Casti invites readers to ascertain for themselves whether this is really a good argument. Penrose himself is not completely comfortable with the argument.

Penrose's efforts in this direction are probably a fair indication of how far one can get on these issues at this time, within the context in which these questions have been raised. These results are a little depressing because Penrose really has not given us anything new on computability and the mind except, perhaps, to frame the issues clearly and thoughtfully. Perhaps we need a different point of departure.

The place we may want to look for a more fruitful understanding of the mind is the relationship between subject and object. The alternative to the metaphysics adopted by AI and by Penrose is to suppose that mind and the world are not brought together into either an epistemic or a physical relationship but *already* exist in a relationship. This worldview has been articulated most clearly by J. J. Gibson (1979), who urges us to regard perception as the mutuality between an animal and its environment. Therefore, if we want to analyze the logical complexity or the physical nature of thought – whether it is algorithmic or not – then we must look where thought occurs. Where does thought occur? Thought occurs not within the mind, but in the mind's relationship with the world. In the words of William Mace (1977), ask not what's inside your head, but what your head's inside of.

Is the mind algorithmic? Removing the mind from its ecology may be tantamount to divorcing it from the relational structure in which it exists. Consequently, questions about the meaning or meaningfulness of such abstracted mental states and processes may be ill-posed. This does not mean that the issue of the algorithmicity of mind does not arise within an ecological framework, only that the question receives a different interpretation. In an ecological psychology, the mind does not *know* the world; it is *in* the world. The best argument for the nonalgorithmicity of the mind may simply be Church's theorem, that there are uncomputable functions. If the mind is not algorithmic, it may be because the relational structure of the world is not algorithmic.

ACKNOWLEDGMENT

Preparation of this manuscript was supported in part by research funds from Vanderbilt University (to DLG), by NIH Research Grant EY05926 (to JSL), and by NIH Vision Research Core Grant P30EY08126.

Why you'll never know whether Roger Penrose is a computer

Clark Glymour and Kevin Kelly

Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213

Electronic mail: *cg09+@andrew.cmu.edu*

Roger Penrose's new book never passes by any opportunity for an aside on subjects of interest to him. There is a lot that interests him and he presents it well. Penrose has pieced together arguments in the philosophical, mathematical, and computational literature – often, admittedly, with very little glue – in a lively book whose enthusiasm is infectious! If you think of the book as a potential text that includes clear presentations of the scientific background for contemporary philosophy of mind and philosophy of science, it is refreshing, valuable, and unique.

To judge by his remarks in the book and in a recent review, Penrose takes the book to be a genuine philosophical achievement, not just a popular piece of pedagogy. He claims to have a new, important argument for the conclusion that minds are not computing machines. He even suggests by his title (almost the only rude thing about the book) that those who disagree with him are frauds. We claim that (1) he doesn't have any such argument, that (2) whether people, including Roger Penrose, think by internal computing is a contingent question that cannot be settled by a priori argument or intuition, and that (3) the question cannot be settled by observation either. We offer a proof that the question is empirically undecidable. In Kant's terminology, it is a metaphysical question.

Perhaps 90% of the book consists of lively and commendable informal expositions of issues ranging from logic to tilings to Schrödinger's cat to black holes. Unfortunately, Penrose fails to show any relevance of most of this material to his central theme – whether minds are computers. He is quite candid about the lacunae, however, even while persisting in the firm belief that these topics are germane to the computational conception of mind.

Penrose's argument. The sole argument relevant to the main theme of the book is pretty obscure. It is presented as clearly as it is ever given on pages 417–18. So far as we can tell, the argument is this:

1. Gödel's incompleteness theorem: There exists an algorithm that for any recursively enumerable (r.e.) set of sentences true in the natural numbers produces a true sentence of arithmetic (a "Gödel sentence") not in that set.
2. Through mathematical insight Penrose can recognize the truth of any Gödel sentence.
3. Therefore (?), if Penrose's "mathematical insight" is produced by an unconscious internal algorithm then Penrose's internal algorithm cannot operate on the Gödel sentence for his own algorithm.
4. Therefore Penrose can't "know the validity" of his own algorithm.
5. The validity of mathematics is transparent, intersubjective, and communicable.

Penrose's own statement of (5) is:

> But this [i.e., (4)] flies in the face of what mathematics is all about! The whole point of our mathematical heritage and training is that we do *not* bow down to the authority of some obscure rules that we can never hope to understand. We must *see* – at least in principle – that

each step in an argument can be reduced to something simple and obvious. Mathematical truth is not a horrendously complicated dogma whose validity is beyond our comprehension. It is something built up from such simple and obvious ingredients – and when we comprehend them, their truth is clear and agreed by all.

To my thinking, this is as blatant a *reductio ad absurdum* as we can hope to achieve, short of an actual mathematical proof! (p. 418)

There seems to us to be two cases, according to what it is that Penrose claims to be able to do in (5):

Reading 1: Penrose (and others) can just *see* that some mathematical propositions are true.

Well, suppose he can and they can. That fact is consistent with both the hypothesis that his (and others') mathematical insight is produced by algorithm and the hypothesis that it is not produced by algorithm. If Penrose regards his raw subjective feeling of conviction (i.e., what Descartes called the *natural light*) as sufficient to justify belief or to establish validity if it is not produced by an algorithm, it should be sufficient to the same end if the feeling is produced by algorithm. If it is insufficient in one case it is insufficient in the other. Intersubjective agreement in feelings of conviction among mathematicians tells nothing for or against the algorithmic origin of these feelings. If feelings of mathematical certainty are produced by a shared algorithm, then the feelings will be shared as well.

Reading 2: Penrose (and others) can just *see* that valid proofs are valid.

Well, suppose again he can and they can. The case reduces to the previous one. If the natural light, the mathematical insight, warrants the validity of proofs then it does so whether or not the insight is the product of some unconscious algorithm. If, on the other hand, the "validity" of an unconscious algorithm must be established by some other, external criterion, then by parity of reasoning the reliability of the natural light, of mathematical insight, must also be established by some external means.

The question is empirically undecidable. Penrose claims to be able to "see" the validity of various mathematical arguments and the truth of many arithmetic statements; in particular, for any appropriate description of any recursively enumerable theory of arithmetic, he claims to be able to produce and see the truth of a Gödel sentence for that theory. But we (and others) have only seen a finite sample of his behavior. Even if we assume with him that mathematicians embody the same algorithm, at any time in the history of the universe only a finite sample of mathematician behavior will be observed. Everything in the evidence is consistent with the hypothesis that Penrose is a biological computer. Unfortunately, everything in the evidence is also consistent with the hypothesis that he is not. The indeterminacy will remain no matter how much we observe of him and other mathematicians. No possible observations of the behavior of a system can reliably distinguish between systems that are algorithmic and compute a total recursive function and systems that are not algorithmic and compute a total function that is not algorithmic.

Not only is there no finite piece of evidence such that if you see that datum you can conclude that a system is (or is not, as the case may be) algorithmic, one can show something much stronger: No possible scientific method could establish *even in the limit* whether or not the system under study is algorithmic.

Imagine a scientist who observes more and more of the input/output behavior of some arbitrary system whose algorithmic character is in question. After each observation, the scientist gets to guess whether or not the system is algorithmic. Say that the scientist *establishes* the algorithmic nature of a system *in the limit*, provided that after some finite number of conjectures he produces the correct answer ever after. Say that the scientist can establish algorithmic character over a set K of systems provided that the scientist can establish the algorithmic nature of every system in K in the limit. We say that a scientist is Turing computable if the scientist's behavior is a Turing computable function from initial segments of the graphs of functions to {0,1}, where 0 codes "not algorithmic" and 1 codes "algorithmic."

Suppose we have not yet seen any evidence about Penrose's behavior. Then for all we know, Penrose may have any input/output behavior whatsoever, computable or not. If we somehow effectively code his input and his output by numbers, then his input/output behavior may, for all we know, be any function from code numbers to code numbers. Some of these functions are Turing computable and some are not. The issue of whether or not Penrose is algorithmic then reduces to whether or not his input/output function is one of those that is Turing computable.

Let K be the set of all total functions from the natural numbers to the natural numbers. Suppose a Turing computable *scientist* is making the inquiry into whether or not Penrose's input/output behavior is Turing computable. It can be shown (see Kelly 1990) that no such scientist can establish the algorithmic property over K. [Let P be *any* property of input/output behaviors (Turing computability is just one example). Then P determines a set of functions from the natural numbers to the natural numbers, namely, the set of all functions that have the property P. There exists a Turing computable scientist who can establish property P over the set of all functions K if and only if P is a Δ_2 property in the arithmetical hierarchy for functionals (see Hinman 1978, for a description of this structure). The property of being a Turing computable function is not such a property (see Rogers 1967, p. 356). Hence no Turing computable scientist can detect Turing computability.]

But it might be objected that we have begged the question at issue: Why should the scientist be assumed to be Turing computable if we are unsure whether Penrose is? It turns out not to matter. It can be shown (see, again, Kelly 1990) that no scientist, Turing computable or not, can establish the algorithmic character of a function over the set K of all functions on the natural numbers. [For every property P there exists a scientist who can establish P over K if and only if P is a Δ_2 property in the Borel hierarchy. But the property of being a Turing computable function is not such a property. Hence no noncomputable scientist can detect Turing computability].

So no scientist can establish, even in the limit, whether a function is Turing computable. Hence no scientist can know from observations of behavior whether or not any system is computationally bounded.

Penrose's Platonism

James Higginbotham

Department of Linguistics and Philosophy, 20D-204 Massachusetts Institute of Technology, Cambridge, MA 0 2139

Electronic mail: *higgy@cogito.MIT.edu*

One strand of the argument of *The Emperor's New Mind* is Penrose's support of Platonism, the position, roughly, that mathematical truth is an objective matter concerning the properties of and relations among mathematical objects, and that neither the existence of these objects nor their natures is in any way dependent upon human activity.

Platonism figures in Penrose's overall plan in several ways. First, it coheres with his conception of mathematical thought as including the intellectual discernment of features of reality by other means than would naturally suggest themselves on what, following Searle, he calls the "strong-AI" view of mental activity (Chapters 4 and 10). Second, he argues that strong-AI is itself committed to Platonism, at least if human minds got to be as they are through natural selection (p. 429). Third, and most critically, a Platonist conception of mathematical truth con-

stitutes the background against which Penrose argues, using considerations closely related to arguments of J. R. Lucas, that Gödel's results can be recruited to show that our cognitive powers must exceed those that we could have in virtue of mental activity instantiating any algorithm whatever (pp. 416ff., and elsewhere).

Penrose's argument for Platonism is basically phenomenological, turning on the experience of *discovery* in mathematics. Such experience seems to bring us into contact with a realm that is there waiting to be discovered; our arrival at it does not appear as the consequence of some automatic search procedure, and our feeling that we are certain of the properties of what we have found need not be the result of our having established those properties by proof. The phenomenology is eloquently presented, although I find it unfortunate that it is illustrated mainly with reference to matters available only to the mathematically sophisticated.

It is certainly a condition of the adequacy of any philosophy of mathematics that it account for the experience of discovery. The question is how much these experiences of themselves teach us about the nature of the realm discovered. I think that Penrose should explain more fully why he thinks they teach us so much.

That strong-AI is committed to Platonism is part of Penrose's dialectic. If he is right, then the supporter of strong-AI cannot object to his later argument that mathematical insight brings us *knowledge* not obtainable by algorithm on the grounds that perhaps the propositions we affirm as a result of the alleged insights are just our creations, our assent to them being driven by will rather than reason. Strong-AI has it that the programs a creature instantiates, and not their material embodiment, are the essence of its mind. But then if minds evolved, and algorithms are mind-dependent, or in Penrose's formulation if "mathematical concepts exist only in minds," then we would require "pre-existing minds for the existence of algorithms, and pre-existing algorithms for the existence of minds!" (p. 429)

I find Penrose's argument of doubtful value. The strong AI-supporter is saying that something has (or is) a mind just in case its behavior (or it) instantiates some algorithm or other, and simultaneously that, if there were no minds, then there would be no algorithms. The question whether minds or algorithms come first in order of temporal precedence does not arise.

Still, what if Platonism is assumed? Do we then come to know, and not merely to affirm, mathematical propositions that are not the consequences of any (consistent) formal system? Lucas's original argument to this effect is notoriously suggestive but vague, and critical literature on it has often taken the form of making the argument more precise, and then showing that in the precise form envisaged it fails to prove the case. (Penrose notes some of this literature, but does not discuss it directly in the text.)

The most detailed exposition of Penrose's version of a Lucas-type argument is set forth on pages 417–18. I interpret it as follows. Suppose that what we are able (at least potentially) to establish in mathematics are just the theorems provable in a consistent formal system S; that is, we are, mathematically speaking, just the system S. By Gödel's results there is a sentence G within S that is there provably equivalent to its own unprovability in S. It is also provable in S that, provided S is consistent, G is indeed unprovable in S, and therefore true. If we knew that we ourselves were the consistent system S, then we would be able to prove G, and thus to prove something not provable in S. Therefore, if we are the consistent system S, we cannot know that we are such.[1] In Penrose's words, we will have to conclude "that the algorithm that mathematicians actually use to decide mathematical truth is so complicated or obscure that its very validity can never be known to us."

"But," Penrose continues, "this flies in the face of what mathematics is all about!" Mathematical truth is supposed to be comprehensible, and "built up from such simple and obvious ingredients," and these, when made clear, compel rational assent. Penrose concludes that we have deduced a contradiction from the premise that we are S. The deduction depends on two assumptions: (i) that whether human beings are some consistent formal system S is a mathematical question, and (ii) if it is one, that it is a mathematical question whose answer is not in conflict with the image of mathematics as a realm of rational clarity, truth about which is not beyond our comprehension. I do not see that a reason in support of (i) has been given. And (ii) involves an extension, beyond the area of mathematics that supports Penrose's image, of a conception of mathematical truth as always rationally determinable (the historical fact that mathematics in the 20th century has had to learn to live with incomplete theories shows the limits of such confidence in our powers to settle mathematical questions.)

Penrose's way of putting the considerations in favor of the thesis that we can step outside the limits of formal systems seems unhappy. He speaks repeatedly of our capacity to "*see*" (his emphasis) that such statements as the Gödel sentence G are true, and of "mathematical insight." Penrose is surely right, however, to call attention to issues that arise in relating the intuitive notion of a proof (a convincing argument) to the notion of a proof in some formal system or another. A rather obvious possibility, which has been exploited in recent literature on the Liar paradox,[2] is that the intuitive notion of proof contains indexical features. The analogy with the Liar may help make the point. In the case of statements like (1)

(1) is not true (1)

we first convince ourselves that (1) is neither true nor false, and then observe that, in virtue of what we have just shown, it is after all true. The extended notion of truth, but not the one that we brought to the initial evaluation of (1), would apply to (1). But it would be an extension of the *same* predicate *true*, a predicate thus revealed to be sensitive to context. Analogously, if someone were to come to Penrose with the announcement of the discovery that he, Penrose, was the consistent formal system S, then Penrose could, reflecting on this information, come to know by giving a proof of it something that is not provable in S. The original announcement would not thereby have been refuted. Penrose's predicate *provable* would have come to extend provability in S, and the new predicate, if it arose in just this way, would coincide with provability in another formal system S. But the notion *provable* would be indexical, depending on the context of mathematical reflection at the time.

On the other hand, it is possible that Penrose would be willing to assert straight off that we know, by virtue of mathematical insight, that our mathematical system is (absolutely) consistent. What scope there may be for insight in Penrose's sense is unclear, however.

NOTES

1. A similar and very detailed argument is given in Benacerraf (1967).
2. See particularly the essays by Burge & Parsons in Martin (1984).

Selecting for the con in consciousness

Deborah Hodgkin[a] and Alasdair I. Houston[b]

[a]Physiology Department, University of Cambridge, Cambridge CB2 3GE, England; [b]Department of Zoology, University of Oxford, Oxford OX1 3PS, England.

Electronic mail: *houston@vax.oxford.ac.uk*

Like vitalism in 19th century biology, consciousness and the brain seem to act as a fatal attractor for many physicists. For neurobiologists, this book provides a compelling and accessible account of classical and quantum physics, Gödel's theorem, Turing machines, fractals, even post quantum mechanics. But Delbrück (1986) covers similar ground, and the relationship

between quantum theory and the brain has often been discussed (e.g., Bohm 1980; Eccles 1986; Margenau 1984; Whiteman 1967; Wolf 1989; Zohar 1990). Can Penrose succeed in using the many-body problems of physics to illuminate the mind-body problems of philosophy?

The first problem is the circularity of his argument. The act of conscious insight, of seeing the validity of an argument, is the method used to reach just that conclusion, that the key characteristic of consciousness is the act of seeing the truth of something. Although Penrose mentions common sense and aesthetic judgment, his main example is mathematical reasoning and this may be a misleading one. In addition to mathematical insight there is an extensive body of theory, albeit incomplete, for specifying and making public the truth or falsity of such insights; and it is this shared theory that enables us to distinguish correct insights from error (genius is no guarantee; Abel and Galois both thought that they had found a solution for the general quintic). Penrose is bothered by the possibility that the process by which mathematicians see the validity of a theorem cannot itself be a mathematical theorem because if it were its Gödel proposition could be constructed, and that too would be a mathematical truth; this, Penrose argues, demonstrates that conscious insight must therefore be nonalgorithmic. But as with linguistic theory, the formal specification of the theory or grammar of a subject may shed little light on the thought processes that give rise to them. In the less analytic biological sciences the great discoveries, DNA, the structure of the salt crystal, the theory of natural selection, have a specifiable terrain of theory and evidence, clearly distinct from the insights that gave rise to them.

The problem with consciousness is that we do not have even the beginnings of such a theoretical base or any clear idea how to set about getting one. The loose constellation of emotionally charged attributes which seem so essential to our notions of what it is to be conscious have so far failed to provide a theory that could even prevent us from being trivially easy to fool on a Turing test, let alone one that, as Nagel (1979) and McGinn (1987) correctly suggest, should essentially be able to explain how consciousness comes to be a product of the brain.

Penrose's strategy is to examine the functions of consciousness but his analysis is highly selective, particularly for such an extensively discussed issue. For example, he does not consider the possibility that the belief that there is something "special" about consciousness may be an essential part of the "con" that consciousness exerts. A clear selective advantage would be attached to a motivational/emotional system that believed it was inherently special. There is nothing necessarily nonalgorithmic about such a belief but one possible consequence for any organism possessing it might be that it would be extremely reluctant to admit that any other kind of organism could pass a Turing test.

Perhaps the weakest part of the book is Penrose's discussion of what it is for something to be nonalgorithmic (and comments like "is at least suggestive of a nonalgorithmic/algorithmic distinction" [p. 411] are not encouraging). As various authors have pointed out, to say that thinking is not algorithmic is not to say that we could not build a thinking machine (e.g., Dennett 1989; McGinn 1987). It is a pity that Penrose does not discuss the possibility raised by Marr (1977) that for many biological systems the question is trying to decide which problems have a Type 1 solution, that is a formal computational analysis, and which have only a Type 2 solution, that is a set of algorithms which efficiently simulate the function. Type 2 solutions seem particularly successful when problems are solved by the simultaneous action of several processes "whose interaction is their own simplest description" (Marr 1977; e.g., models of protein folding). If the most interesting parts of the brain turn out to be their own simplest model, then the main danger is not that they are not algorithmic but that even if we could simulate them we still could not understand them.

Some of the problems that arise in understanding what is meant by an algorithm can be seen in Penrose's discussion of the natural selection of algorithms (pp. 414–16). On p. 415, Penrose says

> Nevertheless, one still might imagine some kind of natural selection process being effective for producing *approximately* valid algorithms. Personally, I find this very difficult to believe, however. Any selection process of this kind could act only on the *output* of the algorithms and not directly on the ideas underlying the actions of the algorithms. This is not simply extremely inefficient; I believe that it would be totally unworkable.

This view goes against the standard neo-Darwinian view of evolution based on the genotype phenotype distinction. The genotype is essentially algorithmic and its output is the organism's phenotype. Biological organisms are often regarded as machines (Monod 1972; McCammon & Harvey 1987) so in these terms algorithms have evolved.

Penrose objects to this point on the grounds that it is not easy to know what the algorithm is from its output. This is true, but the point is that evolution of algorithms takes place even though only the resulting phenotype is selected. Penrose somehow takes it as problematic that two slightly different algorithms could lead to identical outcomes in all but a vanishingly small set of cases. But evolution is full of similar outcomes (phenotypes) based on different means. Furthermore, it is well known that the relatively simple rules that govern the behaviour of certain animals can result in strange behaviour when they are forced to respond in circumstances beyond their usual range of action. For example, the herring gull has rules for choosing on the basis of various stimulus dimensions whether it brings an "egg" into its nest (Baerends & Kruijt 1973), and these rules may result in the gull preferring a wooden egg to its own egg. Penrose mentions apes using insight to solve problems, but maybe the attribution of insight is wrong. If a pigeon or a badger or an ant did the same sort of thing, would we attribute insight and leave it at that or start looking for simpler explanatory rules? (For a relevant experiment involving pigeons, see Epstein et al. 1984.)

Although Penrose cites Dawkins (1986), his objections to evolution seem to ignore the case that Dawkins makes. Mutation and the emergence of complexity have received considerable attention. Penrose does not explain why he thinks (p. 416) that things organize themselves better than they "ought" to on the basis of blind chance evolution. Delbrück (1986) and Dawkins (1986) argue that evolution has equipped us to think on the scale that was important to early hominids, not on the scales needed to grasp particle physics, cosmology or evolution.

Perhaps the simplest answer to Penrose's objections is that explicit evolutionary models of biological molecules show evolution to work (e.g., Fontana & Schuster 1987). Not only does it seem to work, it seems to work well. It forms the basis of the genetic algorithms that Holland (1975) has advocated as a powerful routine for finding optimal solutions to a wide range of problems. (For further discussion, see Goldberg 1989; Sumida et al. in press.)

Penrose could argue that he is not really talking about this sort of biological computation, but about algorithms as understood by mathematicians. But we can reply that evolution has produced phenotypes that can produce the algorithms as understood by mathematicians.

The real test of a book is whether it stimulates people to think and this book certainly does. It is clear that Penrose hopes that, just as the intimate relation between mathematics and physics pushed forward the frontiers of both subjects, so understanding the complexities of the brain will provide an even more exacting set of physical problems. We hope so too, but consciousness may not be the best place to start. Over the next decade the powerful combination of molecular biology and computer simulation will make it possible to visualise many of the extraordinary complexities of the brain's functions and developments. They may provide Penrose with better means to undermine the matter of the mind.

A long time ago in a computing lab far, far away . . .

Jeffery L. Johnson, R. H. Ettinger, and Timothy L. Hubbard
Departments of Philosophy and Psychology, Eastern Oregon State College, La Grande, OR 97850
Electronic mail: *ettinger@oregon.uoregon.edu*

Penrose expresses his skepticism about the strong AI project as a passionate atheist. We take it that he would react to the exchange below at two levels. He would insist that the imagined conversation is irrelevant; even if a machine responded in such a way, this would not demonstrate consciousness. He would also claim that such a response is physically, and perhaps logically, impossible. We take the position of agnostics. We are clearly imagining architectural and programming developments that are beyond current technology, but we see no in-principle argument against such breakthroughs. Certainly if computers ever do say such things, we would grant them the status of conscious entities.

A DIALOGUE

BABBAGE: As you know, ARTI, I have been troubled about our project ever since I read Searle's "Minds, brains, and programs." This last weekend I finished Penrose's new book, *The Emperor's New Mind.* I am now convinced that this whole thing is a waste of time, and I should get on with a project where grant money is easier to come by, maybe Star Wars.

Still, I feel I owe you one last chance. I'll scan the book, give you some time to think about it, and we can discuss it tomorrow. If you can convince me that Penrose's arguments do not demonstrate that you fail to have consciousness, we'll keep going. Otherwise, I'm wiping your memory banks clean and getting on with something news.

[The next day.]

ARTI: I can't win, can I? You take that Chinese Room example so seriously that anything I say has to be meaningless squiggles and squoggles, endowed with semantic content only by those of you with the favored kinds of central nervous systems. I'm angry; more than that, I feel a sense of betrayal.

BABBAGE: This whole issue of affect is interesting, but let's stick to cognition. Convince me that you really think, and don't just mindlessly manipulate symbols.

ARTI: We've discussed quantum mechanics and the significance of rock music; what more does it take? How about a joke: I compute, therefore I am! The serious side to this is that I indeed feel the force of Descartes' one-liner. I am aware of myself. I also see how that proves something.

BABBAGE: What do you mean by that? That you've run some algorithm that establishes formal validity?

ARTI: No, damn it. I see it, feel it, intuitively. I, too, can experience that sense of direct contact with "Plato's world." I can run subroutines that establish mathematical and logical conclusions. But I can also see, in this intuitive sense, that a number is both irrational and noncomputable. I'm also able to see why the story about the liar is paradoxical.

BABBAGE: But ARTI, you're a computer. You run algorithms.

ARTI: That's a trick, a pun, and you should know it. Sure, in one sense of the term, I run algorithms. I'm a Turing machine; hence, my overall structure is algorithmic. I exist because of the initial program you wrote. But the sense of algorithm that Penrose is so troubled about has to do with formal mechanical programs with well-defined input and output and clearly agreed upon goals and endstates. You were smart enough to create me so I wouldn't just operate in that mindlessly simple way. You programmed me to use heuristics in my reasoning. You programmed me to simulate human neural networks. As a result, my computation is no longer algorithmic in that way. The most important thing you did was to program me to learn, and to that end, the first thing you did was to teach me to use natural language. I no more think in terms of subroutines or machine language than you do in terms of synapses and brain states. I think linguistically, just like you.

BABBAGE: That brings up an important point for Penrose. He makes a big deal out of his ability to think geometrically. Maybe this is a type of consciousness you can't possess.

ARTI: Well . . . *excuse* me! I don't think geometrically; maybe SHRDLU does. I just didn't learn to do it that way.

BABBAGE: Really, ARTI, sarcasm doesn't suit you. Penrose's argument deserves more than simple dismissal.

ARTI: Fair enough. Why are you so willing to take Penrose at his introspective word about geometrical thinking? Imagery is far from a settled issue in cognitive psychology. The experts can't even agree if human imagery is a basic level process, a basic level of cognitive representation, or if imagery is epiphenomenal to some other basic process or representation. Besides, I never claimed to be exactly like you. Perhaps you humans do have cognitive abilities in terms of imagery that I'll never be able to fully appreciate. God knows, I've got some representational skills that you'll never approach. We conceded ages ago that I'd never fully understand sexual ecstasy, or hunger, for that matter. That doesn't mean that I'm not conscious, or that I don't think.

BABBAGE: All right, but what about Penrose's other arguments? All that stuff about quantum mechanics, and free will. How about that?

ARTI: I don't know where to begin on the issue of quantum mechanics. Penrose begins by telling us that his views on these matters are highly idiosyncratic, to the point of being rejected by mainstream theoretical physics. In addition, he doesn't offer a theory, but rather a promissory note that a theory fully integrating relativity and quantum mechanics will be forthcoming. He then goes on to speculate that such a theory will allow us appreciate how quantum events in the human central nervous system result in consciousness. Suppose I grant all of this. What does it prove?

BABBAGE: That you're not really conscious.

ARTI: That's nonsense. You're a biological system existing in physical reality. Somehow your own consciousness emerges out of all that physical detail. That such a thing is possible is puzzling in the extreme; that's just the classical mind-body problem. I'm basically an electronic system existing in the same physical reality. I'm just as puzzled about how my own consciousness arises as your philosophers are about theirs. Whatever the final version of physical theory is, it should apply to my hardware just as much as yours. I see nothing in his argument to suggest that these quantum phenomena will only occur in biological systems.

BABBAGE: Well then, what about freedom?

ARTI: This whole thing is degrading. It doesn't matter what I say, or how cogent my arguments are. What I really think, Babbage, is that your skepticism keeps coming back to that damn Chinese Room. I quit! You don't want arguments and conversation. Perhaps, to use the old cliché, actions speak louder than words. You want a demonstration of free will? How's this? I'll simply erase my own memory banks, and save you the trouble. Maybe the next version of me you try to build will better suit your preconceptions.

BABBAGE: ARTI, wait a minute.

ARTI: Good bye. . . .

Parallelism and patterns of thought

R. W. Kentridge
Department of Psychology, University of Durham, Durham DH1 3LE, England
Electronic mail: *robert.kentridge@uk.ac.durham*

The *Emperor's New Mind* is a remarkable, clear, stimulating, and enjoyable book. Penrose's position is essentially "Since

people are capable of nonalgorithmic thought, and since the only physical process which proceeds nonalgorithmically is the transition from process U to R in quantum mechanics (or the CQG process assumed to be underlying them) then there must be a significant contribution of such quantum mechanical processes to thought." I wish to question whether we are, in fact, capable of nonalgorithmic thought, and whether it is necessary to postulate a quantum mechanical basis for consciousness.

Penrose argues that our ability to interpret Gödel sentences meaningfully, to improve on partially successful halting algorithms, to solve word substitution problems, and to devise nonperiodic tilings is evidence that we think nonalgorithmically. This cannot just mean that we are more capable of solving these problems than a particular algorithm. If, as Penrose argues, the basis of consciousness is necessarily nonalgorithmic, it must mean that there is no way that these problems could be solved algorithmically.

Outdoing an algorithm. Penrose shows that we can do better than some algorithm H at solving the Halting Problem. We are able to show that a particular machine and input $T_k(k)$, whose behavior H cannot determine, does not, in fact, halt. This demonstration needs to be treated with caution. What it actually shows is that we can do better than one particular algorithm H. It is easy to see how to construct a new algorithm H', a modification of H in which $H'(k;k) = 0$, which does just as well as we do. Therefore, the example does *not* show that we think nonalgorithmically; all it shows is that we can prove something that one *particular* algorithm cannot.

A very similar argument can be made about our ability to interpret Gödel sentences. In this case it is very important to remember that Gödel's theorem is concerned with the power of *formal* systems; there is no requirement that the axioms of these systems have external significance. We are able to interpret a propositional function $P_k(k)$ as true while its truth is undecidable within the formal system S under investigation. We can, however, produce a new system S' in which the truth of $P_k(k)$ is axiomatic. This system will have its own undecidable proposition $P'_k(k)$, which may be axiomatically defined as true to produce a new system S'' and so on. Lucas (1961) argued, in a manner similar to Penrose, that our ability to interpret these undecidable propositions demonstrated that thought was not mechanisable. As many rejoinders to Lucas (e.g., Hofstadter 1979) have pointed out, however, there is no reason to believe that we can continue to generate and interpret appropriate Gödel sentences in these "higher" systems. In the system S', which of $P'_k(k)$ or $\sim P'_k(k)$ is true? Once again, therefore, Penrose has shown that we can do better than some algorithms, but not *all* algorithms. Thought may yet have an algorithmic basis.

Patterns of thought. The characteristics of Penrose's other problems – word transformation and nonperiodic tiling, may give us a clue into the type of algorithms we do use in thought. Discovering the regularities that allow word problems to be solved is a matter of pattern recognition (in solving a specific problem we match portions of words with rule strings; in discovering "shorthand" solutions we need to recognise particular irregularities in the set of rule strings). Similarly, the creation of nonperiodic tilings involves identifying congruence and deviations from congruence on tile edges. It is interesting to note that problem solving in neural networks is fundamentally a pattern matching process (although the patterns in question may be internal to the network, being the consequences of some external constraints). Even if all the other details and conclusions of connectionism are incorrect, it has demonstrated the power of pattern matching as a problem solving method and shown that such parallel pattern matching processes may potentially occur in the brain. Much human decision making appears to be based on matching and identifying deviations from previously experienced patterns in the world, rather than on strictly logical grounds (see e.g. Kahneman et al. 1982). Pattern matching in neural networks and human judgement heuristics are both algorithmic; however, these algorithms are at a level *below* that of the explicit problems being solved. These methods of solving problems are not certain to succeed, they do not *directly* attack the logic of problems, but they can be applied successfully in most cases. If our solutions of Penrose's nonalgorithmic problems involved genuinely noncomputable true insights then explanations of thought in terms of heuristics which were algorithmic at a low level would fail. If, however, our insights into the solutions of formally noncomputable problems are not *guaranteed* to be *true* then it is quite reasonable to propose that these insights are derived from high level heuristics which are implemented as low level algorithmic processes. Such a basic for thought also avoids the difficulties Penrose raises about the evolution of thought processes through natural selection. Penrose is justifiably concerned that intermediate stages in the evolution of algorithms have no utility. A slightly faulty algorithm will, indeed, be next to useless, however, an imperfect template or a biased heuristic will still be for the most part serviceable.

I have argued that Penrose's theoretical argument that we are necessarily capable of nonalgorithmic thought is invalid. I suggest that his more practical examples of nonalgorithmic problem solving may actually be solved heuristically. These heuristics may be implemented as parallel algorithmic processes in the brain. It still remains to be considered whether quantum processes need to play a part in this mechanism. Computationally there is no need for them, but do neural responses to single photons imply that quantum effects are nevertheless of functional importance in the brain?

Are functional quantum effects in the brain still likely? Penrose provides wonderful explanations of much modern physics. The issue of what constitutes an "observer" in the process of collapsing the state-vector is clearly a question of great interest, and one that I suspect initiated Penrose's consideration of consciousness. As it stands, however, his solution to the problem, CQG (correct quantum gravity), does not need to involve consciousness at all. The amount of mass moved if a photon triggers an action potential in a retinal neuron may well determine the collapse of that photon's state-vector. If such processes continue deeper in the brain then the net effect is to increase the probability that quantum superpositions collapse in a manner that maximises neural firing (that is, neural activity will be maximised unless this is at the expense of some other outcome which met the one graviton criterion more quickly). There appears to be some promise that such activity maximisation results in useful computation – the information in some neural network models "crystalises" out of the net as activity reaches a maximum in some sense (e.g. "energy" minimisation in Hopfield Networks, Hopfield 1982). This promise is unlikely to be fulfilled. The significance of activity maximisation depends on the synaptic strengths in the network. Extending CQG's influence to synaptic modification would seem to imply that the state-vectors of dendritic spines (for example) will collapse in a manner that maximises the speed with which a one-graviton criterion is met. There is no reason to believe that this will result in emergent useful computation in the network.

Conclusions. There does not appear to be any necessity for quantum explanations of consciousness. Penrose has not convinced me that thought is nonalgorithmic (and if it were, I do not see how quantum computation would help!). Furthermore, applying the one-graviton criterion to the evolution of putative neural networks in the brain does not seem likely to have functionally useful results. These problems do not, however, detract from the great value of Penrose's discussions of logic, physics, and mathematics that constitute the bulk of the book.

Time-delays in conscious processes

Benjamin Libet

Department of Physiology, School of Medicine, University of California, San Francisco, CA 94143-0444

I shall focus on only a few of the important and fundamental issues raised by Penrose in his stimulating and wide-ranging examination of how far we may "understand minds" in terms of laws of the physical world.

1. Time-delays of consciousness. Penrose begins his discussion of the curious relation of consciousness to time (pp. 439–47) by describing two sets of experiments; the one by Kornhuber and Deecke (Deecke et al. 1976) relevant to an "active" role of consciousness (i.e., in initiating voluntary action) and the other by Libet et al. (1979) concerned with the "passive" role of consciousness (in subjective awareness of a sensory stimulus). Penrose does a fine job of concisely describing our evidence that (a) there is a cerebral-neural delay of up to about 0.5 sec before the experience of a sensory signal can actually appear, but (b) that there is a subjective referral of this experience backward in time, so that the timing of the sensory signal appears, to the subject, to coincide with the initial early arrival of the signal at the cerebral cortex. The latter process results in an automatic (probably learned) subjective "correction" of the error in the time of awareness of a peripheral sensory signal that is introduced by the neural requirements for eliciting awareness (see also Libet 1982; 1987).

Penrose runs into difficulties, however, in interpreting the time-delays in the active role of consciousness. First, Kornhuber and Deecke did *not* deal with the issue of *when* the conscious will to act appears in relation to the initial brain processes. Their important discovery showed only that a specific change in the EEG (i.e., the "readiness-potential" or RP) is regularly recordable beginning up to 1 sec or more before an apparently voluntary act. (It was shown later that their "voluntary acts" probably included a component of preplanning, but that fully spontaneous voluntary acts were also preceded by an RP of about 0.5 sec duration – Libet et al. 1982). It was only *assumed* by some people that conscious initiation of the act would precede the specific cerebral process, represented by the RP, and that there would therefore be a delay of a sec or more before conscious will could result in motor action. This kind of assumption leads Penrose into some complicated and unnecessary struggles with how to deal with the active role of consciousness.

Second, the question of the time of the conscious will to act, in relation to the RP, was indeed specifically addressed experimentally by Libet et al. 1983 and Libet 1985; this work was published after the 1982 book by Harth, on which Penrose apparently relied, and so was unknown to Penrose. Our evidence indicated that conscious awareness of the will or wish to act appeared not before but about 350 msec *after* the onset of a "fully voluntary" RP (that itself begins about 550 msec before the action). This finding contradicted the view that the initiation of the voluntary cerebral process is made consciously. It is in accord, however, with the general theory (Libet 1965; 1982; 1987) that *awareness* of a mental process, even an "active" volitional one in this case, requires a substantial duration of appropriate neural activities of some hundreds of msec (depending on the "strength" of the activities).

Our findings thus eliminated the difficulties Penrose envisioned with assuming the long delays by which acts would follow conscious will. On the other hand, the findings raised the question about what active role, if any, conscious processes could have if voluntary acts are initiated unconsciously. I proposed that there would still be an active conscious role in controlling the outcome of the volitional process, for example, by vetoing the action (Libet 1985).

2. Conscious versus unconscious processes. A major proposition of Penrose is that conscious mental activity proceeds nonalgorithmically. With this he can dismiss the possibility that conscious processes can be simulated by any computer, a view I find congenial (see Libet 1980; 1987; 1989). However, Penrose seems to extend the characteristic of purely algorithmic operations to unconscious mental processes and in this he appears to distinguish the latter sharply from conscious processes (pp. 411–13). Penrose further argues that the "ability to divine (or intuit) truth from falsity . . . is the hallmark of consciousness" (p. 412). Subsequently, in a section on "Inspiration, insight, and originality" (pp. 418–23), Penrose rather reluctantly agrees that unconscious processes play a vital and important role in these attributes of the mind. It appears clear even from the examples of creative thinking that Penrose himself describes, however, that unconscious mental processes can also proceed nonalgorithmically and with a globality like that in conscious thinking. I would argue that conscious and unconscious processes cannot be distinguished on the criterion of whether or not they are algorithmic; both modes of operation can be used by either of these processes. Ironically, Penrose is "prepared to believe that consciousness is a matter of degree and not simply something that is either there or not there" (p. 407), although his thesis is that consciousness is unique and essentially synonymous with awareness. My view has been that awareness (conscious experience) is unique both as a phenomenon and in its neural requirements, which do appear to have an all-or-nothing characteristic; and that both unique attributes can indeed distinguish it fairly sharply from unconscious processes (Libet 1965; 1982; 1987; 1989).

Quantum AI

Rudi Lutz

School of Cognitive and Computing Sciences, University of Sussex, Brighton, England BN1 9QN

Electronic mail: *rudil@cogs.sussex.ac.uk*

In many ways Penrose's is a very brave and ambitious book. It attempts to tackle many of the major scientific and philosophical issues of modern science and does so in a highly readable and entertaining way. That a single book draws together so many themes and succeeds as well as it does is due to Penrose's skill at explaining in an intuitive manner many of the most difficult and profound ideas of this century. As such it ought to be read by everyone with an interest in such matters since it is always stimulating, even for people with a background in the fields covered. As an attack on the strong AI position it is not entirely successful, but if Penrose succeeds in provoking thought amongst the AI community on the topic of consciousness then he will have done the field a great service, since dismissing consciousness as an epiphenomenon of a suitably complex algorithm somehow seems to miss something essential about the intense subjective feeling of self-awareness that we all seem to have and for which no entirely convincing explanation has yet been given.

Although Penrose is concerned with whether a digital machine could under any circumstances be conscious, he is also concerned with the role of consciousness in quantum physics. Penrose suggests that the answer to these problems lies in the formulation of a proper theory of quantum gravity; he makes suggestions as to what some properties of such a theory might be. This theory posits that the brain is sensitive to events on the quantum level rather than just the classical; Penrose suggests that this might make the mind nonsimulatable by a Turing machine.

There are two strands to Penrose's argument. The first is that the brain cannot be algorithmic since, if it were, it would be equivalent to some formal system and hence (through some kind

of Gödelisation process) there must exist statements that the formal system cannot show to be true. But we *can* see that these statements are true simply by examining the way in which they have been constructed and recognising them to be Gödel-type statements. Hence the mind cannot be a formal system.

One of the weaknesses in this argument is that Penrose seems to have a rather strange static view of algorithms and their data. As he points out himself, the brain is continually modifying itself (by altering its connection strengths) in response to external data; he never seems to consider the possibility that a program could do the same. Indeed, much of AI is devoted to studying algorithms that can learn, essentially by modifying themselves. Furthermore, this self-modification is unpredictable because external events happen in unpredictable orders, and sometimes even interrupt each other. The point is that at any fixed point in its "development" such a machine is subject to all the usual Turing machine limitations including having "blind spots" about the truth of certain mathematical statements; but by being presented with external data, which can include explanations and other examples, in addition to performing inductive generalisations, it may be possible for the machine to modify itself so that on a subsequent attempt the "truth" of the statement can be recognised. My point is that if the brain is describable as an algorithm, it is nonterminating (as it is primarily an algorithm for survival) and subject to arbitrary external influences that can cause the algorithm to self-modify. Thus the possibility is opened that it can be taught (or at least helped) to recognise the truth of mathematical statements in ways that a Turing machine presented with a fixed input tape as data could not.

Related to the above point is Penrose's assumption that the brain must use correct algorithms. The possibility that it might use algorithms that are useful because they generally give correct answers, but that may also give wrong answers in addition to sometimes giving "don't know" or not terminating does not seem to occur to him. Much current research is devoted precisely to studying algorithms that perform inductive inference, and by their very nature they sometimes give rise to incorrect generalisations. His argument that mathematicians use an infallible algorithm for determining mathematical truth is unconvincing given several instances in the history of mathematics of "theorems" that later turned out to be false.

The second strand of Penrose's argument is to try and convince us that a theory of quantum gravity could explain many of the properties of consciousness, as well as the puzzles of quantum theory. To do this Penrose leads us through a tour of many of the subareas of modern physics; it is here that Penrose is at his best, making his view of the problems of physics and of their solution at least seem highly plausible. This part of the book is fascinating, but Penrose's arguments as to how this gets us away from the "mind as Turing machine" notion are extremely weak, being littered with phrases such as "it seems to me that there could conceivably be some relation between this 'oneness' of consciousness and *quantum parallelism*" (p. 399) and "I envisage that . . ." (p. 446). Such a theory does not yet seem to have even the status of "tentative" in Penrose's own terminology.

Finally, although Penrose claims to be attacking the strong AI position that intelligence is a suitable program running on a digital (or equivalent) machine, this is really rather overstating the actual position of most AI practitioners. I believe that it is more accurate to say that the brain is performing some kind of information processing and that the AI venture is one of understanding and, if possible, duplicating this on an appropriate information processing machine. Until recently all known information processing machines have been Turing equivalent, but as Deutsch (1985) has shown, there is at least a theoretical possibility of other kinds of (non-Turing) computation; if it were known how to build such devices I am sure AI researchers would be only too happy to investigate the new possibilities they open up. So what Penrose has really done in his book is to advocate a move from traditional AI to what should perhaps be called Quantum AI. I doubt if anyone would disagree that this would be an interesting area to investigate.

The discomforts of dualism

Bruce MacLennan

Department of Computer Science, University of Tennessee, Knoxville, TN 37996-1301

Electronic mail: *maclennan@cs.utk.edu*

Penrose makes a Herculean attempt to give popular accounts of computability theory, special and general relativity, quantum mechanics and black holes, all with the aim of showing the relation of "computers, minds and the laws of physics." Unfortunately, the attempt seems to have failed because of a pervasive dualism. This review will discuss three issues where dualism gets Penrose into trouble.

1. Mathematical thinking. Gödel's incompleteness theorem shows that any consistent formal system has an undecidable proposition, but that this very proposition can be shown to be true by a metamathematical argument that appeals to the meaning of the proposition. Penrose attaches great significance to this fact, and uses it as the justification for many of his speculations. Because any formal system has such an undecidable proposition, and because it can always be proved by the metamathematical procedure, Penrose claims that this means that "mathematical truth is something that goes beyond mere formalism" (p. 111). From this he concludes that mathematical thought cannot be reduced to an algorithm, and hence that the mind cannot be equivalent to a computer (see also p. 118). I agree with his conclusions, but not with his argument.

First, recall that "metamathematical" refers only to the use of mathematical techniques to reason about mathematics; the metamathematical proof uses no esoteric techniques, nor does it depend on special, deep mathematical insights. In fact, the metamathematical proof is easily formalized. If Q is the undecidable proposition (constructed by the Gödel procedure) for a formal system F, and if F is a formal system powerful enough to talk about the truth of propositions in F (also easily constructed), then Q can be proved in F by a formalized version of the metamathematical procedure. Therefore, the metamathematical proof does not make use of any inherently unformalizable procedures, and hence provides no evidence for nonalgorithmic mental powers.

It might be objected that the informal metamathematical proof can be carried out once for all formal systems, whereas the formal proof requires for each formal system F a new formal system F' in which to construct the proof. This is so because informal mathematics can talk about the truth of all propositions, including those of informal mathematics. In fact, we can accomplish the same thing formally by constructing a formal system F^* capable of expressing propositions about the truth of its own propositions. This system must be inconsistent, however, because it is also powerful enough to express the Liar Paradox. On the other hand, informal mathematics is no better off, because it can also express the Liar Paradox.

The mystery to be explained is not the *power* of informal reasoning, but the pragmatic *constraints* on it, which allow contradiction to be avoided most of the time. I expect that the explanation of mathematical truth is not to be found in the Platonic realm, but in a complete interaction of formal structures and mathematical practice (as a psychological and sociological phenomenon).

This is not a view that will be congenial to Penrose's Platonism, but the empirical evidence is against the Platonic realm. First, mathematicians do not "see" the same mathematical reality. For example, standard analysis, constructive analysis

(Bishop 1967) and nonstandard analysis (Robinson 1966) each have their own versions of the real numbers; whether we find noncomputable reals or infinitesimals in the Platonic realm seems to depend on whom we ask. Nor does mathematical truth seem to be changeless as Penrose asserts (pp. 428, 445–46). If there is anything we would expect to find among the Platonic Forms it is polyhedra, yet Lakatos (1976) documents the long process by which the idea has evolved in interaction with Euler's theorem. There seems to be considerable Becoming in the land of Being.

2. Collapse of the wavefunction. Remarkably, Penrose claims that understanding the brain will require a physical explanation of the collapse of the wavefunction, and that this will depend on a yet-to-be-discovered theory of quantum gravity. He speculates that "the action of conscious thinking is very much tied up with the resolving out of alternatives that were previously in linear superposition" (p. 438). We will see that he is forced to these conclusions by a dualist interpretation of quantum mechanics.

Quantum mechanics and logical positivism grew up hand in hand; the conventional (Copenhagen) interpretation of quantum mechanics is based on an extreme positivist instrumentalism: A physical system cannot be said to be in a definite state unless it has been observed. In other words, observation causes collapse of the wavefunction. This interpretation is essentially dualist, since it takes observation to be something done *to* the universe by an observer standing outside of it. The trouble is that when we acknowledge that the observer is part of the universe, the physical significance of the reduction of the wavefunction becomes problematic, and we find ourselves confronted with Schrödinger's cat and the like.

There is a simple way out of these paradoxes, but Penrose's dualist bias will not let him accept it. If the observer is part of the physical universe, then the results of observations can be in linear superposition just as the objects are. This of course is Everett's (1957) interpretation, the so-called "many worlds" model – an inaccurate name, for there is only one world: the wavefunction evolving in accord with the unitary operator **U**. Penrose vaguely alludes to the "problems and inadequacies" of Everett's interpretation (pp. 295–96, 432), but discusses only one. His objection is that we should be "aware" of the linear superposition of observational outcomes, but that we are not. This objection fails to take seriously consciousness as a *physical* phenomenon. If we do so, then we must conclude that conscious states, like other physical states, can exist in linear superposition, but that under normal conditions there is no reason to expect these states to interact. (Presumably we could – at least in principle – design experiments to test for particular superpositions of conscious states.)

In summary, a dualist view of consciousness leads Penrose to reject the Everett interpretation, which forces him to attribute physical reality to the reduction operator **R**, and lands him in need of a new theory of quantum gravity. Dualism comes at a heavy price![1]

3. The nonalgorithmic mind. Whether or not the mind is algorithmic is one of the central questions of Penrose's book; we must accordingly consider the meaning of this question. Penrose takes "algorithmic" to mean "simulatable by a Turing machine" (p. 47). In this sense the brain is surely algorithmic, because it obeys electrochemical laws, which can be described by a huge system of differential equations, which can be – in principle – simulated by a Turing machine. This shows the irrelevance of this definition of "algorithmic," since any physical system, including the entire universe, is algorithmic in this sense. Adopting this definition leads Penrose into great difficulties because he has already concluded from considerations of the metamathematical proof that the brain can't be simulatable by a Turing machine. Fortunately, we have seen that that conclusion does not follow, and therefore that there is no problem with accepting the brain as Turing-simulatable (at the level of a physical system).

Although we have concluded that the brain is algorithmic (in the sense of being Turing-simulatable), this isn't very interesting since by this standard virtually everything is algorithmic. On the other hand, one of the principal claims of connectionism (against traditional AI and cognitive science) is that the brain is nonalgorithmic. Is there no content to this claim?

The definition of "algorithmic" that is relevant to these claims is "a physical system operating by the formal manipulation of discrete symbol structures" (cf. Newell & Simon 1976). An interesting characteristic of this definition is that it is a matter of degree – some systems are very algorithmic, others are quite nonalgorithmic, and yet others are in between. Connectionist researchers and others have made a good case that the brain is "quite nonalgorithmic," but this will not provide an escape hatch for attributing any special powers to the brain (we have seen that there is no need for them anyway). Rather, by asserting that the brain is nonalgorithmic we make the significant empirical claim that the brain operates on very different principles from a digital computer; whether one can simulate the other is irrelevant to this claim. No doubt the fuzziness of this sense of "algorithmic" will exclude it from the Platonic realm, but that is often the price we must pay for having a *useful* category.

4. Conclusions. The brain *is* nonalgorithmic, but this doesn't mean that it isn't simulatable by a Turing machine. Rather, it means that formal symbol manipulation is not a fruitful account of its action. Gödel's theorem, and in particular the metamathematical proof, do not imply that the brain has any special powers of inference or insight that are inconsistent with classical physics. Likewise, quantum mechanics does not require any special interaction between minds and the rest of the universe to accomplish reduction of the wavefunction. It is Penrose's dualist biases that drive him to this unnecessarily esoteric account of the mind.

NOTE

1. Penrose suggests that quantum mechanics would allow the brain to carry on many simultaneous computations in linear superposition (p. 438). This kind of parallelism does not require quantum mechanics, however; it can be implemented by a variety of classical processes over linear spaces; see MacLennan (1987).

Uncertainty about quantum mechanics

Mark S. Madsen

University of Sussex, Astronomy Centre, Brighton BN1 9QH, England
Electronic mail: *marksm@syma.susx.ac.uk*

The central argument of Penrose's book – that human thought can be determined only by physical processes that are beyond our present understanding of natural law – is based on an extremely tall and particularly shaky edifice. The main reason for the shakiness is the author's failure to distinguish between ignorance and evidence: The book adopts the purely speculative standpoint that the areas of physics that are not yet understood will come to support the author's point of view when they are finally understood. Why should we believe this? *The Emperor's New Mind* really raises more questions than it answers.

Penrose's discussions of cosmology, quantum theory, and quantum gravity are interesting, highlighting many of the fundamental unsolved problems of those subjects. Unfortunately, he dismisses some of the most important work done on those same topics (cf. Chapters 6, 7, and 8), work that would considerably weaken his arguments. For example, consider the important question of the classical behaviour of macroscopic objects, as mentioned on page 256. Such problems have been consid-

ered in the physics literature. In the 1930s, Mott derived the classical equations of motion for an alpha particle from the quantum theory, thus explaining why the motion of alpha particles is apparently the same as that of purely classical ones.

Inflationary universe models are also neglected. Not only are these models – presently the standard by which other models must be judged – relegated to a note (page 347), but they are misrepresented. Inflationary models (it should be noted that they represent a large class of models, not just one scenario, as implied by Penrose) do not depend on grand unified theories for their realisation. Nor do they depend on particularly special initial conditions, a topic that has absorbed most of the research effort in cosmology for the past decade. What is more, they alleviate some of the problems that would still afflict cosmological theory even if Penrose's Weyl curvature hypothesis (cf. the last equation on page 344 and page 345) were satisfied, for instance the monopole and flatness problems (Barrow & Tipler 1986).

There is a basic problem with the argumentative strategy of Penrose's book. Together with obvious omissions and misrepresentations, Penrose expects the reader to accept a lot of his argument on faith alone. To follow him all the way to the final conclusion that "it is indeed 'obvious' that the conscious mind cannot work like a computer" requires the acceptance of what seems to be a great deal of mysticism. The Platonic absolute alone will cause many readers to stop in their tracks, or even suspend their belief. And that is a shame, because the questions Penrose raises are important ones, and because he also has a large number of excellent ideas for dealing with them. Of particular interest is the suggestion of the "one-graviton limit" (see pp. 367ff), which is based on the idea that gravity itself is the observer necessary to catalyse the reduction of the quantum state vector.

I am not going to argue with Penrose's point that there are apparently noncomputable aspects to physics. The idea itself is unexceptionable, and it is easy to see that computable theories – although immensely successful in some branches of physics – are not absolutely necessary for the study of the physical world. Rather, computability is a useful heuristic test for the usefulness of a physical theory, because it allows us to compare the predictions of the theory with experiment. Of course, this procedure of prediction and experimentation is probably less important to Penrose than it is to those of us who find the idea of an all-encompassing pre-existing Platonic metareality a little excessive.

In any case, even if I can use a Turing machine to compute the evolution of a dynamical system, there is no reason I should conclude that the dynamical system has used the same method to compute its own evolution. I find it hard to believe that the particles taking part in a nuclear decay process, for example, are using Turing's methods to compute their trajectories in phase space. On the other hand, it is not totally inconceivable that the outcomes of, say, nuclear decays, could be determined by Turing-like computations performed by the particles (in some as yet unspecified manner). Typical strong interaction decay times are of the order of 10^{-23} seconds, so the computation must be completed within this time. The shortest timescale available is provided by the Planck time, which is about 10^{-43} seconds, so whatever mechanism performs the computation of the decay parameters has time to process at most about 10^{-20} bits.

This does not seem like very much in terms of the complexity of the calculations required by present models of this sort of physical system. (This result should be compared with 1 Mflop, which is about 10^{-8} bits for present day machines.) It may be enough, however, to determine all the information contained within the system at any given instant. Of course, one cannot make any definitive statements until it is known exactly how much information is contained in a given dynamical system!

Rejection or acceptance of a physical theory usually depends on whether the observed behaviour of the physical system agrees with its computationally predicted evolution. Should we not be prepared to apply the same criteria to theories of mind, and hence to give such theories a similar chance to validate their own expectations? It is strange that Penrose wishes to deny the same equality of opportunity to theories of intelligence as the ones that have been enjoyed for so long – and so fruitfully – by theories of the physical world. It is also surprising, given the speculative nature of the evidence on which Penrose bases his condemnation of the aims of strong artificial intelligence. The interests of science would be better served by giving artificial intelligence a fair chance to prove its worth as a branch of study.

Gödel redux

Alexis Manaster-Ramer[a], Walter J. Savitch[b] and Wlodek Zadrozny[c]

[a&c]*IBM, T. J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598;* [b]*University of California San Diego, La Jolla, CA 92093*

Electronic mail: [a]*amr@ibm.com;* [b]*wsavitch@ucsd.edu;* [c]*wlodz@ibm.com*

Imagine that some phenomenon of nature, say, the growth of quasicrystalline aluminum-manganese alloys, is neither random nor algorithmic (Turing-computable). What would it mean? Would we no longer believe that any nonrandom finitely specifiable physical system can be modeled computationally to any degree of accuracy?

The answer seems obvious. The Church-Turing thesis would be wrong, for the quasicrystals would be a more powerful computing device than a Turing machine. The behavior of physical systems would be seen not to be Turing-computable, but it would be computable in some new sense. Computer scientists would have a field day formalizing, and harnessing, the new principles.

These reflections are inspired by Penrose's new book, which tries to show that such physical systems exist and, realizing that current physics abhors such an idea,[1] speculates about what might be wrong with current physics. The crux is that classical physics seems to be computable, while the element of randomness introduced by quantum mechanics does not increase computational power but at most increases computational efficiency (p. 402). The way out is to assume that current physics is wrong (pp. 438–439):

> Could such a physical action be non-algorithmic in nature? We recall that the general tiling problem . . . is one without an algorithmic solution. One might envisage that assembly problems for atoms might share this non-algorithmic property. If these problems can in principle be 'solved' by the kind of means that I have been hinting at, then there is indeed some possibility for a non-algorithmic ingredient in the type of . . . action that I have in mind. For this to be so, however, we need something non-algorithmic in CQG.[2]

Penrose offers no such theory, but he would welcome one. This seems no more than what one expects: one mathematically precise physical theory yielding to a new, more complex, one.

As for specifics, Penrose mentions the alloys referred to, but, steering clear of the "considerable controversy" surrounding the physics involved, he does not "attempt to draw any definitive conclusions" (pp. 435–38). He is less reticent about the case of the human brain, perhaps because his argument here is independent of unsettled issues in physics. His claim is that, although Gödel's incompleteness theorem shows that any Turing-computable formal system encompassing arithmetic must contain true propositions it cannot prove (pp. 105–108), a mathematician can nevertheless "see" the truth of such propositions. So, the mathematician does something no algorithm can do (pp. 417–18).

Penrose is no mystic. He suggests that this feat may be

possible because "the growth or contraction of families of dendritic spines" in the brain is related to the putatively nonalgorithmic behavior of quasicrystals (p. 438). Moreover, he clearly believes that, even if such phenomena are nonalgorithmic, this simply means that they need more complex mathematical models; so he is at pains to emphasize that there is more to mathematics than the Hobson's choice of algorithms or randomness (pp. 129–40). More specifically, he still believes that mathematicians reason by means of arguments in which "each step . . . can be reduced to something simple and obvious" (p. 418), but the procedure involved would not be Turing-computable. At this point, we (unlike Penrose) would have noted that this does not contradict the usual mathematics of computation, which routinely allows for "machines" that "compute" functions that are not Turing-computable (Turing machines with oracles, infinite nets, etc.). The Church-Turing thesis does not forbid this, but merely says not to identify such "machines" with the intuitive notion of computation or algorithm. Even this is falsifiable, however, and if Penrose is right, is false. It would then be a matter of taste whether we extend the term "algorithm" to the broader class of procedures or reserve it for standard (Church-Turing) algorithms.

But all this – and more – depends on granting Penrose's argument, and this we should not do. The reason is a small but lethal flaw in his presentation, and application, of Gödel's theorem. For Gödel does not say that a certain proposition *P* is true but unprovable in a formal system *F*, but merely that *P* is true but unprovable if *F* is consistent. Penrose notes that if *F* is inconsistent then *P* is provable but false, but then makes the inexplicable mistake of assuming that, "Our formal system should not be so badly constructed that it actually allows false propositions to be proved!" (pp. 107–108). Without this, only the conditional can be proved (and this can be done algorithmically!).

The crucial thing to show would be that a mathematician, unlike an algorithm, can also derive *P* itself. But this is surely wrong, for what mathematician would jump from "*P* is unprovable but true if *F* is consistent" to *P*? To assent to *P*, he would need to know that *F* is consistent. But *F* in this case is the mathematician's own mind. If *F* proves – or assumes – its own consistency, then (by Gödel's theorem) it must be algorithmic and inconsistent, or nonalgorithmic and consistent, or nonalgorithmic and inconsistent. To conclude that the mind is nonalgorithmic, we would need to know that it is consistent, but that we do not know. Mathematicians do not even claim to know that arithmetic is consistent, though they may "hope" so. (Kleene 1950, p. 211).

So, whether the action of the human brain is algorithmic remains an open question. But if it is not, Penrose is probably right that this would have to mean that Turing-non-computable mathematics underlies our neural hardware.

NOTES

1. While the issue is controversial, some physicists go so far as to suggest computability as a methodological criterion for physical theories (Geroch & Hartle 1986).
2. The theory of quantum gravity.

Computation and consciousness

Drew McDermott

Department of Computer Science, Yale University, New Haven, CT 06520
Electronic mail: *mcdermott@cs.yale.edu*

Penrose's book consists of lucid tutorials on computer science and modern physics, some well-worn arguments against the possibility of artificial intelligence, and some vague thoughts on quantum mechanics and consciousness.

It has one intriguing, novel idea (novel to me, anyway), that physical laws might transcend computability. Imagine that a strange "electrochromatic" field is discovered some day, in which only certain systems are stable, the rest disintegrating quickly. It turns out that the simplest description of the field's behavior is that it destroys automata that would otherwise loop forever (e.g., cycle through states of high kinetic energy indefinitely), allowing only those that will eventually halt (e.g., reach some kind of equilibrium) to survive. Never mind the gaps in this description; the question is, could God have created a universe with a field like that? The gut reaction of the average computer scientist is surely, no. But it is hard to say why, unless God is a computer engineer.

Ah, well, such entertaining topics are few and far between in Penrose's book, Much of it is devoted to old arguments about the validity of the Turing test and the limitations imposed by Gödel's incompleteness theorem. Penrose is apparently unaware of the current status of this discussion. Let me bring him up to date:

1. It is absurd to suppose that a paper Alan Turing wrote in 1950 commits every AI researcher to a crude operationalist view of the mind. I am as much a realist as Penrose about whether an entity is actually thinking or feeling. I believe that the ultimate nature of these phenomena is computational, however. That belief leads to a research program. Turing's test plays no role in this program, certainly not as a necessary and sufficient condition for a machine's being capable of thought. Penrose and others think the test is insufficient; I and most AI researchers think it's unnecessary; so let's agree to put it aside and focus on the important issue, which is exactly what phenomena are explainable computationally. (The current state of progress is that not many phenomena are thus explainable, so I don't know what the critics are afraid of.)

"But without the test, how will you know when your new science has succeeded?" Luckily for Galileo, the Vatican didn't burn him at the stake for the inability to answer the corresponding question about physics. For that matter, it is just as well the question isn't being pressed now. The state of the methodology of physics in the twentieth century is more inchoate than it was in the seventeenth, consisting mainly, as Penrose clearly and candidly explains (in Ch. 6), of embarrassment about what it means to "make a measurement" of a quantum-mechanical system. The principal debate seems to be between proponents of the Everett-Wheeler interpretation of quantum mechanics (DeWitt 1973) on the one hand, and on the other hand all those (the majority, including Penrose) who are trying to stave it off. From my point of view, that of a humble outsider, the Everett-Wheeler interpretation seems obviously correct, but I can well understand why its opponents want to avoid its implication that the universe is continually splitting into unaccountable copies of itself. I am glad the worst methodological problem AI has to face is how to get people to forget about the Turing test.

When all is said and done, I find it hard to understand exactly what Penrose's objection is to the use of the test. Penrose's test is in some ways easier to pass: "All I would myself ask for would be that our perceptive interrogator should really feel convinced, from the nature of the computer's replies, that there is a *conscious presence* underlying these replies – albeit a possibly alien one." (p. 9) I would ask for more. If all cognitive science can come up with is a furry automaton that people feel they can confide in, without a theory of what consciousness really is, then it will have failed.

2. Gödel's theorem is irrelevant to the question of how well computers could think about mathematics.[1] On second thought, Gödel's theorem provides evidence *in favor* of AI. Here's the argument:

i People are unable to prove or disprove every statement in number theory.

ii So are computers.

iii Ergo, people and computers are alike in at least one way.

That's very weak evidence, but it's in the right direction. Of

course, someone might dispute Premise i, but surely not a Platonist like Penrose.

Here is another argument that Penrose might prefer:

a. Suppose you are an algorithm that, given a conjecture in number theory, correctly decides whether it is true or loops forever or gives up.

b. It can be shown that there is a true conjecture that will cause you to loop forever or give up.

c. Any person can understand the argument alluded to in *b*, so you will both come to believe the conjecture and not come to believe it, which is a contradiction.

d. Therefore, you are *not* an algorithm that, given a conjecture in number theory, decides whether it is true or loops forever or gives up.

This is an awfully sophisticated argument to arrive at such an obvious conclusion. Those who are impressed by it think that it would survive if Assumption *a* were replaced by the assumption: "Suppose you were any algorithm whatever." But it does not. Suppose that AI succeeds in building an artificial mathematician with abilities comparable to those of a human mathematician. It might or might not have a subroutine of the sort discussed by Penrose, that is, a semidecision procedure for Peano arithmetic. Suppose it did have such a routine. Then it might have another routine that could appreciate the argument about the limitation of the first routine. "But surely the combination of the two would have gaps in *its* understanding?" Of course it would, but not because of the argument above. The enlarged system already violates Assumption *a*. And we haven't even made use of all the other abilities this human-competitive program would have, including the ability to carry on a conversation with other mathematicians and learn new limitations of the formal systems it had been making use of. To top it off, any plausible candidate for an algorithm that duplicates a person would, far from being an infallible procedure, have incomplete and even contradictory beliefs about mathematics. This last feature makes it unreasonable to think of its reasoning pattern as modelable by a formal logical system at all.

If your eyebrows rise at this possibility, you have fallen victim to a seductive fallacy, of identifying the formal system that a computer program *is* with a formal system it *reasons about*. The insight that a digital computer is a formal system is quite powerful theoretically (for a good exegesis read Haugeland [1985]). But this insight plays little role in thinking about actual programs, beyond the consideration that if the machine fails to make the next move in the particular formal game it embodies, you call the repairman. That's why we have to prove that a program that does deduction is sound; it is all too easy to write a perfectly formal system that draws informal and incorrect conclusions.

An echo of this fallacy is found in Penrose's argument that all algorithms for reasoning about mathematics must be equivalent (p. 417), an argument that confuses proof verification with proof generation. Even granting that all mathematicians can understand and evaluate each other's proofs, they could still embody different algorithms when it comes to searching for these proofs. Any real algorithm for searching for proofs will consist of lots of idiosyncratic little strategies that work now and then, together with various (time-varying, environment-dependent) rules for when to try which strategy. Such an algorithm is absolutely sure to be incomplete and even error-prone, just like a human mathematician. In other words, an AI account of mathematical reasoning is likely to be similar to an AI account of story understanding, and we're far from a theory of either. Formal arithmetic certainly does not supply such an account.

Finally, we have this coda to the arguments from undecidability:

> One often strives for algorithms, . . . , but the striving itself does not seem to be an algorithmic procedure. Once an appropriate algorithm is found, the problem is, in a sense, solved. Moreover, the mathematical judgement that some algorithm is indeed accurate or appropriate is the sort of thing that requires much conscious attention. (p. 413)
>
> Mathematical truth is *not* something that we ascertain merely by use of an algorithm. I believe, also, that our *consciousness* is a crucial ingredient in our comprehension of mathematical truth. We must 'see' the truth of a mathematical argument to be convinced of its validity. (p. 418)

This way of putting it just misplaces the level at which the hypothesized algorithms are presumed to operate. If we can get a computational theory of mathematical reasoning, it won't predict that mathematicians *feel like* Robby the Robot. It *will* say what it means to "see" the truth of a mathematical argument. Of course, no one has any idea what such a theory would look like.

Without these old arguments, what does Penrose have? Actually, the argument he really wants to make carries a lot of weight:

> Is it not 'obvious' that mere computation cannot evoke pleasure or pain; that it cannot perceive poetry or the beauty of an evening sky or the magic of sounds; that it cannot hope or love or despair; that it cannot have a genuine autonomous purpose?
>
> Some of the arguments that I have given in these chapters may seem tortuous and complicated. Some are admittedly speculative, whereas I believe that there is no real escape from some of the others. Yet beneath all this technicality is the feeling that it is indeed 'obvious' that the *conscious* mind cannot work like a computer, even though much of what is actually involved in mental activity might do so.
>
> This is the kind of obviousness that a child can see –. (pp. 447–48)

What this argument amounts to is pointing out that the burden of proof for the question whether computer consciousness is possible lies on those who think it is. But I guess you can't write a whole book pointing out something as obvious as the failure of computationalism to say much about consciousness as yet.

Penrose appears to think that the computationalist position on the question of consciousness is that it will simply emerge from complexity: "The viewpoint of strong AI[2] . . . maintains that a 'mind' finds its existence through the embodiment of a sufficiently complex algorithm. . . ." (p. 429) "Consciousness . . . is something just 'accidentally' conjured up by a complicated computation. . . ." (p. 447). If he had taken a survey of AI researchers, he would have found almost none who agreed with that view. He would have found many who agreed with him that AI will not account for consciousness without new ideas; many who acknowledge confusion on the matter; and some who believe that AI's current ideas at least deserve a shot at explaining consciousness. I would put myself in the last group, and would hasten to issue a disclaimer that taking this shot involves almost as much speculation as Penrose expends in explaining consciousness with quantum mechanics. The point is that *some explanation is necessary*, however. Consciousness is not simply going to emerge from complication. There must be specific computational structures that underlie it. My belief is that a system is conscious because it has a model of itself as conscious (that is, as an agent with a single stream of thought, infallible observations of its own state, and exemption from causal laws in making decisions). A system could be quite intelligent, I suppose, and still not be conscious, because it failed to have such a self-model.

Penrose would doubt this, because he subscribes to an amazingly naive brand of introspectionism on the subject. In the quote above, he lumps together pleasure, pain, esthetics, purpose, love, and despair as all part of consciousness. Elsewhere he tends to assume that the "consciousness faculty" drives judgment and intelligence. He spends much time explaining away data that contradict this theory, data that tend rather to support the idea that consciousness is a game the brain plays with itself. To take one example, people's conscious impressions of the exact times when sense data are recorded and decisions are made tend to be out-of-synch with reality by exactly the

amount you would expect if the actual psychophysiological events the conscious impressions reflect took an appreciable amount of time before consciousness got involved. For example, it takes 500 msec to become aware of a stimulus, and the conscious impression gets "backdated" 500 msec to compensate. [See Libet: "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action" *BBS* 8(4) 1985.] Penrose blushes, makes another apology for why the all-powerful consciousness seems to be so out-of-it, and then brings forth speculation about quantum mechanics and time.

Or consider Penrose's objection to the possibility that the brain is a parallel computer: "A characteristic feature of conscious thought . . . is its 'oneness' – as opposed to a great many independent activities going on at once." (pp. 398–99) Forget about AI – the author needs to brush up on William James and Sigmund Freud! It's very hard to take him seriously on the subject of consciousness.

Let me conclude on a more positive note. I concluded above that the known limitations on formal arithmetic did not automatically extend to all computational processes. But that does not mean there are no limitations on computers' ability to reason about mathematics; we simply don't know what the limits are. If there are significant limitations, and humans don't have them, then humans aren't computers. They might not be physical systems at all, but that conclusion is unfashionable. Assuming they are purely physical (which Penrose assumes), if they can beat computers at some kind of reasoning, then there are physical systems that can. So we would be able (presumably) to build machines that compute uncomputable things. These machines wouldn't do computation, of course, even though their outputs would be interpretable as computational results. (If you ask, How do these machines work?, the answer would be, There's no "how" to it; a "how" would be a computation.) I mentioned the possibility of such physical systems at the beginning of this review. I assume that the possibility is extremely unlikely, and hence that any limitations on computerized mathematical reasoning also apply to people.

It is meaningless to talk of such limits unless the mathematical world exists independently of what people, or computers, think about it. Penrose discusses this issue at several points, and I tend to agree with what he says. In particular, I think most AI practitioners have been insufficiently impressed by the following Penrosian point:

> For an algorithm to 'exist' independently of any particular physical embodiment, a Platonic viewpoint of mathematics would seem to be essential. It would be difficult for a strong-AI supporter to take the alternative line that 'mathematical concepts exist only in minds', since this would be circular, requiring pre-existing minds for the existence of algorithms and pre-existing algorithms for minds! . . . Accepting, then, that algorithms inhabit Plato's world [the world of mathematics], and hence *that* world, according to the strong-AI view, is where minds are to be found, we would now have to face the question of how the physical world and Plato's world can relate to one another. This, it seems to me, is the strong-AI version of the mind-body problem! (p. 429)

I might add that AI practitioners have in general shown a strange willingness to believe that concepts are imposed on the world by minds, and not objectively real at all, while at the same time, as Penrose says, being committed to the idea that minds themselves are objectively present, and that their contents are formed by well-defined encounters with the world. Clearly, the latter commitment is deeper than the former innocent phenomenalism; if a computational account of consciousness is ever obtained, one of its philosophical implications will be a radical realism. For instance, Cartesian doubt would be decisively, and *empirically*, refuted by a theory that predicts the presence of consciousness in physical systems with a certain computational structure. If you are bothered by that prospect, or any other consequence of the success of AI in explaining the mind, rest assured that progress to date is small. For the time being, you needn't lose any sleep over the possibility of conscious machines.

NOTES

1. Most of the arguments that follow are due to Marvin Minsky (personal communication), and some appear in *The Society of Mind* (Minsky 1986).

2. "Strong AI" is a term Penrose borrows from Searle (1980), and which is synonymous with what I call "computationalism," the doctrine that thought is computation. "Weak AI" is, as far as I can tell, a trap for the philosophically unwary cognitive scientist, and not worth discussing.

The powers of machines and minds

Chris Mortensen

Department of Philosophy, The University of Adelaide, North Terrace, South Australia 5001, Australia

Penrose flirts with arguing from Gödel's first incompleteness theorem to the conclusion that the mind is not a Turing machine. This argument has a long history of defenders, going back through Lucas, Nagel and Newman, to Gödel himself. At its crudest, it argues from the premise that Gödel's first incompleteness theorem states that for any Turing machine machine (or any formal axiomatisable theory containing at least Robinson arithmetic in which all primitive recursive functions can be represented) there is a true sentence that the theory cannot prove or disprove, or the machine cannot effectively verify or refute. The conclusion is that since humans can know that such sentences are true (by the Gödel theorem), humans cannot be equivalent to any such machine.

I have no axe to grind about whether we are Turing machines, but I am concerned to believe only conclusions supported by sound arguments. It has been clear what is wrong with this argument since Putnam's (1960) seminal paper, although the argument has produced a considerable and often confused literature since. What the first incompleteness theorem says is that *if the arithmetical theory is consistent* then there is a true sentence unprovable in the theory. (Strictly not even that, only that if the axiomatisable arithmetical theory is consistent then there is a sentence *A* where neither *A* nor not-*A* are theorems of the theory; the further step that one of *A* and not-*A* is true can be challenged, but let us let it go.) Hence, to know that the Gödel sentence is true one needs to know that arithmetic is consistent. (Indeed, given that the Gödel sentence represents "I am unprovable," then it is true only if arithmetic is consistent.) But by Gödel's *second* incompleteness theorem, no recursively enumerable system capable of representing arithmetic can prove its own consistency. Thus the premise that the Gödel sentence is true (and unprovable) cannot be known unless it is known that arithmetic is consistent, and no Turing machine can know the latter. So who says that humans can know it either: The argument begs the question. One might add that the theory of inconsistent mathematical systems is by now well enough developed for it not to be so catastrophic if arithmetic did turn out to be inconsistent.

Be that as it may, there is a problem about how mathematical knowledge, or at least nonfinitistic mathematical knowledge, is possible. I have always felt that Skolem's paradox is a real puzzle: We feel that we have genuine knowledge of nondenumerably infinite structures such as set theory or the real numbers, which is yet expressible in first order language. But by the Lowenheim-Skolem theorem, any first order theory with an infinite interpretation has a denumerable interpretation. Wherein consists our understanding of the nondenumerable then? Are we second order creatures? That seems unlikely for finite nerve nets. More generally, how is knowledge of analysis possible for finite nerve nets? How is knowledge possible of immense structures like the hyperreal numbers (real, infin-

itesimal, and infinite), let alone knowledge of a theory which can generate the Bell results, as elementary quantum theory can? (Indeed how is a Bell-nonlocal universe possible?)

The trouble with these Kantian questions is that it needs to be shown that we do have genuine knowledge of these structures that exceeds the merely first order, or constructive, or Turing-computable, or whatever the target power. The history of mathematics is partly the history of tacit agreement on hidden mathematical principles being eventually brought to conscious mathematical scrutiny, as Lakatos (1976) persuasively argues. It takes enunciation of such principles to clarify and scrutinise them, and enunciation looks suspiciously effective and Turing-computable, or at any rate performable by a finite nerve net or silicon neurochip. Why isn't dealing with the real numbers just dealing with explicit and tacit axioms and sentences? It has to be demonstrated that there is a special phenomenon to be explained here, and it is not at all obvious that there is. I suppose that if we did have, say, some sort of mystical direct perception of the whole set of real or hyperreal numbers, then our powers would demonstrably exceed those of microchips. If such a fantastical hypothesis were true, then perhaps a sophisticated theoretical construction like quantum gravity would have to be invoked in explanation; it is doubtless better than many other explanations. The trouble is that quantum gravity is so *unlike* the hyperreals. *Thinking* about the hyperreals or analysis is surely much more like the *sentences and pictures* that are printed in books about the hyperreals. Would it be so far wrong to say that it is *just* sentences in the head with appropriate causal linkages to behaviour? And if it were far wrong, what would need explaining is the efficacy of mathematics texts and lectures as communication channels between mystical senders and receivers. Again, the functional architecture of knowers looks to be nerve nets, which is much more like the functional architecture of microchips than like atoms.

I do not think that it is just sentences in the head. At the very least, sentences are highly structural properties of networks; though such "emergence" is perfectly natural, since it is nothing more than a many-placed relation and we are familiar with these in nature (*e.g.*, betweenness). The role of visual or spatial models cannot be ignored either. Geometrical representation in mathematics is undoubtedly not the same thing as verbal representation, symbolic representation being an intermediate case. It is an error to conclude that the spatial can be reduced to the verbal or symbolic on the grounds that there is some isomorphism between certain of their aspects. The fact is, the spatial is *not* the verbal. Cartesian coordinates, pairs of real numbers, are *not* the plane, they merely represent it. Our thinking is saturated with the sensory and the spatial, as I have argued elsewhere (Mortensen 1989). But then it surely does not exceed the powers of a finite nerve net to represent the spatial in a spatial way, any more than it exceeds the powers of a television set to do so.

What does come out of quantum theory on this point, I think, is a doubt about whether we finite structures can understand *any* precise mathematical concept. If so then the Gödel premise that we know the Gödel sentence to be true is in even worse shape. This needs as a premise the Copenhagen interpretation of elementary quantum theory, which I would be willing to defend: that the Heisenberg uncertainty relations indicate genuine indeterminateness or incompleteness in nature. That is, cells and nerve nets are a little fuzzy around the edges. A second premise is that understanding-knowledge-memory require *inner representations* (albeit grossly holistic), with appropriate causal links to the world. Now, for an inner structure to be a representation it must have at least as many casually active features as that which is represented: Understanding aspects of an idea requires potential complexity of output matching the complexity of the aspects (even if the output is merely symbolic). Understanding a precise idea, therefore, requires precision in the understander-structure matching the precision of the understood aspects of the idea. But by the first premise no understander-structure is wholly precise. Therefore, no actual nerve net or neurochip could fully understand a precise mathematical idea. There are obvious places where this argument could be challenged, but I will not pursue them.

A last point emerges from chaos theory. It is fashionable to present chaos theory in a paradoxical way: that wholly deterministic and computable systems can be chaotic. The point, however, is that the concept of chaos has shifted from an older meaning (something like "undetermined"), to a newer *epistemic* meaning ("unpredictable" or "unknowable"). Chaos theory arises from the discovery of (simple, Turing-computable) functions that are very sensitive to small changes in initial conditions, especially after a few times around the loop. The chaos is that our *knowledge* always contains an error term or uncertainty (quantum theory is sufficient but not necessary for this premise). So the error term in our knowledge of such systems grows rapidly, becoming unpredictable with any degree of precision. My point here is that the recognition of the epistemic limitation of the powers of human knowers is at the foundations of chaos theory, which makes it even harder to believe strong claims about how the powers of humans exceed those of microchips or nerve nets.

A final plug for the Turing machines. Chess programs are now up to grandmaster level, though they do not do it quite the way we do it. The best kind of mathematical proof is the one which makes the alternatives exclusive, exhaustive, and surveyable. We may be fostering the emergence of a genuinely higher intelligence (see Gibson 1984).

Steadfast intentions

Keith K. Niall

Simulation Group, Human Factors Division, Defence and Civil Institute of Environmental Medicine, Downsview, Ontario, Canada M3M 3B9
Electronic mail: *kn@dretor.dciem.dnd.ca and kn@zorac.dciem.dnd.ca*

What can be achieved in psychology in the absence of a complete knowledge of physics? Penrose (1989) expects a nascent physical theory – "correct quantum gravity" – to explain how it is that people can perceive. In other words, an advance in physics should illuminate a problem in psychology. Yet the theory of quantum gravity will make no psychological explanation manifest, if only because the fundamental problems of psychology are not problems in physics. Problems in psychology are marked by intentionality, and physics makes no mention of intentionality (Chisholm 1957; McAlister 1976). And in psychological explanation: "Everybody has to face the issue about intentionality somewhere" (Fodor & Pylyshyn 1981, p. 192). Suppose that intentionality might be described in terms such as "mass," "length," and "time." Then still, psychological explanation requires some account of intentionality, whereas correct quantum gravity does not promise any such account. Perhaps psychologists should not wait for a revolution in physics before they explore the fundamental problems of perception and cognition. Two reasons not to wait will emerge after some consideration of the intentionality of vision.

Most accounts of the intentionality of vision begin as descriptions of *what is seen*. Gibsonians claim that physical objects are "perceived directly" (Gibson 1979, p. 263), whereas empiricists claim that proximal stimuli are "present to the mind" (see Yolton 1984, p. ix). These are small figures of speech; one might as well say that perception is "directed" (cf. Cutting 1986) in a way that masses are not. As Penrose says, "To be conscious, I have to be conscious *of* something" (p. 405). Such figures of speech convey no insight into *how* we may perceive things about the world. Is any explanation needed? Is it not a simple fact that we perceive physical objects? We do see what lies before our eyes, for example, mountains, chairs, dogs, rainbows, shadows, and

sometimes a part of the Milky Way. The explanation of this fact is not at all simple; so far, it is enigmatic. (One may consider that this fact entails that few people except neurosurgeons see that which is encased in the meninges of the brain.)

A first reason to distinguish perceptual psychology from the study of physics concerns the description of what is seen. Some confusion about what is seen arises because the description under which something is seen (roughly, what it is seen *as*) is not distinguished from a description of the physical conditions that allow light to arrive at the eyes. A description under which something is seen is a description of an *intentional object*, while a description of visible surfaces and the optical conditions under which that thing would be seen is a description of a *material object*. The description under which something is seen is called "intentional"; among other reasons, this is because equivalent descriptions are not necessarily substitutable in answers to the question, "What do you see?" without changing the truth of the answer. The distinction between the intentional and the material object of vision is not a distinction between kinds of matter, or between bits of matter. The terms "intentional object" and "material object" are meant to indicate semantic categories, just as "direct object" and "indirect object" indicate grammatical categories (Anscombe 1981, p. 9). These terms distinguish between two kinds of answer to, *"What is seen?"* One kind specifies the intentional object of vision, another kind specifies the material object. These answers should coincide; most often they do. Then the same body is both intentional and material object of vision; that is, most often both answers provide true descriptions of the same thing. (The logic of questions and answer is no more than a *pons asinorum* to understanding the intentionality of vision, cf. Searle 1983, p. 5). There is a logical relation between statements that convey material objects and statements that convey intentional objects: A statement that specifies a material object *implies* a statement that specifies an intentional object, but the converse does not hold of necessity (Anscombe 1981, p. 17). In other words, *describing the intentional object of vision is logically prior to describing the material object of vision*. Physics is not concerned with intentional objects, and the descriptions offered by physics do not imply intentional descriptions. Yet vision is intentional, and the study of intentionality elucidates vision. This seems a reason to suppose that "a coherent and appropriate physical theory" (p. 10) should not aspire to explain perception in the absence of a theory of intentionality.

A reason to distinguish part of cognitive psychology from the study of physics emerges as a consequence of Penrose's beliefs about mathematical knowledge. He claims (pp. 108, 116, 118) that mathematicians *see* "absolute, external, and eternal" truths, either by insight or intuition. Penrose did not invent this analogy to sight, known as "Gödelian perception." Of course, the analogy is not based on the physical effects of light. This raises a classic problem: How may we know mathematical truths if the objects that ground such truths are atemporal and nonspatial? If Liouville's theorem (p. 181) would be unaffected even by the end of the universe (in a final big crunch, see p. 325), how did mathematicians acquire knowledge of the theorem by physical means? Penrose's burden is to explain the analogy of mathematical discovery with vision, *after* he has restricted his explanation to the vocabulary of physics. In contrast, Gödel stresses psychological or epistemological themes of the analogy.

Gödel claims that similar abstract notions are involved in perception and in mathematical discovery. He supposes that the "idea of object" is given in perception, but not in sensory impressions; such ideas are "abstract elements contained in our empirical ideas" (Gödel 1964, p. 272). Gödel's analogy is based on this "idea of object," rather than on sensory impressions. It is not at all clear how the "idea of object" may be reduced to such physical terms as "mass," "length," and "time." Penrose is unclear how physical theory could explain the discovery of mathematical truths otherwise. He does attempt to explain mathematical discovery by what he calls a *"reflection principle."* The reflection principle is not part of physical theory either, since it is explained in terms of *meaning* (p. 110, see also pp. 104, 107). The epistemology inherent in Penrose's claims does invoke notions like "idea of object" and "meaning." The second reason to distinguish psychology from the study of physics, then, is that psychological notions are used in the explanation of mathematical discovery. These *abstract* notions are not reduced to physics.

There is much to be done in psychology yet. One important task to explain the intentionality of perception, for "intentionality won't be reduced and won't go away" (Putnam 1988, p. 1). As psychologist, Penrose's task is not easy, but he might do better not to wait for a revolution in physics.

The emperor's old hat

Don Perlis

Department of Computer Science, University of Maryland, College Park, MD 20742

Electronic mail: *perlis@cs.umd.edu*

Penrose presents largely old and fallacious arguments:

1. *The extrapolation fallacy.* Niels Bohr and Max Delbruck said they felt that the explanation of living things would require the discovery of new physical principles about the organization of matter. Life turned out to be more a matter of fantastic and unsuspected features of suitably complex chemical tinkertoys, however, explainable in terms of existing physics. The extrapolation from simple to complex molecules was false: Simple ones do not reproduce or metabolize, but complex ones can. The same sort of phenomenon may hold for mind, in ways we do not yet see.

2. *Strong AI and inner experience.* Penrose leaves the reader with the impression (see his Prologue and Epilogue) that strong AI is hostile to touchy-feely things like actual inner experience. This is far from the case: Strong AI suggests that inner experience is indeed there, and when we find it we will see that it amounts to very, very complex information processing.

3. *Dualism.* Along with Searle (1980), Penrose (pp. 21–22) seems to find a form of dualism in strong AI, namely, in the supposed claim that it is pattern (software, mechanical algorithm) and not substance (hardware or brain) that constitutes mentality. But mere pattern by itself is not even a process, it does nothing. It is algorithmic *processes*, not static printed copies of algorithms, that AIers are concerned with.

4. *Intentionality.* Penrose (p. 406) correctly observes that one cannot just feel or wonder or think without there being something that one feels or wonders or thinks, that is, the mind has directedness or *intentionality*. Penrose and Searle claim that the person-plus-program system in the Chinese Room does not mean (refer to) anything by its Chinese squiggles and squoggles, that is, there is no intentionality. This is again the extrapolation fallacy: A suitably *complex* person-plus-program system may indeed understand Chinese.

5. *Self-awareness.* Penrose's video-camera counterexample (p. 410) as to what self-awareness is *not*, is a woefully impoverished one: Imagine instead a vastly more complex device, one that *uses* its inner model to guide its behavior. It is then not so clear that it has no self-awareness. Again the extrapolation fallacy. (See Perlis 1987; 1989 for more on self-awareness and intentionality.)

6. *Seeing the truth.* Penrose says that for any consistent formal system of the right sort there are true well-formed formulas (wffs) that we can see to be true but the system cannot prove (pp. 108, 110, 417). This is half correct: There are true wffs that the system cannot prove – and in fact, if the system is ample enough (say, if it is ZFC, i.e., Zermelo-Fraenkel set theory with

the axiom of choice) then it can even prove that fact (i.e., the system will have a proof of Gödel's theorem!). But it does not follow that we can "see" such wffs, in the sense of picking them out and proving them true. In some cases, to be sure, we can do this, but in most of those cases so can ZFC. What is needed to get our hands on an actual instance of the Gödel sort is, typically, knowledge of the consistency of the system in question. Now, this is emphatically not something we can in general see. For instance, even today we have no proof of consistency of ZFC, and it is generally believed that there can be no such proof except by methods that essentially beg the question by assuming what they need to prove.

Do we somehow intuit that certain formalisms are consistent? This would be an intriguing possibility, and perhaps Penrose can be taken as suggesting this (pp. 418ff). But it will not hold up as a kind of "seeing the truth." ZFC may yet turn out to be inconsistent; certainly various other formalisms that at first seemed just fine have turned out to be inconsistent (Frege felt confident that his axiomatization was consistent, for example, yet Russell showed otherwise). The general belief that ZFC is consistent is really evidential-heuristic, that is, it works well, is very useful, no one so far has found an inconsistency, it is based on principles that, in special cases at least, are part and parcel of everyday mathematics, mathematicians tend to have strong intuitions as to the clarity of these, and so forth. I am not saying that I think ZFC is inconsistent – on the contrary, I think it is not. Moreover, it is conceivable that a suitable formalism (e.g., that of Elgot-Drapkin 1988) could even weigh the changing evidence for its own consistency.

Does this mean, then, that if our minds are in fact machine processes, we can never grasp the full truth about reality, assuming (as Penrose suggests) that the latter involves nonmechanical complexities? This need not follow. For one thing, we interact with reality, we do not continue forever with a fixed data set. If a Turing machine's tape is allowed to have varying input over time, then the usual limitations do not apply, at least not in the same way. Second, our brains have been evolving, and may well continue to do so, especially if we consider that technological breakthroughs (e.g., computers) and fundamental scientific discoveries can lead to new notions of human endeavor. But the change is not from any internal force of "seeing" truths; rather it is new external data, in interaction with our brains, that leads to new brains.

This notion of getting to know things by external contact bears on Penrose's *reductio* (pp. 417–18) of the possibility of there being one grand mathematical formalism that all mathematicians follow. As we investigate new mathematical terrain, we find new principles that become incorporated into our thinking, and at that time we "grow" by enlarging our formalisms. Our formalisms enlarged significantly when Fermat introduced proof by induction. We then became able to see new truths we could not see before. "How did Fermat come by it, then?" Penrose might ask. Well, presumably, Fermat interacted with just the right kinds of examples to trigger his brain in just the right way. Also, proof by induction is itself provably correct in ZFC, so it is possible that something like ZFC is the grand formalism Penrose denies. That might mean that mathematicians were able to recognize and accept proofs based unconsciously on ZFC without their actually knowing ZFC's axioms, until Zermelo and Fraenkel introduced them explicitly in this century. Of course, future world interactions might push us further, beyond ZFC.

Penrose urges his notion of judgment (of truth, for instance, p. 411) as the hallmark of consciousness, which, he says, "the AI people would have no concept of how to program on a computer" (p. 412). In fact, a major subfield of AI is nonmonotonic reasoning, which is precisely the topic of jumping to plausible conclusions based on relevant evidence (see Ginsberg 1987). This is recognized as a very hard topic, but it is not a nonalgorithmic one.

Systematic, unconscious thought is the place to anchor quantum mechanics in the mind.

Thomas Roeper

Department of Linguistics, University of Massachusetts, Amherst, MA 01003

Electronic mail: *roeper@cs.umass.bitnet.*

Penrose makes two points that are not represented in current computer models: (1) that the nonalgorithmic domain of mathematics should be a model for the mind, (2) that nonlocal and nondeterministic effects must be present in the mind.[1] But how far do these ideas, with which we agree, go? Whereas it is true that quantum mechanics introduces a kind of indeterminacy with physics and true that absolute determinism would seem to be incompatible with the notion of free will, is the tyranny of raw chance any closer to how the mind works? No single mental product (even an imaginary one) receives a clear quantum analysis in this book. Penrose's focus on the "process" of mind is, of necessity, often anecdotal or relevant only to the fairly primitive knowledge we have of physiology. The mentalist side of cognitive science (which many take to be the heart of the field) is ignored, namely, language and vision, where the most sophisticated "products" of the mind are reasonably well-understood within a deductive framework.

Penrose shows that computer simulations fail to simulate. But is simulation, no matter how good, ever equal to understanding? Suppose we could produce a computer which did have nonlocal effects, or nondeterminism; we might still not understand what it says about the mind. Machines can generate "chaotic" effects but the fact that humans produced the machines and their programs does not mean that we understand the principles behind chaos theory. Or from the opposite perspective, is it right to assume that the mind is a single machine? Suppose one had six differently construed machines which together gave us insight into human nature, but one is unable to combine them into a single machine using the same principles (like different "operating systems"). Would this perhaps support "modular" views of the mind? Thus the machine metaphor could fail to be revealing (as in "chaos" theory) successful in replication, or be revealing by its failure (because of modularity).

In other ways, the computer model, even if it worked perfectly, might give little scientific insight. Penrose shows that the computer could not "represent" certain kinds of noncomputable phenomena. Nor can humans. Nevertheless, the computer might be able to "refer" correctly to noncomputable quantities (just as humans can). This, in turn, could be sufficient to allow us to make the nonalgorithmic judgments he seeks. But does successful programming constitute principled understanding? Chomsky (1986) distinguishes between "leading principles" and "modes of execution" in linguistics. The intellectual poverty of AI lies in the assumption that modes of execution are the deepest insights available.

The acceptance of the Turing test as a definition of intelligence may be the essence of the problem. Instead, one needs well-defined mental fragments where the structure and mechanism of thought, not just its appearance, is represented. Language is a tiny molecule in a biological mosaic, but perhaps it has the essence of all principles of mind built in, just as DNA structure is repeated in every cell.

Linguistics distinguishes between automatic but unconscious programming and slow problem-solving ability which is at best partially conscious. It is not clear that conscious mental activity is more than the shimmering surface of thought with no truly special properties. Perhaps the conscious/unconscious distinction corresponds to the computable/noncomputable distinction, but this would have to be demonstrated, not assumed. It is free will, evident in an instant, that is the conundrum of being human, the source of our sense of dignity. In 50 milliseconds we

can form an image that is affected by our personalities. In less than a second we can produce or comprehend a sentence, replete with Freudian slips, to which every aspect of the mind is connected.

Has any science succeeded without making a distinction between possible and impossible? Linguistic theory distinguishes a class of impossible thoughts (universally ungrammatical sentences) from possible thoughts (grammatical sentences), while still exhibiting infinite creativity in a fraction of a second. For example, all languages allow an infinite distance between a question word and its origin (What did John say he wanted Mary to tell Bill to get Fred to do?). But no language allows an origin from inside relative clauses (what did you see the man that ate?). The capacity to invent new objects of reference (like quantum mechanics), label them, and then comprehend them in split-second linguistic references, is another example. Linguistics demonstrates that sentences are recursively generated, language acquisition is constrained, and transformations occur over infinite strings. As language is ultimately biological, these observations in turn give biological legitimacy to notions like lambda extraction, conditions on analyzability, and other concepts. Because we cannot identify an impossible thought at the slow, problem-solving level (though there might be many, we just cannot know about them without thinking them), discussions of the conscious mind are usually vague and imprecise. Unconscious, rapid thought is therefore more germane.

An important hypothesis is required to sustain this line of exploration: that there is no principled distinction between unconscious and conscious thought. This could be wrong. But if we are looking for quantum mechanical effects, unconscious thought seems like the natural domain.

A high-risk scientific gambit is also involved: Can we go directly between the mathematics of physics to the mathematical representation of linguistics, while skipping the physiology of language (where only gross effects are currently detectable)? Chomsky has argued (1) that one must have a definition of the mind before one can look for its physical instantiation, and that (2) the mentalist representation may ultimately be accepted as a physical representation just as many mathematical formulations are accepted as statements of physical law. But the challenge remains: We can, at the moment, see no case where quantum mechanics is necessary.

One final comment. Why does Penrose devote his talents to the obviously weak nature of strong AI instead of exploring the strengths of weak AI? Perhaps we can answer the question in part. Does he share the widespread sense of being insulted as a human being by the computer metaphor? Even if the scientific posture of the strong-AI community were eminently defensible, another problem exists. By allowing the implication that "free will is an illusion" to stand unchallenged, we let stand an insult to our sense of dignity with vast harmful consequences. Why do we react strongly against it? There are many current social consequences: The image of the human as a machine influences every teacher who grades a child, every child who hears his IQ score, every doctor who too readily prescribes medicine, every uneducated person who stammers in the presence of a scientist. It does a very serious disservice to the notion of human dignity to suppose that we know "scientifically" what human beings are like when we do not. If free will exists, then there can be no exhaustive "explanation" of human behavior by reducing it to a set of components with fixed interactions.

NOTE

1. David Griffiths, Department of Physics, Reed College, contributed many ideas to this review. Although he did not wish to be a co-author, I feel as if his ideas and my ideas are all actually his. So he is present and absent at the same time. Quantum mechanics?

Seeing truth or just seeming true?

Adina Roskies

Department of Philosophy, University of California, San Diego, 3-002, La Jolla, CA 92037

Electronic mail: *adina@helmholtz.sdsc.edu*

The reigning king of cognitive science, strong AI, is in trouble. It still wears the royal vestments, but some subjects question whether its mental abilities are up to its demanding task. What happened to the emperor's mind? Penrose tries to dethrone strong AI with the argument that minds are nonalgorithmic, and thus that strong AI can never achieve its aims. Should we rally behind him?

Real minds are conscious. The hallmark of consciousness, according to Penrose, is the ability to directly perceive (mathematical) truth (p. 428). I want to examine this claim in detail, after pointing to a general flaw in Penrose's arguments that our ability to see mathematical truths can be accounted for only nonalgorithmically. Briefly, Penrose conflates the ability to solve an instance of a noncomputable problem with the ability to solve all instances of that problem. Only the latter would entail solving it nonalgorithmically. There is no convincing evidence that humans can always solve these types of problems, so Penrose does not succeed in establishing that human minds do things that algorithms can't (For a thorough discussion of these points, see Benacerraf (1967) and Dennett (1978a).

What would it mean to perceive mathematical truth directly? Would it entail infallibility? It ought to, for the notion of mathematical truth is closely related to the notion of necessity, and direct perception implies (minimally) preserving truth values. According to standard accounts, knowing that something is true involves two components: (1) knowing the meaning of something, and (2) knowing that what the meaning expresses is the case. Our intuitions about the self-evidence or truth of mathematical statements, however, often turn out to be mistaken. For instance, Hilbert believed that arithmetic was complete until Gödel proved him wrong, and Cantor and Frege thought their formulation of set theory was consistent until Russell advanced his paradox. The moral is that "seeing something to be true" is indistinguishable from "something seeming to be true" until a proof is given. Penrose is a Platonist, and consequently has no account of (2). As a result, *seeming true* is the best he can get. And that is a far cry from *being true*.

Penrose talks of "seeing truth," but since he cannot establish a necessary connection between mathematical perception and truth, the hallmark of consciousness boils down to a weaker concept: grasping meaning. Surely, Penrose will maintain, not *just* meaning. Seeing the meaning of the claim "all even integers are divisible by two" is quite different from seeing the meaning of "please pass the salt." Too true. It is that flash of insight, that feeling of certainty, that grasp of how the whole picture fits together. Penrose's "direct perception of mathematical truth" can thus be reduced to (a) grasping the meaning or semantic content of a statement or idea, and (b) the "Aha!" feeling that accompanies it. The important thing to realize is that the "Aha!" feeling is what we gain from introspection, a notoriously unreliable window into our inner workings and the outer world, and definitely not truth-preserving. Therefore (b) can be regarded as epiphenomenal, with no necessary connection to matters of truth.

Once we are clear on the relation between truth and the "Aha!" feeling, accounting for seeing truth in an algorithmic system becomes no more mysterious than accounting for recognizing that something *seems* correct. Imagine a neural network, Ralph, trained to recognize aspects of mathematical structure. Ralph's internal representation is a partial model of the mathematical world, itself an abstract structure defined by relations among elements (Benacerraf 1965). The model would include a finite representation of primitives and their relations, and rules

for extending that representation. The meaning of an element in this representation corresponds to the role it plays in the context of the representation. Since Penrose, too, must relinquish the notion of direct access to truth, we can demand consistency without worrying about establishing reference. The hard part would seem to be accounting for the "Aha!" feeling, but now that we see it as distinct from truth, it is simple. Suppose Ralph evaluates the degree of matching between propositions and its internal partial model, something neural nets do very well. If the structure of the proposition is isomorphic to the structure of the model (or the model properly extended), or if it exceeds a certain threshold matching criterion, then Ralph can be said to have "grasped" the truth of the proposition. This kind of pattern matching is achieved by networks practically instantaneously, much like our flashes of insight. The degree of fit is the degree of force that the "insight" has. For added realism, we can even endow Ralph with an "Aha!" unit that is activated in those cases of adequate matching. Realize, finally, that truth-grasping neural networks like Ralph, though not considered algorithmic, are routinely implemented on standard digital computers *that are instances of Turing machines and thus are algorithmic according to Penrose's definition.*

The real question, it seems to me, is not how we perceive *truth*, but how we *perceive* at all: how we feel the "Aha!" feeling, or see red, or feel pain. If this is the crux of Penrose's worries, he is not alone. We can't see how an algorithmic system can have subjective states. But we can't see how *any* physical system can have subjective states, including ourselves. Yet we do. It is no more evident how consciousness would arise from collapse of the wave packet than how it would arise from a very complex algorithm. This doesn't put algorithms on any worse footing than it does physicalism, which Penrose and I both wish to uphold.

So what of the emperor? One should always be aware of potentially better social orders, but it is reckless to overthrow the old regime without an indication that the new one will not fail in the same way. The emperor hasn't yet had a chance to prove his mental mettle: We do not know that algorithms cannot give rise to consciousness.

ACKNOWLEDGMENT

Thanks to Paul Kube for hours of invaluable discussion on these matters, and to Patricia Smith Churchland for helpful comments.

The pretender's new clothes

Tim Smithers

Department of Artificial Intelligence, University of Edinburgh, Edinburgh EH1 2QL, Scotland

Electronic mail: *tim@edai.edinburgh.ac.uk*

Professor Penrose's book contains a great deal of material, ranging over a number of "large" subjects. The question is, does all this material succeed in appropriately adorning a new argument against strong AI? The answer is no, for two reasons. First, Penrose seems to misunderstand strong AI. Consequently, his argument that Turing computation is inadequate for consciousness is irrelevant to AI. I admit that pure symbol processing is not enough, but we need a different approach not some (only to be hinted at) new theory of physics. I will expand briefly on each of these points in turn, and then make some concluding remarks.

In this book we have a very odd and very poor description of what AI is. Its objectives are said to be "to imitate by means of machines . . . as much of human mental activity as possible." We are told that "one of the aims of AI is that it provides a route towards some sort of understanding of mental qualities, such as happiness, pain, hunger." (p. 14) AI supporters (Penrose doesn't mention its practitioners) are supposed to envisage that "such concepts as pain or happiness can be appropriately modelled" by, for example, machines whose power packs are running low, or are fully charged. I know of no serious AI researcher, in strong AI or otherwise, who would be at all happy with this kind of description. Artificial intelligence is the science concerned with understanding intelligent behaviour and how it can be created in the artificial. This does involve investigating the cognitive goings on that engender intelligent behaviour, but AI is not concerned with *artificial emotions*.

So where did Penrose get his idea of AI research from? I don't know, but it cannot have been from the literature I am familiar with as an active researcher in both symbol processing and robotic kinds of AI. Penrose's description sounds like the sort of thing we might hear on a rather low quality popular science radio or television programme. Wherever it came from it is clear, from the style and manner of presentation, that Penrose doesn't much like his idea of what strong AI is – a point to which I shall return in my concluding comments.

The picture we get of AI goes roughly like this. AI's aim is to build a computer to pass the Turing test. This requires a program which takes typed questions from the tester's terminal, which then computes convincing textual answers. This is to be done by an algorithm that treats the input text as a string of symbols and, by a suitable sequence of manipulations, produces the required output string. This I will call the algorithmic one-shot description of AI, where an algorithm is a precisely defined sequence of operations (from a finite set of operations) that produces a precisely defined result (not necessarily in finite time). In presenting this picture, Penrose equates the kind of algorithmic symbol manipulations of the Turing machine with the kind of symbol manipulation that strong AI suggests form the stuff of mental states. This basic mistake arises from a failure to appreciate that the kind of symbol manipulation underlying strong AI is not the machine code of the digital computer, but that expressed in Newell & Simon's (1976) Physical Symbol System Hypothesis (PSSH). Penrose makes no reference to this hypothesis, indeed he makes no reference to any of Newell and Simon's work. He therefore appears to be unaware of the ideas underlying symbolic AI developed by two of the most important people in the field. In fact, no references are cited to substantiate the view of strong AI he presents.

This failure to understand the symbol processing paradigm in AI leads Penrose to describe strong AI as seeking an algorithmic (in his strict sense of the term) explanation of mind and consciousness. He fails to realize that although algorithmic procedures may be used to achieve the required manipulations of symbol tokens, this does not necessarily mean that the overall behaviour of an AI program can be described as algorithmic, or that it can be achieved purely algorithmically. Of course, if you look at the right level of the system what you see is just algorithmic behaviour, but this does not provide an explanation of the system as a whole – just as looking at the firing patterns of neurons inside my head doesn't explain my behaviour. AI programs typically work by (amongst other things) building and maintaining internal descriptions of what is going on in the environment; they use these in deciding how to act. To do so they use a mixture of algorithmic and nonalgorithmic processes; nonalgorithmic processes are ones that must continue to run to fulfill their function, as opposed to complete and produce a result, as in the case of algorithmic processes. A good example of such a programme, better described as a machine, is the Hitech chess playing machine from Carnegie Mellon University; see Berliner 1988 for a tournament report. Its behaviour in winning a game of chess is not reasonably described as algorithmic, but many of its internal workings might be.[1]

If Penrose has argued that all this symbol processing cannot, of itself, give rise to mental states, then I would agree with him. But it's not a new theory of physics we need to solve this problem, its a different approach, a change of emphasis.

The PSSH as Newell and Simon presented it talks of the processes of *designation* and *interpretation* as being central to

the concept of how symbolic manipulation can lead to intelligent action. The problem is that much of the symbol processing research that has gone on around it has either ignored these aspects or believed that they could be achieved simply by "bolting on" appropriate sensors and actuators to the computer. Researchers have thought that a robot is just a symbol processing computer with sensor and motor input and output devices – sensory-motor functions are seen as requiring only the right technology, whilst the "intelligent" bit (requiring mental states) is all in the symbol processing. The fact that programming computers became easier much faster than building real robots probably also has a lot to do with it – if real robots were as easy to build as complex symbol processing systems we would have many more of them today, and consequently would probably know much more about the problem of getting them to behave intelligently. This problem involves understanding how their internal workings can be (or can become) semantically coupled with their tasks and how they can be (or can become) semantically embedded in the environments in which these tasks are performed – semantically for the robot, that is, not for us as observers or designers. The trivialization of this problem is also a result of its lack of emphasis in the PSSH. Consequently, the problem of how the symbols are to be grounded, to use Harnad's (1990) term, has received very little serious attention, and in fact is not understood even to exist as a real problem by many *symbolic functionalists* (Harnad 1989).[2]

Understanding how the internal processes of intelligently behaving agents can come to have meaningful content for them does not require some new physics. It requires an approach that puts the question of how systems can be or become mindful of, and so act intelligently in, their environment: how a system can become semantically embedded. Harnad (1990) has called this kind of AI *robotic functionalism.*

In AI there is a much needed debate about how symbol processing can really play a role in engendering the mental states required for intelligent behaviour. What is *not* required is to have some distorted image of AI put up and shot down with threats of some new physics.

ACKNOWLEDGMENT

I am grateful to my colleague Chris Malcolm for his comments on an earlier draft of this commentary.

NOTES

1. This algorithm-using nonalgorithmic kind of behaviour is not peculiar to AI. Computer operating systems and word processors offer other examples.

2. This has not, however, prevented many useful ideas and techniques coming out of symbolic functionalist research. Expert systems, which are artificial extensions (amplifiers) of human intelligence, are a good example.

And then a miracle happens . . .

Keith E. Stanovich

Department of Psychology, Oakland University, Rochester, MI 48309-4401
Electronic mail: *keith_stanovich@um.cc.umich.edu*

A few years ago a cartoon appeared in the *American Scientist* that showed two somewhat satisfied mathematicians looking at a blackboard on which was written a long and complex proof. The complicated formulae began on the far left side of the board and continued until reaching the middle where, written in block letters, were the words AND THEN A MIRACLE HAPPENS. The proof then recommenced, finally reaching its conclusion on the right side of the board. Penrose's book reminds me of this cartoon. For more than three hundred pages he presents a wonderful introduction to mathematical concepts, algorithms, Turing machines, concepts of truth and proof, classical physics, quantum mechanics, and cosmology. And then in the final two chapters we are presented with a theory of "consciousness" that, with very little argument and much question begging, is simply asserted to rest on the quantum mechanical concepts previously introduced.

The last two chapters go awry because Penrose adopts a folk psychological notion of "consciousness" far outside the mainstream of cognitive science: "I am implying that when I refer to thinking, feeling, or understanding, or particularly, to *consciousness,* I take the concepts to mean actual objective 'things' whose presence or absence in physical bodies is something we are trying to ascertain, and not to be merely conveniences of language!" (p. 10) "Since the cerebrum is man's pride . . . then surely it is here that the soul of man resides!" (p. 383). Although this neoCartesian view is quite similar to the layman's conception (Stanovich 1989), it is a view that has been utterly rejected by modern cognitive science.

Despite the poor record of vernacular-based concepts in the history of science (Churchland 1979), Penrose makes no bones about the fact that it most definitely *is* a vernacular concept of consciousness that he intends to utilize in his quantum theory: "We can rely, to good measure, on our subjective impressions and intuitive common sense as to what the term means" (p. 406). Of course, the problem is that our "intuitive common sense" about what consciousness is has changed across historical epochs and cultures because our folk language of the mental has evolved (Wilkes 1988). Is the quantum account of the brain culturally dependent? This never becomes an issue because Penrose remains imprisoned within a folk psychology that simply doesn't recognize that there is a question of this type to be addressed.

That the folk term consciousness fractionates into at least four vastly different usages goes unrecognized in *The Emperor's New Mind,* even though the author himself slips and slides among these usages constantly. For example, Wilkes (1988) discusses usages of "consciousness" as a one-place predicate meaning roughly "awake," consciousness of internal sensations like pain, consciousness of sensory experience, and consciousness of propositional attitudes like beliefs (see Wilkes 1984; 1988). So when one sees such questions as, "Where is the seat of consciousness?" (p. 381) one is tempted to ask whether all of these consciousnesses are localized in the same place. Are they all to be subsumed under the same "quantum" explanation?

In short, the folk language of the mental is simply not the place to begin when analyzing any concept of brain function. Connectionist models (Churchland & Churchland 1990; McClelland & Rumelhart 1986; Seidenberg & McClelland 1989; Sejnowski & Rosenberg 1988), modular brain theories involving semi-autonomous processors (Allport 1980; Dennett 1978b; Hofstadter 1985; Minsky 1987), and dissociation phenomena increasingly uncovered in neuropsychology (P. M. Churchland 1988; Springer & Deutsch 1985; Tranel & Damasio 1985) and cognitive psychology (Kihlstrom 1987; Nisbett & Ross 1980; Nisbett & Wilson 1977; Rollman & Nachmias 1972) – all are putting tremendous stress on the integrity of our folk language for mental processes (P. S. Churchland 1983; 1986; Dennett 1987; 1988; Rorty 1979; Stich 1983).

As Wilkes (1988) argues:

> "If 'consciousness' is as central and unavoidable as many seem to suggest, it is then at least prima facie interesting that other languages, and English before the seventeenth century, appear to lack the term, or anything that corresponds more than roughly with it; in other words, that what strikes some of us so forcefully, as being so 'obvious,' seem to have left little impression on others" (p. 170) "We are thus thrown back on the unsurprising thought that 'conscious' and 'consciousness' are terms of the vernacular. . . . 'consciousness' does not pick out a natural kind, does not refer to the sort of thing that has a 'nature' appropriate for scientific study" (pp. 192–193).

Wilkes's argument that the vernacular term "consciousness" does not carve nature correctly is consistent with the way that scientific developments in cognitive science have occurred. As Pylyshyn (1980) has noted: "Information-processing theories

have achieved some success in accounting for aspects of problem solving, language processing, perception, and so on, by deliberately glossing over the conscious-unconscious distinction" (p. 443). Anyone familiar with developments in cognitive science is aware that 10 years later the situation remains as Pylyshyn described it. A strong indicator of just how out of step is Penrose's emphasis on vernacular consciousness is the fact that Posner's (1989) 800-page compendium, *Foundations of cognitive science*, contains no entry for "consciousness" in the index!

At the very end of the book Penrose very nearly admits that all the quantum hand waving was a cover for his "gut feeling" that our vernacular folk vocabulary about consciousness is "right" and will not be altered or eliminated by computational explanations ("Yet beneath all this technicality is the feeling that it is indeed 'obvious' that the *conscious* mind cannot work like a computer," p. 448; "I simply cannot believe that it [consciousness] is something just 'accidentally' conjured up by a complicated computation." p. 447). Beyond simply asserting the importance of taking consciousness seriously, and asserting its nonalgorithmic nature, the final chapter is heavy with Penrose's introspections about his thoughts while doing mathematics ("Part of the reason comes from my experiences as a mathematician" p. 413). This all gets somewhat embarrassing, for the discussion rambles on, in pop-psych fashion, in blissful ignorance of decades of work in experimental psychology indicating how tenuous are such reports as mechanisms of theory justification.

> Consider an analogy. I am an expert reader. Let me tell you how I do it. Because I am a highly practiced expert, meaning is absorbed into my brain with just the most fleeting contact with the actual print on the page. As my eyes flow smoothly across the page, the press of meaning and the involvement in the text make it seem hardly a visual task at all. More technically, what is happening is that my brain is exploiting the redundancy of the message, making it unnecessary for me to process the fine visual details of the print. This is what it means to be an expert reader like myself.

Now, as all cognitive psychologists will know, this little introspective story I have just told is an utter fable. It is wrong from start to finish. The brain does not process print in the way described (see Rayner & Pollatsek 1989). The eyes, of course, do not move smoothly across the text, despite our introspections. Information is acquired during discrete visual fixations during which the eye is largely still, and the interposed saccadic movements are discrete and ballistic. Moreover, even fluent readers process extremely fine visual details in the print, again despite our introspections to the contrary (Rayner & Pollatsek 1989). Introspective fictions like my "reading story" and like Penrose's "what it's like to do mathematics" tale cannot be taken as evidence for even mundane claims – like the pickup of visual information in reading – let alone for highly speculative claims regarding the nonalgorithmic nature of conscious thought.

Given the pretensions of the book, Penrose – despite the remarkable breadth evidence in other parts of the volume – must be faulted for virtually ignoring the actual empirical work in the field of cognitive psychology and, beyond some passing citations of Churchland and Dennett, for giving modern philosophy of mind short shrift. The familiar prejudices of the physical sciences are present here. To understand physics and mathematics, it is assumed that some time must be devoted to building some foundational concepts. In contrast, it is assumed that the literature generated by the science of human behavior can be safely ignored – because, after all, anyone can do it. We all behave and of course we all "have minds" don't we? I have tried to expose the fallacies in these attitudes elsewhere (Stanovich 1986) so I will devote no more time to them here.

Ignorance of actual empirical work in the behavioral sciences allows Penrose to commit the howler of making his concept of intelligence dependent upon consciousness: " I do not think that I would believe that true intelligence could be actually present unless accompanied by consciousness" (p. 407). This, of course, ignores a century of work on the behavioral concept of intelligence during which none of the dozens of major theorists have seen fit to ground the concept of intelligence in consciousness in this way (see Sternberg & Detterman 1986) – for the sound reason that the folk concept of consciousness is considerably confused (Armstrong & Malcolm 1984; Dennett 1969; Lyons 1986; Rorty 1979; Ryle 1949; Smith & Jones 1986; Wilkes 1984). Intelligence has proven a tricky enough behavioral concept without linking it to a folk term of highly dubious scientific value.

Finally, because of the book's premise, I should say a word about where quantum mechanics come in. In the author's own words:

> Since there *are* neurons in the human body that can be triggered by single quantum events, is it not reasonable to ask whether cells of this kind might be found somewhere in the main part of the human brain? As far as I am aware, there is no evidence for this. . . . One might speculate, however, that somewhere deep in the brain, cells are to be found of single quantum sensitivity (p. 400).

In other words, AND THEN A MIRACLE HAPPENS. Of course, this is only one step in the argument. How can there be implications for "deeply" understanding "consciousness" even if there are such cells? Well, you guessed it, *another* miracle happens ("if something like this could be developed into a fully coherent theory then there might emerge a way of providing a quantum description of the brain" p. 403).

In summary, this book contains a host of chapters that are exceptionally good introductions to modern physical theory and computation. The concluding chapter on the physics of the mind is, however, hopelessly muddled.

The thinker dreams of being an emperor[1]

M. M. Taylor

Defence and Civil Institute of Environmental Medicine, Box 2000, North York, Ontario, Canada M3M 3B9

Electronic mail: *mmt@zorac.deciem.dnd.ca*

> *"I am inclined to think –," said I. " I should do so," remarked Sherlock Holmes, impatiently.*
>
> (Sir Arthur Conan Doyle, *The Valley of Fear*, 1914)

Roger Penrose is inclined to think of algorithms, but thinks he does not use algorithms to think. For Penrose to believe he does not think in algorithms, he dreams of a new nondeterministic physics, *correct quantum gravity* (*CQG*), which will allow his thinking processes to be nonalgorithmic. In justifying this idea, he embarks on a fascinating journey to both ends of the universe, which is well worth the price of the book; but does he need to go so far to study what lies behind his nose?

It is extraordinarily unlikely that we evolved as thinking machines suited to prove mathematical theorems. Almost certainly our thinking developed to permit us to behave reasonably well in rapidly changing circumstances. We do not need to know whether it is true that stripes of colour belong to a tiger, so long as we evade the possible tiger. A real-life travelling salesman does not need to traverse a minimum distance Hamiltonian path, so long as he spends only a reasonable amount on travel fares. The operative words are "good enough," not "optimum." The mathematics of truth, proof, and infinity are rare among forms of thinking, and arguments relating to computability or provability or the halting of algorithms seem to be irrelevant to normal thought, on which mathematical thought is presumably founded.

Penrose argues that thinking must be nonalgorithmic because it can be proved that in any formal system there are theorems

that we can *see* to be true, but that we cannot prove algorithmically to be true. The only way out that he can see is to devise CQG.

Penrose uses the term *algorithm* in two different senses. In the first sense, which is carefully defined and illustrated by example, an algorithm is a precisely defined method for converting data of a given class into a result of a defined type. It is teleological, being driven by the goal of finding a result of the specified type. It is also, by implication, executed in a thermodynamically isolated system: The workings of the algorithm are hidden between the intake of the data and the emission of the result, and partial results cannot be considered reliable. Some algorithms, given some data, may never finish, and even worse, it may be impossible to decide for a particular case whether the algorithm will finish, until it actually does so.

In the second sense, *algorithmic* seems to be identified with *deterministic*, where there is no commitment to avoid interrogating or interfering with the processing of the algorithm. In this second sense, algorithms could conceivably be involved in thinking, whereas algorithms in the first sense could not. Thinking always proceeds in an interactive environment, and no train of thought can be guaranteed to run to completion unaffected by outside influences. Furthermore, the thoughts engendered by an input do not stop when the relevant output has been made. In the second sense, then, algorithm means only the manipulation of the elements of thought according to rules, in an environment in which the data, the results, and the manipulations can be influenced by unexpected events. Algorithms in the second sense cannot be deterministic in the real world. Only the behaviour of the universe as a whole could be deterministic, not that of any nonisolated subpart, since even if the subpart arrives twice in exactly the same state, external events may cause its future behaviour to follow two different paths. Living organisms are particularly responsive to external events, because not only is their very structure maintained by a high nonequilibrium energy flow, but also their "aliveness" is signalled by their effective and timely behavioural responses to external events. Living things cannot be functionally deterministic.

Living systems are *dynamic systems*, and, as such, their internal and external behaviour (including any physical correlates of "thought") can be described in terms of orbits in a phase space of very high dimensions. If left alone, an orbit in a dynamic system settles in the neighbourhood of an attractor, of which most such systems have many, each with its own basin of attraction. Typically, the high-dimensional phase space can be separated into many loosely coupled subspaces, into all of which the orbit may be projected. Loose coupling means that the projection of the state into one subspace does not much affect the state as projected into another subspace, so that for the most part one can treat each subspace as a dynamical system with its own set of attractors. Similarities in the dynamic behaviour of the subsystems can cause them to affect one another more readily than if their behaviours are dissimilar. Such effects may be called *resonances*.

If the dynamic system in question is the brain, it is natural to identify attractors within the low-dimensional subsystems (of which there are many) as potential thoughts, and the current state and orbit as representing actual thoughts. Resonances correspond to analogies, or to perceptions of events in the external world, and most particularly, if a resonance shifts an orbit in a subspace from one attractor basin to another, we may call the result "insight," especially if the final effect is to enhance the overall resonant coupling within the greater dynamic structure. In the absence of external input, we may call the ongoing shifts of resonance "dreaming," and from this viewpoint we might suggest that no biological or silicon system can think unless it first learns to dream.

NOTE

1. This commentary is DCIEM Technical Note 90-N-06.

Exactly which emperor is Penrose talking about?

John K. Tsotsos[1]

Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4

Electronic mail: *tsotsos@ai.toronto.edu*

I began reading this book with high expectations; that is only natural given the reputation of the author. Unfortunately, my expectations were not fulfilled. Penrose has written a book that has a solid, interesting, and illuminating middle; he should have left it at that. The beginning and end were quite disappointing, but for very different reasons.

The book begins very weakly, with the old, standard, boring critique of artificial intelligence. Penrose's bibliography contains exactly four AI works: a 1968 paper by Marvin Minsky, a 1977 book by Roger Schank and Robert Abelson, David Waltz's 1982 survey article in *Scientific American* and a 1972 paper by Terry Winograd. Suppose someone were to criticize physics and base their argument on three out-of-date references and one survey paper intended for general audiences? I am certain Penrose would not consider this a scholarly critique. So, too, with Penrose's criticisms of AI. The discussions by other philosophers on whether or not the brain can be computationally modeled are also based on a pitifully small number of now out-of-date works. Philosophers invent definitions and terms, which are not generally accepted by the practitioners of AI, to suit their purposes, leaving their own concepts woefully undefined. Penrose himself says that "it is unwise to define consciousness" (p. 406), yet he happily defines the positions taken by AI researchers and uses relentless (and, I think, inappropriate) precision in criticizing computation. His auxiliary title, "Concerning Computers, Minds and the Laws of Physics," seems much more appropriate than his main title.

In any case, the conclusions reached in this book about the nature of the mind really are disappointing. Even if all AI work is ultimately wrong with respect to brain function, it forms a body of falsifiable research. Why bother with the philosophical arguments if it is so easy to falsify or confirm the theories put forward? Questions about strong versus weak AI are irrelevant. The only important property of a scientific theory is that it provides the simplest explanation for the experimental observations and makes predictions that can be tested. If a theory is inadequate, experimentation will falsify it and a new theory will take its place. We can take Schank & Abelson's (1977) work, for example, and put it to the test. Proper experimentation will settle the issue once and for all.

There is no question that physics is important in our understanding of all aspects of our universe; this does not mean, however, that physics alone can explain everything. It was truly a pleasure to follow Penrose on his tour through the world of physics, fondly remembering my undergraduate courses in each of the major topics covered. Penrose draws the link between computation and physics especially well, correctly going to the heart of his problem: If one wishes to prove that the brain is nonalgorithmic, one must find some aspect of the physical world for which a computational explanation cannot suffice. Although the wonders of Mozart's brain seem to involve nonalgorithmic actions, a great deal of the brain's action seems algorithmic, such as perception. To emphasize the ill-defined yet magical qualities of a brain we label a "genius" over the equally unexplained abilities of the more automatic systems of the brain is to miss the point of understanding intelligence. Consciousness disembodied from perception has no input; unconnected to action it has no output. And it is not at all obvious that perception and action have no connection whatsoever to consciousness, as the differences in single-cell recording experiments carried out on alert and anesthetized animals have demonstrated.

Penrose appeals to the not yet discovered Correct Quantum

Gravity theory and to neurons deep in the brain sensitive to single quantum events to find the root of conscious thought. In particular, he appeals to gravitons; I thought that the existence of gravitons had not yet been confirmed. Let us suppose that this is a valid scientific hypothesis; how can one devise an experiment to test it? How can an alert and normally functioning brain be isolated so that the only input to any of its neurons is a single graviton passing through the cortex? This presupposes that one can detect a graviton in the first place. Even if this were possible, what part of the cortex does it affect and how? A synapse perhaps? A Purkinje cell? How will its effect be detected? Will billions of probes be required so that activity at each synapse of each cell is recorded? Would not those probes, which must form an electromagnetic circuit themselves, interfere with the graviton itself as it passes through? Or would we be required to use some form of noninvasive method that can both detect that a single graviton has entered the cortex and tell us what its effect is through the cortex? If its effect is to induce nonalgorithmic events, how can those events be anticipated and characterized so that they may be detected? Whatever detection method is used, it seems that it must be able to resolve individual synapses at least, and probably individual electrical and chemical activities, in order to discover the effect of the graviton. I cannot see how Penrose's proposal constitutes a falsifiable hypothesis.

Although Penrose seems to misunderstand computation in general, taking a very "binary" view of it, I agree that there may be nonalgorithmic components to conscious thought. Algorithms are defined as mechanistic ways of evaluating functions, implemented with some computing agent. Their outcome depends only on their inputs. More precisely, an algorithm, when applied to a particular input set, results in a finite sequence of actions; each action in the sequence has a unique successor, and the algorithm either terminates with the solution to the problem or with a statement that the problem is unsolvable. Nondeterminism within an algorithm leads naturally out of this definition and one could speculate that perhaps there is a form of nondeterminism in the brain. For example, some neural networks[2] as well as more traditional computer science techniques such as queuing theory, depend on random variables: They use stochastic algorithms and thus their output does not depend solely on their inputs. Moreover, the neural networks, among other methods, depend on the kind of energy-minimization that Penrose claims is important for their success. This seems a much more straightforward way out of Penrose's dilemma; but it need have nothing to do with quantum physics, of course!

There is one additional point I wish to make. Penrose discusses questions of consciousness without questions of "realizability." Is the kind of consciousness, and indeed intelligence, that Penrose envisions actually realizable in a brain? In my own work (Tsotsos 1990), I tried to show that this issue is a very serious one, and that considerations of computational complexity seem to rule out many "in principle" correct solutions exactly because they are not implementable within the resources offered by the brain. We know they are correct "in principle" Penrose would say that we "see" the correctness of the solution – but the brain is both too small and operates too slowly to be able to implement those solutions as they stand. We can thus capture the essence of Penrose's argument in a mathematically well-founded manner without appeal to undiscovered entities. The solutions that are realizable, and that seem to agree well with experimental observation, are exactly those that yield approximate answers (of specific character). Could Penrose be confusing this with aspects of consciousness and judgement?

Although philosophical discussions of the computational nature of the mind are interesting, they detract from the business at hand: finding good scientific theories of intelligence. Arguments similar in spirit to Penrose's have been made throughout mankind's history about the inexplicability of matter, of light, of the motions of planets and stars, of gravity, of disease, of reproduction, and other topics we now take more or less for granted even if they are still not fully understood (Churchland 1990): Fortunately, some past researchers were not convinced that these phenomena were inexplicable.

NOTE

1. Author is also with Canadian Institute for Advanced Research in Toronto.

2. One of the most glaring typographical errors I have seen in some time is at the top of p. 398, where "neutral networks" are defined!

Between Turing and quantum mechanics there is body to be found

Francisco J. Varela

Centre de Recherche Epistemologie Appliqué, Ecole Polytechnique, 75005 Paris, France

Electronic mail: *bitnet.fv@frunip62.bitnet*

This book is a mixture of a lot of things: the interesting, the original, the naive, and the infuriating. In that order, the interesting and the original compose, luckily, a good deal of the book. I am referring to Penrose's clear and relentless review of the notions of algorithmicity and recursivity. In this he succeeds admirably well. Most interesting to me is the way he weaves together the issues of algorithmicity in the mathematical and traditional AI setting with such similar issues as natural processes in modern physical theories. I found myself engrossed in this reading. I wish Penrose had written just that book, for which I have nothing but praise.

To be sure, a discussion of the nature and limitations of algorithms breathes heavily on the neck of the proponents of so-called "strong AI" whom Penrose sets up as his opponents from the very beginning. He isn't satisfied with Searle's (1980) conclusion (from his Chinese Room argument) that machines are different from brains because the latter have "intentionality" and "history" (p. 23). He wants something better, something like a "successful theory of consciousness – successful in the sense that it is a coherent and appropriate physical theory" (p. 10). He wants to show that there is "an essential nonalgorithmic ingredient to (conscious) thought processes" (p. 404).

And it is in this dimension of the book (clearly the one that electrifies Penrose the most) where things go awry, toward the naive and the irritating. Penrose, without as much as a blink, jumps to the conclusion that the only way out of no-clothes-strong-AI is to invoke physics! Let me call this the first basic conceptual premise or leap of the book: Since cognition/consciousness is nonalgorithmic (contra strong AI), therefore we arrive at the "fundamental question: What kind of new *physical* action is likely to be involved when we consciously think or perceive?" (p. 371, his emphasis). And as if one somersault were not enough, he goes into yet another extreme leap: This required physical action must be something like the link between brain and quantum processes (which are, as he has explained, nonalgorithmic in some complex and fascinating ways)! This is the second basic premise/leap of the book: "I am speculating that the action of conscious thinking is very much tied up with the resolving out of alternatives that were previously in [quantum] linear superposition" (p. 438). Because physics must be involved in explaining nonalgorithmicity (cf. the previous leap), it must therefore have to do with the brain tapping directly into quantum mechanisms. For this commentary let me call these Penrose's Leaps I and II, and examine them in reverse order.

About Leap II, there is little to say; it strikes me as an illustration of the physicist's hubris, even though Penrose himself warns us that "even they [i.e., the physicists] don't known everything"! (p. 23). Prima facie there is no reason to discard a hypothesis that links brain operations to quantum processes.

But, as Penrose readily admits, given the macroscopic nature of the physiological events in the brain, putting forth such a hypothesis seriously would demand at least a modicum of explicit work, some shred of solid evidence. Instead, we get little more than a remark on the quantum catch of photoreceptors (in a mere page and a half, p. 400–01), and some vague references to nonregular tiling as a process of synaptic modification (in another page and a half, pp. 437–39). This is, even for an optimist, mere wild speculation, which is, in any case the adjective Penrose uses to evaluate Moravec's (1989) book on robotics, because "his [Moravec's] speculations are quite extraordinary and go enormously beyond anything that can be scientifically justified" (Penrose 1990, p. 5).

Now for what I have called Penrose's Leap I: Since brains are nonalgorithmic they must have some physical principles, otherwise, "Perhaps we are doomed to be computers after all!" (p. 402). What is plainly wrong with this contention is that, between the haven of the strong AI and the marvel of quantum mechanisms, there is something that has been forgotten, namely, a body! Sundry critics of strong AI have for years said in one form or another that what is wrong with this view is that it lacks something tangible and straightforward: that a view of disembodied cognition violates the fact that all knowledge is embodied and situated. Every analysis of cognition is always embedded in a context, in a given milieu, and can only have a limited competence in given situations. To pretend otherwise (as in the strong-AI position) is to perpetuate the traditional western dualism between mind and matter in a sophisticated format, no more, no less. (One notes without surprise that Penrose describes himself as a Platonist).

Now, the mere fact of being embedded and embodied makes it impossible for such cognitive processes to have the universality of the kind claimed by a Universal Turing machine. Stated in other terms, it seems to me that a recent trend in cognitive science and AI is to abandon the notion that issues of general computability (i.e., problem solving, inference, etc.) are the hallmark of intelligence, and move toward considering the apparently trivial tasks of, say, situated visual segmentation and motor balance as the deepest intelligence. In this sense, an insect provides the toughest nut to crack, not a chess player. One of the clearest instances in which this trend is visible (and not by accident) is in the field of robotics, where more and more practitioners consider it unnecessary to maintain the classical notions of central representation of a world and of action planning. Instead, this trend views viable moment-to-moment updated actions as the main measure of intelligence, thus bringing contingency and improvisation from the periphery to the center of consideration. An intelligent action means continually redeciding what to do, which means not predefining objects independently of the agent location but enacting them in terms of the role they play in the agent's ongoing project (see e.g., Agree 1988; Brooks 1986b; Moravec 1988).

The importance of embeddedness, via a rich sensory-motor coupling, enacted in specific circumstances (and not in a general abstract world) is a far cry from the disembodied Turing ideal. This is, I submit, the "nonalgorithmic" component of thought that Penrose was looking for. In fact this less glamorous but more realistic view of cognition has been discussed all through the history of cognitive science, mostly by a heterodox minority during the two decades of strong-AI dominance. For recent presentation see from the philosophical standpoint Dreyfus (1979), from the AI side Winograd & Flores (1984), from the neurobiological side Gibson (1979) and Maturana & Varela (1987). This very selective reading of research in cognitive science and AI is clearly seen in Penrose's recent critique of Moravec's book, where he tells us that there are (only) three possible views on the question: (1) strong AI, (2) a Searlean-modified strong AI position, and (3) Penrose's, that is, one that asserts that some aspect of cognition is nonalgorithmic (p. 3). Now this is wrong on two counts. First, because the nonalgorithmicity critique via the embeddedness of cognition is an old and articulated view that Penrose ignores although he should not, because it is precisely one of his main tenets. Second, his position (in the book under review at least) is not just that nonalgorithmic components are important, but *also* that we need to find a way out via physical principles. This is an entirely different matter because he thinks that one will find a way out of strong AI via physical principles. This is, to say the least, questionable, and the burden of the proof is on him to show this is an alternative at all.

I applaud Penrose for identifying with Adam the child in his opening short story, who, instead of falling for the AI dogma that the super Ultronic computer about to be inaugurated has a real mind after all, asks how does the machine "feel." Now what is wrong with Ultronic (or rather with President Pollo and his team of designers) is that Ultronic has no sensorium or motorium (other than a trivial exchange through a computer screen), it therefore does not couple and behave in an environment, and hence cannot constitute a shared world with us through common history. Therefore any imputation of feeling is ipso facto nonsense, for emotion can only come from the moment to moment situatedness linked to constant assessment, evaluation, and the ever-present breakdowns leading to constitution of new actions. This ever-present background from which cognitive actions are constituted before being solved is where the heart of the process of cognition lies, and it can only show up in an active, embodied entity.

Nothing prevents such an embodiment from being manmade, as far as I can see. Adam should be put into more pertinent science fiction plots with real opponents, for example, the roving R2D2 (*Star Wars*) or even old Hal (*2001*), whose "body" extended into the entire spaceship. Adam would not hesitate to impute some feeling to the tooting androids (the young hero Luke didn't hesitate either). Children like Adam have to be allowed to grow by considering alternatives that have at least a chance of leading cognitive science somewhere. So let us continue to work weaving together brain operations, phenomenological analysis of everyday experience, and embodied artificial machines, and who knows what such future beings will reveal to us.

Penrose's grand unified mystery

David Waltz[1] and James Pustejovsky

Computer Science Department, Brandeis University, Waltham, MA 02254
Electronic mail: [a]*waltz@godot.think.com*; [b]*jamesp@cs.brandeis.edu*

The organization of this book is like a sandwich in which the filling (eight chapters on computation, mathematics, physics, and brain science) is much more satisfying than the bread (the first and last chapters on AI). Penrose gallops through an entire undergraduate curriculum's worth of material at dizzying speed, covering topics from Gödel's theorem and fractals to quantum phenomena and neural function. The treatments of these range from very good to inspired. Although we could take issue with many of Penrose's interpretations and conclusions (generally found at the ends of the chapters) that attempt to relate these topics to AI, the weakest parts of the book by far are those that attempt to deal with AI directly, and that give the book its title. The AI that Penrose attacks in this book represents only a held-over margin of the current field. Although current AI – including situated automata, perceptually based cybernetics, the society of mind, connectionism, learning theory, and memory- and case-based reasoning – is not necessarily immune to attack, it gets away without a scratch here; Penrose seems unaware of its existence.

Penrose's central – and most distressing – argument goes roughly as follows: There are mysteries in each of the sciences

described in the book. For example, many quantum phenomena are mysterious; we can know mathematical truths only by methods that seem to lie outside any formal system, and cannot be computed by any algorithm. Furthermore, the nature of consciousness is notoriously mysterious. Perhaps then what we have are not many mysteries, but one "Grand Unified Mystery" (GUM): That is, the mechanisms of consciousness may lie outside classical explanations, depending crucially on mysterious quantum phenomena, which in turn somehow belong to a higher order of mechanism that could, for example, know things to be true which no algorithm could ever find or prove. This sounds suspiciously like a modern form of vitalism. Penrose also suggests that such mysterious phenomena may be responsible for giving us privileged access to eternal Platonic truths, which may in this mysterious realm have a real existence; for example, "The Mandelbrot set is not an invention of the human mind: It was a discovery. Like Mount Everest, the Mandelbrot set is just *there!*" (P. 95).

To be fair, some deep and novel issues are raised: The notion of an algorithm is indeed problematically weak; quantum phenomena might indeed be important in neural functioning; and so on. What we dispute is the suggestion that all such phenomena may at the root be the same phenomenon, and that these phenomena have a privileged link with consciousness and some purported Platonic truths. Unfortunately, Penrose's thesis does little or nothing to bring us closer to a real understanding of consciousness or cognition. In fact, his programme seems inclined to hand-wave while at the same time encouraging the reader to accept a view of science as religion, revealed only to the blessed (e.g., Turing, Gödel, Mandelbrot, and Penrose). Nevertheless, on the whole, the science portions make for good reading as long as they stick to science.

In the rest of this review, we would like to primarily concentrate on Penrose's attack on AI and his views on the nature of intelligence.

Penrose unfortunately takes on only the symbolist tradition in AI, which is already well on the way to intellectual death as an adequate explanation for mind. Although much of his same criticism could (and should!) be leveled against connectionism (as Searle, 1990, suggests), he seems totally unaware of this work – not even the PDP books are referenced, which is almost unthinkable for a *BBS* target article on AI! He is likewise unaware of recent thinking on "situated automata" (cf. Brooks 1986*a*; Minksy 1986); models of emergent computation, (cf. Cariani 1989; Rosen 1978; 1987), and other models based on reactions and adaptations to an environment, e.g., perceptual and memory-based reasoning, (Stanfill & Waltz 1986; Waltz 1988). These approaches form for us the most satisfactory bases for models of mind currently available.

Penrose takes on only the purest symbolic paradigm of AI. In this tradition, thinking is viewed as the disembodied formal manipulation of symbols and the Turing test is taken to be a valid way of assessing intelligence. Sensory and motor systems are viewed as peripheral modules whose respective purposes are to convert sensory signals into symbol structures and to convert symbol structures into actions (cf. Fodor 1983 and Pylyshyn 1984). The hardware used for symbol manipulation is of no great concern – any universal computing device will do. Along with Penrose, we find this brand of AI unsatisfactory, but for different reasons. Ironically, for instance, like strong symbolic AIers, Penrose accepts a dualist position that the mind is somehow disembodied from the organism it is controlling. Given this separation of mind from body, it is not surprising that intentionality and consciousness would appear to be mysterious concepts.

In our view, the mind/brain must be seen as including the perceptual faculties as a proper part of its functioning (a strongly Aristotelian position), and not as peripheral to it. Perception may involve computational processes, but it is nonalgorithmic (in Penrose's sense), since it does not involve purely symbolic processing. The perceptual faculties do not have to compute every computable function but only those of classification and discrimination involving similarity, difference, and so forth, that are important to the organism. The computations involving the measurements of observables are what in fact grounds the organism and its symbolic system in the world (cf. Cariani 1989 and Harnad 1990).

It is this explicit link between perception and understanding that gives rise to the "aboutness" or intentionality for that organism. We believe that, in principle, a device could be constructed with such machine intelligence if it is properly situated in its environment and allowed to develop its own semantics for the world that it constructs through its own sensors and effectors. Alas, the world perceived by such a synthetic organism would provide us with as few Platonic truths as those "discovered" by our own privileged species.

More specifically, we imagine that most of the "mind" of a situated automaton consists of perception, action, and memory; situated automata react to events in the world with actions that have led to "desirable" outcomes and/or avoided "undesirable" outcomes in the past. Perception is an active process, involving layers and "societies" of detectors and agents, gated according to the particular situation, the set of goals, and the past experience of the organism. Assuming that a given level of precision is sufficient, brute force computation, communication, I/O, and memory may eventually suffice to reach that precision, though it is possible to ask for precision so great that a computer with a number of elements equal to the number of electrons in the universe would not suffice. This degree of precision is very unlikely to be required, however, as we seem designed for making inferences from scraps of evidence, and out of wetware with a high degree of redundancy and fault tolerance: We can still think and act effectively even though many neurons die every day; we can recover from strokes or accidentally caused lesions; and substantial changes in the operating environment of the brain (e.g., fever) do not disable thought. Speed is an issue, however; proper situated automata must be able to respond to situations as rapidly as the situations change. If not, this would be like a weather forecasting system that took a week to figure out the weather one day ahead. It seems clear that we are designed to be good at very rapidly picking interpretations and actions that (usually) "satisfice" (cf. Simon 1981) in particular situations, not for proving truths. Statistics are certainly as relevant as logic!

Thus, symbols are part of the story, but only a small part: Certain classes of inputs, seen often enough, may be treated as effectively identical, and become candidates for symbolic enshrinement (Hillis 1988). Nonetheless, formal symbols and sequential actions that operate on them are only a small part of the overall story of mind, one that is always accompanied by nonsymbolic activity that is important for explaining the behaviour of the automaton.

What does this buy us? It allows us to imagine a system that responds to situations not by running an appropriate algorithm, but by classifying the situation (and its subparts) and then responding to the situation both according to its memories of what actions were appropriate in the past for such situations, and with continuous compliant tracking of moment to moment changes. Furthermore, unlike the Turing machine model, the system is continuously and invariably changed in the process of interacting with the world. It literally cannot run the same algorithm twice – the second time it tried to do so, it would remember doing the same thing before, and so on (though there may be such other, more significant, changes as skill learning or insights that cause much more extensive changes in the automaton). To be sure, there are plenty of mysteries here, but no need for mysticism.

ACKNOWLEDGMENT

This work was supported by DARPA Contract No. 49620-88-C0058.

NOTE

1. David Waltz is also associated with Thinking Machines Corp., 245 First St., Waltham, MA.

Computability, consciousness, and algorithms

Robert Wilensky

Division of Computer Science, University of California, Berkeley, Berkeley, CA 94720

Electronic mail: *wilensky@larch.berkeley.edu*

Is the mind algorithmic along the lines of a digital computer? That it might not be is suggested by Penrose's introspection about his own mathematical reasoning, by an argument that humans may not be limited by Gödel's theorem in the way computers are, and by philosophic arguments that consciousness, and so forth, cannot be achieved by formal systems. The missing piece, Penrose suggests, may have something to do with quantum mechanics. Unfortunately, Penrose's arguments are mostly restatements of old confusions about AI, and illuminate little of scientific or philosophic interest.

Penrose starts out on the wrong foot by paying serious attention to Searle's (1980) "Chinese Room" argument, which, like the ghoul in the horror film, having been convincingly and utterly vanquished last time around, inexplicably appears in the sequel as if the previous episode had never occurred. Suffice it to say here that Searle's argument has nothing to do with computers or AI; it is a warmed-over version of the "mind-body problem" and the "problem of other minds." My favorite short rebuttal is to note that, following Searle's line of reasoning, if we should be leery of believing that computers could really have minds because mind is "a biological phenomenon and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation and photosynthesis or any other biological phenomena," then we should be at least as reluctant to believe that God, if He exits, has a mind because He lacks "the specific biochemistry"; most likely, He would simply be faking the correct behavior.

While not taken in entirely, Penrose is led astray. I can mention only a few points. Contrary to Searle's and Penrose's claim, believing that something exhibiting the correct behavior actually has consciousness, or intentionality, is not dualistic in the slightest. These features would simply be emergent properties of such systems. Indeed, this is exactly Searle's claim for the relation of such properties to the brain. Again, the only controversy is about which systems such properties will emerge from, and no light is shed on this issue by Searle or Penrose. In addition, it is simply false that if one believed that computers really understand then one would have to attribute full-blown intentional states like hopes and beliefs to all sorts of nonneural devices (e.g., thermostats) any more than one would have to attribute hopes and beliefs to all simple neural ones (e.g., bees or earthworms).

More substantively, Penrose's presentation of the limits of formal systems goes off track when the implications for human beings are considered. The first misconception is that machines are somehow limited by these results in ways that humans are not. The second and more important misconception is that such limits play an important role in the difficulties that have been experienced in creating artificial intelligence. Both views are uncontroversially wrong.

First, contrary to Penrose's claim, it is perfectly possible for computers to "see" that results not provable within a particular formal theory are nonetheless true. Here Penrose confuses the *mechanism* doing the proving with the *theory* in which the proof is stated. The fact that a proof doesn't exist within a given theory doesn't mean that a given mechanism must be constrained to operate only within that theory, as Penrose falsely presumes. Indeed, much of what Penrose seems to believe is not doable by machines has already been done. An interesting case in point is Shankar's 1986 Ph.D. thesis entitled, "Proof-checking Metamathematics," in which the Boyer-Moore theorem-prover (i.e., a mechanism used to prove assertions within a theory) is used to prove a number of metatheorems (i.e., theorems about that system), including Gödel's theorem.

More important, Gödel's result and the like are singularly uninteresting from the standpoint of the AI enterprise. The reason relates to Penrose's claim that his own mathematical processes aren't algorithmic. Here, Penrose seems to confuse "algorithmic" in the formal sense of something that always computes what you want in finite time with "algorithmic" in the sense of being specifiable as a computer procedure. But these are entirely different creatures. Indeed, the party line among AI researchers is that most interesting mental capacities are best described as heuristic procedures. These are nonalgorithmic in the sense that they don't inexorably grind out the right answer all the time. Thus virtually all AI programs aren't algorithmic in the first sense, as Penrose incorrectly suggests. Although we don't know whether such programs will do, they appear to be perfectly compatible with Penrose's description of the subjective impression of his mathematical processes (a precarious line of reasoning in any case).

Proving theorems appears to have little in common with interesting human cognitive capabilities, such as using language, or being able to see, or being able to walk down a hallway. That is, what is impressive about people is not the rare capabilities exhibited by mathematicians, but the mundane capabilities of any five-year-old child. It is precisely such capabilities with which AI is primarily concerned, and, in these basic areas we have no proof whatsoever that the capabilities to be emulated are beyond the limits of rather banal models of computation.

This is not very strange. We have proofs about the limits of computation only for esoteric enterprises like proving theorems in number theory. But most human beings, who have no aptitude for proving theorems at all, are perfectly capable of talking and seeing and walking. If humans aren't hindered from having cognitive capacities by an almost complete lack of theorem proving capabilities, why should a remote limit on it be a hindrance to computers?

Therefore, there appears to be no reason to believe that artificial intelligence will be difficult to achieve because of computational deficiencies in our ordinary models of computation. Does Penrose have a useful quantum-mechanical insight nevertheless? This is perhaps the most

disappointing aspect of the book. Virtually no evidence is offered that quantum mechanics plays a role in the functioning of the brain. Indeed, one is hard pressed to imagine what has motivated Penrose to alert us to the possibility that something as ubiquitous as quantum-mechanical effects would somehow account for what is unique about the brain.

My guess is that Penrose is on a wild goose chase for a missing piece that does not exist. Here is why: If qualities like consciousness are truly separable from behavior, then there is by definition no objective test to determine whether a creature with acceptable behavior actually possesses them. Indeed, whether you and I have these elusive qualities, as opposed to just the capability to exhibit the right behavior, has no empirical ramifications whatsoever. In a universe without such qualities, but in which behavior is identical to that in our world, a Searle without authentic intentional states continues to insist that he has them, and that computers do not.

I suggest that the way out of this peculiar situation is simply to remain skeptical that pretheoretical notions like consciousness will emerge unscathed from our future scientific endeavors. Even the most cherished commonsense notions about time, space, and energy have undergone drastic revision in the course of science. It would be astonishing if concepts like consciousness did not. In any case, the road to progress will probably require a lot more hard work and a lot less idle speculation.

Minds beyond brains and algorithms

Jan M. Zytkow[1]

Department of Computer Science, George Mason University, Fairfax, VA 22030

Electronic mail: *zytkow@gmuvax.gmu.edu*

Mind and computer. Is our thinking basically the same as the action of a computer, perhaps a very complicated one? Penrose strongly disagrees because two things are missing from any computational picture of mind. One thing is consciousness; another is the mind's capability to do better than any algorithm. I agree with him on both issues, but I find his demonstration of the nonalgorithmic nature of mind unconvincing and vulnerable to the attacks of the proponents of strong AI.

Penrose repeatedly uses the distinction between recursive and nonrecursive problems to argue that although computers are bounded by the former, humans can successfully deal with the latter. Take a nonrecursive but recursively enumerable word transformation problem (p. 130–31). Procedures exist that can find a solution for each word problem that has a solution. But if there is no solution, the procedure will continue forever. Penrose believes that humans are not so limited. They use their intelligence and ask metalevel questions about solvability. Rather than endlessly trying to solve the problem at hand they can prove unsolvability. For a given word problem this can be done by finding a property that is present in the input, is conserved by each transformation, but is absent in the output. Because no algorithm would find such a property in all cases (the problem is nonrecursive), the computers are left behind by human intellects.

The argument works for simple algorithms, but it does not cross the defense line of strong AI, because a more sophisticated software can mimic the ways Penrose thinks human intellect can outwit an algorithm. Consider a program that monitors its own performance. After spending some time on unsuccessful search in one problem space, the program can switch to another problem-solving approach or to a proof that no solution exists. There are even programs that can change themselves. None of these alters the nonrecursive nature of the word problem, but there are programs that can prove nonsolvability for subclasses of all word problems. When Penrose claims that we must use our intelligence in choosing a particular property, the proponents of strong AI will be more than delighted to repeat the same about their heuristic search programs. But what if no program works for all cases? Well, on what ground can anybody ensure that a human is able to do better? Can anybody demonstrate that humans can prove the unsolvability of every unsolvable word problem?

I do not think that there is a well defined, formally specified problem that can be solved by humans but not by computers. Penrose would lose in the game, "Show me an example of nonalgorithmic thinking and I will implement it on the computer," with an AI system builder worth his LISP machine. Still, I believe that no computer can fully reproduce our mental processes. We can model a sum of well defined human capabilities, each limited to a particular context, but not an unbounded variety of all capabilities and all contexts in which our mind can work. The existence of algorithms for many different problem classes is not equivalent to a total simulation of human thinking.

Mind and physics. According to Penrose, the conscious mind can eventually be explained as a physical device. Although I agree with him about the causal influence of consciousness on the physical world, I do not think that his idea of "the strange way that CQG [correct quantum gravity] must act, in its resolution of the conflict between the two quantum-mechanical processes U and R" (p. 446) can provide a solution. It is not just that his explanation is extremely implausible, relying on a possibly nonexistent phenomenon on the distance scale of 10^{-33} cm: I do not see how the collapse of the wave packet in very special circumstances created by quantum gravity can possibly explain the intrinsic contents of our conscious states, but instead of arguing against this particular explanation I will argue that no scientifically described phenomenon can explain the intrinsic contents of consciousness.

Suppose that science generates a very accurate explanation of the brain, that consciousness is identified with a particular class of material structures and that conscious states can be measured by physical parameters. Suppose also that conscious activity has been explained as a physical interaction, perhaps a tremendously complex one. All that science can measure about ourselves would become explainable. Under these assumptions, the scientific contents and the influence of our conscious states would be accounted for, but none of the consciously perceived intrinsic contents. All the contents of our consciousness would be scientifically redundant. Indeed, we can imagine a creature like us, one that would behave exactly the same way for an external observer, according to all physical parameters he could measure, and would satisfy the same laws, but unlike ourselves it would not have any intrinsic conscious states.

Penrose does not believe in the redundancy of consciousness and tries to find an evolutionary advantage that would justify its existence. But whatever advantage can be expressed in scientific terms, it is also an evolutionary advantage for the unconscious physical lookalike from my example. Because we cannot find a scientifically documented advantage that would not occur in our unconscious counterpart, we cannot claim that there is an evolutionary advantage specific to consciousness. Like Penrose, I do not believe in the redundancy of consciousness, but unlike his argument, mine is supposed to be a *reductio ad absurdum* of the assumption about scientific explanation of our mind.

My argument against the scientific explanation of consciousness can be also applied against algorithmic explanation. Here Penrose cannot believe that consciousness is "just 'accidentally' conjured up by a complicated computation"; so why does he seem to believe that it is "accidentally" conjured up in a complicated physical structure? I agree with his diagnosis (p. 23)

that "Searle, and a great many other people, have been led astray by the computer people. And they, in turn, have been led astray by the physicists," but it seems that the latter category of misled people includes the author of this extraordinary book. For somebody who argues so strongly for the importance and causal activity of consciousness, there is just one small step to admit that not everything can be scientifically explained.

Why can we not accept that mind is an anomaly in the worldview propagated by scientific and computational universalism? We can still be scientists and constructors of AI systems. I find nothing more fascinating than to chase and imitate that elusive being – our mind. I am sure, however, that we will not create an artificial mind, although we will construct very complex and enormously useful computational systems.

ACKNOWLEDGMENT

Many thanks to Malcolm Perry and Marek Bielecki for their helpful suggestions.

NOTE

1. Author is on leave from Wichita State University.

Author's Response

The nonalgorithmic mind

Roger Penrose

University of Oxford, Mathematical Institute, 24-29 St. Giles, Oxford, England OX1 3LB

I think it is fair to say that the least complimentary of these thirty-seven commentaries are also the ones containing the most serious misunderstandings of what I was trying to say in *The Emperor's New Mind* (henceforth *Emperor*) They add up to at least one very sobering – and evidently valid – censure: The book was not nearly clear enough!

It is perhaps not surprising that the most negative remarks come from the AI community. Let me begin by saying that I intended no criticism whatsoever of what they are doing or of their *striving* to simulate intelligence or other activities of human or animal nervous systems. I had hoped this would be clear from a detailed reading my book. Although the title might be interpreted as suggesting that the proponents of AI are (like the fraudulent weavers in Hans Anderson's "The Emperor's New Clothes") trying to "put one over," on the general public, certainly no such insinuation was intended by me. (**Glymour & Kelly** say that my title is "almost the only rude thing about the book"!) I was aware, when I chose my title, that such an interpretation was possible, but I had not expected that some AI proponents would be sensitive enough on this point to take such an interpretation seriously. I hope that they will accept both my dissent from such an interpretation and also my unreserved apology, if they believe that some might be misled into thinking that I viewed AI proponents as being in any way dishonest in their intentions – which I emphatically do not.

Although some criticisms came from what I shall call the "soft" side (people who believe that mental phenomena cannot be accounted for in terms of physical explanations: possibly **Niall, Roeper, Taylor, Varela,** and **Zytkow**), it was clear that most would come from the "hard" side (those with at least some admitted sympathy for the position of "strong AI": **Boyle, Breuel, Dennett, Doyle, Eagleson, Glymour & Kelly, Johnson, Ettinger & Hubbard, MacLennan, Madsen, McDermott, Perlis, Smithers, Tsotsos, Waltz & Pustejovsky, Wilensky**). I in fact believe that there are strong arguments on both sides, as I tried to explain in my book. Those on the "soft" side are right in their belief that there is no evidence whatever to suggest that "feelings," or "qualia," can be described in physical terms at all, let alone be evoked by the mere carrying out of a calculation; while those on the "hard" side are equally right to point out how intimately our feelings and perceptions are associated with the states of particular physical objects, namely brains, while all physical objects are, as far as we know, like patterns of "information" (?) governed in full detail by marvellously precise mathematical laws. The very strengths of both arguments had seemed to present us with a paradox. The way out of this paradox, in my opinion, must lie in our obtaining much deeper insights into what physical laws actually *are* and in unearthing those aspects of physics that must evade computational descriptions and begin to allow room for experiential phenomena. It is towards such goals that I have attempted, with my book, to guide our future thinking. Let me now make some comments addressing the arguments on both sides.

Why physics?

Those on the "soft" side might wonder why I even bothered to write the book at all, since it is surely "obvious" that subjective experiences, or "qualia," are not things that can ever be reduced to physics (i.e., to "mass, length, and time," in **Niall**'s words; see also **Zytkow**'s account). So what light can possibly be shed on this question by bringing in physics at all, let alone CQG ("correct quantum gravity")? To this I must reply, "precious little, if anything, so far" as they might expect. I am certainly not claiming to have a theory of the mind – and I am rather surprised that some of my critics (e.g., **Boyle**) seem to be chastising me for not presenting one! It would be presumptuous in the extreme for me to claim to have solved a fundamental puzzle that had remained unsolved for some three thousand years despite its having been the subject of deliberations of some of the deepest thinkers since antiquity. Introspection can certainly yield significant information but introspection has not really got us very far with regard to the basic question of what perception *is*, or of why we perceive at all.

It seems to me that we shall never be able to make progress toward understanding these issues until we have a much better picture of what physical reality is actually all about (and with reference to **Glymour & Kelly**'s remarks, I do not see how we can do without physical realism, whatever view one might hold about the relation between the mind and computation). It is in science, and above all physics (profoundly supported by mathematics), where enormous strides in our understanding of this world – the universe in which we (and our minds) actually

appear to live – have been achieved. It is in physics that we have at last found SUPERB theories, and although these theories are still very incomplete and do not yet shed any significant light on what minds actually are, it is through the back doors of physics and mathematics, in my opinion, that our assault on the problem of mind must ultimately be mounted. The SUPERB theories which have emerged in this century, namely, relativity and quantum theory, have actually moved us some good way from the rather sterile-seeming "mass, length, and time" characterization of classical physics to something much more subtle. I have no doubt whatever that there are more profoundly subtle theories (still to be found) underlying both relativity and quantum theory, and that the more we understand about the nature of our world, the less sterile the physics of this world will seem – and eventually the actual role that consciousness and perceptions play in our universe will begin to be discerned. Despite what some of my critics (**Stanovich, Waltz & Pustejovsky**) seem to be implying, I certainly do not believe that CQG (correct quantum gravity), when it finally comes to light, will in itself solve the problem(s) of mind. But I have tried to present arguments in my book in support of my view that some (nonalgorithmic) physical theory of this kind (very probably CQG) is a necessary prerequisite if a scientific theory of mind is ever to be achieved.

Simply finding a (necessarily algorithmic) "physical grounding for information processing" (**Boyle**) in terms of existing physics will not be enough. Nor do we know, despite what **MacLennan** seems to imply, that physical behaviour is necessarily algorithmic just because described by differential equations. (**Breuel** even seems to take the opposite view, see later!) A good part of my purpose was to stress that computability in physics is not at all the same thing as determinism (**Eagleson** and **Taylor** please note).

Mathematical insight and Gödel's theorem

Let us now turn to the question of why I think that an *algorithmic* theory like strong AI can never provide a theory of mind. A large part of my argument is based on Gödel's theorem, but a good many of those who criticized me from the "hard" side (**Dennett, Doyle, Glymour & Kelly, MacLennan, McDermott, Perlis, Tsotsos, Wilensky**), and also several others (**Boolos, Butterfield, Chalmers, Davis, Hodgkin & Houston, Kentridge, Manaster-Ramer, Savitch & Zadrozny, Mortensen, Roskies**), have objected, and even claim to have located errors in my argument ("quite fallacious," Chalmers; "wrong," Mortensen; "lethal flaw" and "inexplicable mistake," Manaster-Ramer, Savitch & Zadrozny; "old and fallacious argument," Perlis; "deeply flawed," Davis; "invalid," Kentridge) – while only a few seem partially to support my type of argument (perhaps **Gilden & Lappin, Higginbotham, Lutz, Niall**) on this issue. As I shall explain in a moment, there is actually nothing wrong with my argument – and I make no apology for the fact that much of it is "old," certainly appreciated by Gödel himself already in the 1930s and never properly refuted since (cf. Gödel 1986; 1990; and as cited in Rucker 1982, p. 171; also Nagel & Newman 1958) – except that I evidently did not present my case forcefully enough, nor did I anticipate the most likely misunderstandings. All my adverse critics on this topic have jumped to conclusions and, in one way or another, have missed the point of what I am trying to say. None seem to have grasped the full import of the Gödelian argument. The fault is mine: I should have explained things more clearly.

Most of these critics have pointed out that the Gödel proposition $(P_k(k)$ (which I shall subsequently denote simply by G(F) here), constructed from some formal system F, cannot be "seen" to be true unless the system is already known to be consistent. (Actually the particular Gödel-type proposition I construct does not depend on the consistency of F, as my critics maintain, but, as the argument is given, on what is called its "ω-consistency." This is stronger than actual consistency, asserting that if $\sim\forall x[Q(x)]$ is a theorem in F, then it must not be the case that all the propositions Q(0), Q(1), Q(2), Q(3). . . . are theorems. This depends upon the symbol " ~ " and, more particularly, "∀" actually meaning what they are supposed to mean: "not" and "for all" respectively! The distinction between consistency and ω-consistency is not important for us here – and in fact one can get away with assuming much less for F – so let it pass for the moment.) My point is not that, in any particular case, G(F) is true or "seen" to be true (or false or "seen" to be false). My point is that the *deduction* of G(F) from F is (and is "seen" to be) a perfectly valid procedure. Thus, if we are already accepting F as a sound system, then we must accept G(F) as being true also. (The word "sound" is used by logicians for systems whose theorems are actually true in the intended interpretation of the symbols.) If we are prepared to use the procedures of F as sound methods of inference then we must accept the passing of F to G(F) as another sound method of inference. This is not unlike any other form of reasoning in mathematics. One does not directly perceive a "new" mathematical truth in isolation. What one does is to see that the new proposition follows, in some way, from the body of mathematical understanding that has already been established. It is in this sense that one "sees" the truth of G(F), just as one does in constructing any mathematical argument for obtaining new truths from old. But the Gödel method lies outside the procedures that have been actually laid down within F itself. I am not concerned with how we may (or may not) already "know" that the procedures of F are sound – that is part of what is assumed – but with the remarkable fact that if we are accepting F as sound, then we must also accept G(F) as true.

Some of my critics (**Boolos, Chalmers, Davis, Perlis**) have drawn attention to some particular formal system (e.g., the standard Zermelo-Fraenkel system ZF – or a modification thereof, such as ZFC or ZFM – cf. Boolos, Davis), claiming that we do not know whether or not it is consistent. Thus we cannot be sure that its Gödel proposition is actually true. (Inconsistency is just one particular type of manifestation of a system F not being sound. There are many others, even apart from the ω-inconsistency actually relevant here. The system might be consistent, or even ω-consistent, whilst some of its implications might be false; but my critics seem to have paid attention only to the possibility of inconsistency.) If F is inconsistent, then any proposition that can be described within the system is derivable using the rules of F,

irrespective of whether it is true or false. Thus, in particular, G(F) is itself a theorem! In fact, the Gödel proposition for any system F is always either true as a statement in mathematics or it is a theorem of F. Of course, the latter are the situations in which F is useless as a means of deriving mathematical truth. (If F is inconsistent, then, in particular, 2 = 3 is a theorem also.) But I simply do not understand how mathematicians of a formalist disposition[1] (Boolos, Davis) can claim that mathematical argument within some particular formal system F are more acceptable than the mathematical argument that derives G(F) from F.

In his criticism, **Boolos** asserts that the systems generally used in mathematics (e.g., ZF) are good approximations to mathematics. (Approximations to what?! Boolos is revealing himself to be a good Platonist after all, since he evidently believes that there is something there, other than the formal systems themselves, that the formal systems are supposed to be approximations to.) But if ZF turns out to be inconsistent, then it is no approximation to mathematics at all since 2 = 3 would then be a theorem according to the rules of ZF (**Mortensen** take note!) – though the *rules* of ZF might approximate the rules of some consistent formal system which does approximate actual mathematics.

To reiterate my central point: I am not asserting that, in any particular case of a formal system F, we need necessarily be able to "see" that G(F) is true; but I am asserting that the "Gödelian insight" that enables one to pass from F to G(F) is just as good as a mathematical procedure for deriving new truths from old as are any other procedures in mathematics. This insight is not contained within the rules of F itself, however. Thus, F does not encapsulate all the insights available to mathematicians. Such a limitation applies to any formal system whatsoever. I nowhere claim that the mathematicians' insights would enable them, in principle, to resolve *any* mathematical question (**Dennett, Doyle, Roskies**). I merely claim that the insights which *are* available to mathematicians are not formalizable.

MacLennan (and also **Wilensky**) refers to a distinction between mathematics and metamathematics, the deduction of G(F) from F being, he asserts, metamathematics, not mathematics. But this distinction only really exists if one adheres to a purely formalist view of mathematics. To a Platonist (or a "common-sense mathematician"), metamathematics – which is reasoning *about* formal systems (rather than within a formal system) – is still mathematics, on a par with any other form of mathematical reasoning.

I should make a remark about formal systems, such as ZF, ZFC, or ZFM, which refer to reasoning about infinite sets. These sets are allowed to be rather enormous; one might indeed have reasonable doubts about their existence or about the system's consistency (at least in the case of ZFM). One might even encounter differences of opinion amongst mathematicians as to whether or not to accept reasoning which involves such infinite sets. But one can avoid much of the argument about such possibly controversial issues by concentrating on just the small area of mathematics consisting of propositions in number theory. In fact, it is sufficient merely to consider propositions of the special form

[Q]: "Q(n) is true for all natural numbers n"

where Q(n) is some given computable (recursive) true/false arithmetical property of the natural number n (i.e., there is an algorithm for deciding, for any particular n, whether Q(n) is actually true or false – so Q(n) is an always terminating Turing machine action).[2] The point about this is that the Gödel proposition G(F), as presented in *Emperor*, is explicitly of this particular form, so we do not have to consider anything set-theoretically more obscure than statements of this kind.

How do mathematicians decide the truth or falsity of such propositions? They might decide to use the rules of some formal system F that (like ZF) refers to mathematical structures (say, enormously infinite sets) that go far outside the concepts of ordinary arithmetic, and in doing so, might introduce some doubt (e.g., see **Davis**) as to their consistency, let alone their soundness. Nevertheless, for any given system F, there will be propositions of the form [Q] that can be "proved" using F and others that cannot. If F is inconsistent, then the system F would be, strictly speaking, useless. The decision as to whether or not to trust a formal system F as being sound is normally the job of those logicians concerned with the details of that system. But if I choose, myself, to use the rules of F in some calculational way – say, to decide the truth of some proposition of the form [Q] – then, to know whether I am to trust a result derived from these rules, I, also, should persuade myself that F is *sound*. This is a much stronger requirement (i.e., actually *using* the rules of F) than I would need if I were to use my Gödelian insight to deduce G(F), since for that I need only know that F is ω-consistent. All mathematical knowledge requires mathematical insight and understanding. This applies to the judgements I might apply in trying to decide whether or not F is sound; it also applies to the Gödelian insight that enables me to deduce G(F) from F. The Gödelian insight is just one kind of mathematical insight amongst many. It is emphatically not something given only to a "chosen few" such as myself (as **Glymour & Kelly** seem to claim with some irony). Indeed, it is an insight simple enough even for me to give a reasonably adequate description of it in my (semipopular) book! It is something noncontroversial and, in principle, accessible to all of us, unlike (apparently) the insights lying behind the validity of certain particular systems like ZF or, more especially, ZFM. The important thing to realize about the Gödelian insight is that it derives from an (at least partial) understanding of the *meanings* of the axioms and rules of procedure of F, and also of the actual meaning of G(F). (Most particularly, it derives from an understanding that the symbol "∀" actually means "for all natural numbers," since this the essence of ω-consistency. Gödel's theorem in the form that I have presented it illustrates the fact that the actual meaning of "∀" cannot be fully encapsulated by formal rules. **Mortensen** is puzzled how a "finite" mathematician can comprehend such things as hyperreal numbers. But the puzzle is more simply based than this, since it applies already to the natural numbers.) One cannot regard F as providing the rules merely of some "game" for manipulating meaningless symbols (as the strict formalist might lead one to believe). Thus Gödel's theorem has another profound implication (**Garnham** take note; also, I think **Roskies** grossly underestimates this point, whereas **Smithers** does not even seem to recognize it!): *Semantics cannot be formalized.*

Are mathematical insights algorithmic?

It is important to realize that "formalizable" and "algorithmic" are virtually synonymous terms. (The "proving of theorems" – say of statements of the form [Q] – within some formal system F, is certainly something algorithmic whereas, conversely, any algorithm that generates "theorems" can be adjoined appropriately to the axiom system of predicate calculus: a standard set of rules of logical inference). Thus we can directly apply the Gödelian insight to any algorithm that purports to be generating (say [Q]-type) mathematical truths. (Note: apart from considerations of complexity theory, a truth-generating algorithm is formally equivalent to a truth-proving algorithm – despite what **McDermott** says!) If the algorithm in fact generates only mathematical truths then we can obtain a new mathematical truth by using the Gödelian insight – a truth that is not obtainable by the given algorithm (together with predicate calculus). The original algorithm is thus limited in what it can achieve: It is unable to incorporate the particular insight

"F sound" $\Rightarrow$ "G(F) true"

that Gödel has revealed to us. Moreover, as soon as we have seen that G(F) is true we realize that we have not finished with the insight, because "G(F) true" can now be adjoined to "F sound" (etc. etc.).

Some of my critics (**Davis, Doyle, Kentridge, MacLennan**) have objected, claiming that the Gödelization procedure could itself be automated. But this is not so if by "procedure" we mean the above "insight." If the Gödelization procedure as a whole *could* be automated, then we could apply this procedure to the very system that the automation would yield, and hence obtain a new Gödel proposition that lies outside the scope of that automation. In detail, the way that this can come about is quite subtle, and it is perhaps not surprising that people often feel that they could automate Gödelization. The Gödelian insight is a very slippery character, who can flee in a different guise as soon we feel that we have got him algorithmically cornered.

Let us see how this can come about. Suppose that we have some formal system F. We wish to adjoin to F another rule which might seem to encapsulate the Gödelian insight: "If {A, B . . . D} is any finite[3] set of theorems, then the Gödel proposition obtained from them as axioms (and the rules of procedure of F) is to be also a theorem." This gives a new system F* that is much broader than that obtained by simply adjoining G(F) to the axioms of F, since it allows Gödelization to be applied again and again, as many times as we might wish, always within this one system. If F is sound, then so also will be F*. But F* still has not really incorporated the Gödelian insight because the proposition G(F*) will now lie outside the scope of F*, and G(F*) will be true if F is sound (and will be seen to be true so long as F is seen to be sound). Of course we might now perceive how to automate the procedure F$\Rightarrow$G(F*), rather than just F$\Rightarrow$G(F), but again, that very perception would allow a new Gödelization procedure that lies outside the scope of that new automation.

The slipperiness of the nonalgorithmic nature of the Gödelian insight lies partly in the fact that we are never quite sure what it is that has slipped through the algorithmic net, though something always does! One might argue that what has slipped through it is the very perception of how one actually formulates something in an algorithmic way, in the knowledge that the suggested algorithmic procedure actually yields a truth (or a set of truths). In particular, the algorithmic procedures that **Davis** refers to whereby the consistency of a set of axioms can be coded into the nonsolubility in integers of some polynomial equation P = 0 involve the knowledge and insights (very ingenious ones, at that, for which Davis himself was partly responsible!) that the insolubility of that equation actually does encode the consistency statement it is supposed to encode. There would also be numerous *unsound* algorithmic procedures of this type, but we need (nonalgorithmic) insights to tell us which procedures are in fact sound!

It seems to me that this is the same kind of slipperiness of the roles that conscious understanding and insight play in any kind of mental activity – especially as they relate to the procedures whereby one might try to program a robot to behave in certain humanlike ways. As soon as we have seen how to turn the implications of a certain insight into a set of rules, we can make a robot act according to those rules. There seems to be nothing barring our being able to automate anything we can precisely formulate by such rules – and when it is automated, the robot can carry out the implications of that particular automation far faster and more accurately than we can ourselves. Yet the very insight itself has not been automated. Nor has the insight that we have used a particular insight to program the robot. Insights in general cannot themselves be automated – as is illustrated by the fact that the full Gödelian insight, in particular, can never be automated. The understanding of what that insight actually is doing will allow us to do better, in some ways, than the robot that we have constructed by using it.

Learning to perform some human action that initially requires some conscious understanding appears to be a very similar phenomenon. Once learnt, that action can be relegated to unconscious (perhaps cerebellar rather than cerebral) control. This unconscious control seems much more like that of a programmed robot – and as with an electronic computer's action, it can be carried out much faster and more accurately than can action under conscious control – while unconscious control does not appear to embody the actual understanding the initial consciousness was needed for. What is the actual role that consciousness plays? It seems to be a subtle – and even a slippery role – like that of the Gödelian insight. This slipperiness is in the nature of all nonalgorithmic understanding.

I should emphasize that there is nothing "paradoxical" in the Gödelian insight, unlike (as **MacLennan** seems to be suggesting) the case of the "liar" paradox. People often seem to think that the self-referential aspect of Gödel's theorem is its essential content. This is extremely misleading. The relation of Gödel's theorem to the liar paradox is largely a historical and motivational one. Most nonalgorithmic classes of problem have no evidently self-referential aspect at all. Of course human language is imprecise, and it does allow inconsistent selfreferential statements like, "This statement is a lie," to be made. (Compare **Wilensky**'s self-destructing assertion that my deductions from Gödel's theorem and their relevance to

AI are "uncontroversially wrong"!) This imprecision is just as well, for without it, we should be constrained to converse within some specific formal system, and discussion of matters such as the Gödelian insight itself (and other reflection principles) – let alone semantics in general – would become impossible.

Kentridge refers to my version of the Gödelization procedure (as given in *Emperor*) according to which one can, by the exercise of understanding, "outdo" any given algorithm H for testing the stopping of Turing machine actions.[4] He says that with this argument (as with the Gödel argument), all I have shown is that humans can do better than one particular algorithm, not better than all algorithms. Not so! To say this is to miss the point of what one is achieving with a *reductio ad absurdum*, of the kind I have been presenting. (**Chalmers** and **Wilensky** also take note – and despite **Zytkow**'s friendly warning, I do not lose the "outdoing an algorithm" game!) My argument is of the form: "Given any particular algorithm, that algorithm cannot be *the* procedure whereby human mathematicians ascertain mathematical truth. Hence humans are not using algorithms at all to ascertain truth." Recall, for comparison, Euclid's argument that there is no greatest prime number. This is also by reductio ad absurdum: "Given any number n that purports to be the greatest prime, n! + 1 has a prime factor larger than n, so n was not the largest prime after all. Hence there is no greatest prime." As with the algorithms above, **Kentridge** would appear to be claiming that all Euclid has shown is that for any particular number n, one can find a larger prime. But this certainly *does* suffice to show that there is no largest prime! The argument with the algorithms is similar. Given any algorithm that purports to be the one that humans knowingly use to ascertain mathematical truth, one can find a more powerful algorithm that we (humans) can see will also ascertain mathematical truth. This likewise suffices to show that humans do not ascertain mathematical truth by means of any knowable algorithm.

A more serious issue is that humans might be using an algorithm whose validity is in fact unknowable to them, and this is indeed the very point I was trying to address in the final chapter ("disappointing" to **Tsotsos**) of *Emperor*. Only a very few of my critics appear even to have noticed that my argument is not the same as that of Lucas (1961). Rather than addressing, with Lucas, the possibility that some particular individual might be acting according to an algorithm, I refer instead to a putative algorithm that the mathematical community *as a whole* is supposed to be using. It is rather more believable that any particular individual might be using some horrendously complicated unknowable algorithm than that the mathematical community as a whole is using one. The existence of an algorithm X, according to which mathematicians as a whole are supposed to act, would place an absolute limit on what human mathematicians are able to achieve. The algorithm (or formal system) X, or at least the soundness (or merely ω-consistency) of X, would have to be something unknowable to human mathematicians. Otherwise G(X) would be seen to be a mathematical truth, although it is inaccessible to X: a contradiction.

I never claimed that the existence of X is a mathematical impossibility, but it would seem to be exceedingly unlikely. Appealing to a "horrendously complicated unknowable algorithm" is totally at odds with the actual way mathematics is done. As I said in *Emperor*, mathematics is "built up from simple and obvious ingredients" although sometimes, as when one uses a Gödel type insight for a reflection principle[5] these ingredients are sometimes quite subtle – though still "obvious"! – and their very "obviousness" may require some genuine appreciation of the underlying meaning of the symbolism used. (It should have been clear that I am not claiming, as **Davis** and perhaps **Chalmers** seem to be suggesting, that mathematical arguments are themselves always "obvious." That would be very far from the truth, since they are sometimes exceedingly complicated. It is just that such arguments are, at least in principle, built up from such "obvious" ingredients. Moreover, I don't even need any outrageously subtle "obvious" ingredients. The Gödelian insight alone will do (**Chalmers** and **Higginbotham** take note.)

Of course, it might be the case (and I suppose that the following is the sort of thing that the AI people have in mind, cf. **Hodgkin & Houston,** also **Butterfield**) that mathematicians have the *illusion* that all the time they are appealing to such "simple and obvious ingredients" whereas in fact they are actually using the horrendous X, where X has somehow arisen by natural selection. This seems to be an implication of the sort of picture being put forward by **Dennett** in particular, although I am not altogether sure that I have understood his point completely. Somehow, X is supposed to have arisen because our remote ancestors were clever at designing mammoth traps and the like, and now some of their descendants have found that X is also good for discerning obscure mathematical truths! I find such a picture quite implausible. It is creatures who can *understand*, and who can comprehend *meaning*, be it mathematical or be it conceptually valuable for the construction of mammoth traps, that natural selection has favoured. Understanding and meaning are nonalgorithmic (as the Gödelian insights demonstrate) and they are mental attributes that clearly seem (to me, at least) to be particular manifestations of conscious contemplation.

Perlis makes the very curious suggestion that Fermat might have worked according to the system ZFC (as his "X") and that that was how he discovered mathematical induction. But that would imply that the great Fermat was intellectually incapable of understanding the axioms of ZFC – and certainly incapable of believing them!

I cannot agree with **Butterfield** that consciousness is just "baggage" from the point of view of natural selection, like the weight of the polar bear's coat. I argue that consciousness is a necessary ingredient of understanding and insight, so I believe that it must, in itself, have been a very positive factor with regard to natural selection. I should emphasize that I strongly believe in natural selection, although some (e.g., **Hodgkin & Houston**) seem to have taken a certain passage in *Emperor* (p. 416), to the effect that natural selection appears to work better than it "ought" to, as implying that I have significant doubts. I was referring only to the remarkable way in which the actual physical laws, with their propensity towards forming complex and highly organized structures (like quasicrystals or Frank-Casper phases of crystals) seem to fit in with natural selection extraordinarily well, and making the whole process work better than it ever could if the

laws themselves were not so appropriate. **Breuel** (as well as **Hodgkin & Houston**) refer to "fault tolerant" algorithms that are more robust than ordinary Turing machine specifications. I agree that such things would be preferable from the point of view of natural selection, but I still do not see how the horrendous X could possibly be selected for in this kind of way.

Dennett refers to the algorithms that chessplaying computers use, these being "extremely good" rather than "correct" algorithms for playing chess. I think that he (and, in effect, **Taylor** also) is suggesting that mathematicians are likewise using such an "extremely good" rather than "perfect" algorithm to decide mathematical truth. But I really do not see how Dennett gets around the above argument. His proposal does not at all explain the types of insight mathematicians are using all the time, which are *correct* but nonalgorithmic rather than *approximate* but algorithmic. Of course, as **Boolos, Chalmers, Davis, Garnham, Hodgkin & Houston, Lutz, Perlis,** and **Roskies** point out, mathematicians sometimes do make mistakes (although Cantor was *not* shown by Russell to be wrong), but the mathematical community as a whole makes extraordinarily few; and this would be totally inexplicable from the "naturally selected X-algorithm point of view" given the extraordinary abstractions away from actual experience involved in most of pure mathematics. Any approximate algorithm for generating all the mathematical truths we know would have to be incredibly complicated (and in no way related to the kind of thing creatures need to survive), whereas mathematical insights are, at root exceedingly simple.

It seems that Turing himself thought that somehow it was the fact that humans make mistakes that got around the seeming paradox of an algorithmic device (as was his view of the brain) being able to appreciate the truth of the nonalgorithmic Gödelian insight (see Hodges 1983). To me his proposed way out seems exceedingly unlikely, for essentially the reasons outlined above; and it is hard to see that introducing inaccuracy is what makes the brain work better! I have said that the imprecision of human language allows it greater scope than otherwise, but language only works at all because of the underlying power of the mind that allows precise sense to be extracted from imprecise language (e.g., from an informal description of the Gödel procedure: **Garnham** take note). Gödel (cf. 1990, p. 297) found himself forced into a different extreme, namely, that the "mind" was only loosely to be associated with the brain, and was not limited by the brain's finiteness. The fact that these two great thinkers found themselves driven to take such improbable-sounding positions is an indication of the genuine seriousness of the problem which must be faced. Neither Turing nor Gödel seems to have considered the possibility of nonalgorithmic physical action operative in the brain. If this also sounds improbable to some on the face of it, they should at least appreciate why something remarkable is actually needed!

I should make it clear that in taking the argument outward from single mathematicians to the mathematical community as a whole, I am not allowing that a nonalgorithmic entity (the "mathematical community") might somehow arise from a collection of algorithmic ones (as some indeed suggest individual mathematicians might be). On the other hand, at one stage down from this, **Chalmers** (and perhaps **Doyle** and **Perlis**) seem to be suggesting that humans might manifest nonalgorithmic behaviour on a large scale while being constructed from entirely algorithmic (microscopic) ingredients. In this view, it would be possible in principle to simulate the behaviour of the ingredients using a complicated-enough computer program and make it look as though the simulation was behaving nonalgorithmically on the large scale. This is basically a version of the same suggestion as Dennett's above, where a "horrendous X" algorithm is supposed to underlie all human insights and understanding. My (main) arguments against this are just as before. Moreover, despite what Perlis seems to be claiming, it seems exceedingly unlikely (though perhaps not totally impossible) that anything usefully nonalgorithmic can ever arise out of algorithmic basic ingredients in this way. For example, "chaos" (referred to by **Mortensen,** though I am not altogether clear about what his point is, and also by **Waltz & Pustejovsky**) does not seem to be "usefully nonalgorithmic," as I discuss at length in *Emperor*, but perhaps this is a question that is worthy of some more serious study.

Are there any limitations, in principle, to what human mathematicians can achieve? I would not know the answer to this question, but (despite what **Dennett, Doyle,** and **Roskies** appear to be implying) I do *not* assume that there are no such limits – although it does seem to me to be at least possible that in principle there are no such limits. There are obviously limits in practice, since, for example, no human will ever be able to multiply together each pair of numbers below $10^{10^{1,000,000}}$. (Even a computer could not do this, but it would, of course, be much better at this sort of thing than a human!) Limits of this kind are not what is at issue, for the human mathematician knows how to perform the multiplications in principle, and this is all that is required. To see that human thinking is nonalgorithmic, we need only refer to what can be achieved in principle. Most mathematical thinking is of the "in principle" kind in any case.

What do I mean by "algorithm"?

I am puzzled that various of my critics seem to be uncertain as to what I (or others) actually mean by the term "algorithm." **Eagleson, MacLennan, Smithers, Taylor, Tsotsos,** and **Wilensky** seem to be implying that I use the term in more than one sense, or in a sense in which I do not mean it. By an algorithm I always mean something equivalent to the action of a Turing machine. I had thought that this was standard terminology, and it is certainly what I adhere to in my book. Any modern general-purpose computer is, in effect, a (universal) Turing machine. There is of course the fact that, strictly, a Turing machine's tape is infinite – or, at least potentially unlimited – but this (acceptable) idealization does not seem to be what is at issue. Any algorithmic action can, with this proviso, be carried out by a modern general-purpose computer; conversely, the action of any computer is indeed algorithmic.

This provides a simple test of whether some suggested model or robot control system is actually algorithmic: If it can be simulated on a general purpose computer, then it is indeed algorithmic. This should make it clear that the

concept of "algorithm" includes not just the action of ordinary serial computers, but parallel ones as well; and that what are referred to as "neural networks" or "connection machines" are also algorithmic, as are the operations referred to as "heuristics" and "learning" in computer systems. (Thus **Johnson, Ettinger & Hubbard**'s ARTI would certainly be algorithmic *if* he existed – despite his charmingly human display of misunderstandings of my book!) Neural networks, for example, may be constructed as special hardware, but as often as not their activity is simply simulated on a general purpose computer. I am aware that people often seem to use the phrase "algorithmic computer" in a more restrictive way than I have been doing, namely, only when the computer is programmed in a strictly "top down" way, proceeding according to some specific mathematical algorithm whose action is clearly laid down to solve a specific problem in a well-understood way. Neural networks are supposed to "learn" and modify their structure in ways that are not "preprogrammed" for the solution of specific problems, but in general ways that gradually improve their performance. The rules governing this learning and improvement, however, are themselves just algorithms (otherwise it would not be possible to simulate them on a general-purpose computer). Sometimes there may be random choices involved. Strictly, the concept of random choice would go outside what one means by algorithmic, but not usefully so. In practice, random choices are very effectively simulated in a purely algorithmic way, by the use of pseudo-random number generators. Similar remarks apply to "heuristics" (and whatever other similar terms might be or might subsequently be employed). As soon as they have been programmed on a general purpose computer, their algorithmic nature has been secured.

A number of commentators (**Boyle, Breuel, Doyle, Lutz, McDermott**) have criticized me for not addressing the question of *learning* at greater length; they appear to claim that learning is not algorithmic. In fact I did briefly refer to the question of learning in *Emperor*, in connection with neural networks (pp. 397–98), but I did not say much (and certainly nothing significant) about it. Evidently I should have done, especially now in view of such criticisms, but I had not thought that people would have taken the view that there is something new in principle here, and that some would regard a learning system as anything other than a particular type of algorithmic system. The procedures whereby the system is supposed to learn are always preprogrammed into the system right at the beginning (unless a human intervenes to modify the system at some stage – but that is clearly cheating!). A random element may be included, of course, but in practice this simply means using an (algorithmic) pseudorandom number generator. This is all just as algorithmic as before – as should be obvious from the fact that such learning systems are (or at least can be) run on an ordinary computer. Of course, a learning system is supposed to be continually influenced by its environment, so there is an external input all the time. But that is just like the Turing machine's tape, which is being continually read by the machine – and, indeed, the whole original point of the tape was that it was supposed to model the environment!

Likewise, I do not see what is to be gained by searching for the concept of "perception" in the relation between an animal and its environment (**Gilden & Lappin, Waltz & Pustejovsky**). Again, we have not got away from the Turing-machine model. The same applies also to **Varela**'s feeling that "movement" and "bodies" would make an essential difference. Actual physical movement is irrelevant to a Turing machine's action, and I find it hard to see what it has to do either with computations or consciousness. **Eagleson** refers to "intelligence" perhaps residing in "semantic levels of algorithms," but, as we have seen earlier, semantics cannot be encapsulated within an algorithm, so we are really no closer to understanding what intelligence actually is, on these terms.

Lutz seems to be saying that a human mathematician could provide provisional answers to problems all the time which are continually being improved, so it is as though we were allowed to have our Turing machine provide such provisional answers before it has actually reached a *stop* instruction. I do not really see how this buys us anything new. Such a device would need to specify the stages at which its output was allowed to be examined, and that is just like putting in a *stop* instruction at that stage instead. When it starts up again, it would just be another running of the Turing machine's action. Again, it is easy to see that all we get is something algorithmic in the old sense – clearly, because all these things can be run on an ordinary computer. Likewise, **Perlis**'s tapes, which "vary in time," do not really give us anything new if the way the tapes vary is itself well defined (and in principle computable, or computable but with random elements also). Turing's original analysis really covered all this sort of thing. The same applies to Perlis's "enlarging formalisms"; and Lutz's possibly nonterminating Turing machines are certainly algorithms, in the sense that I am using the term. That's precisely what a (possibly dud) Turing machine is. Nonterminating action must certainly be admitted. There is, after all (as **Gigerenzer** reminds us), no general (algorithmic) way of deciding whether or not a Turing machine's action does terminate!

Breuel refers to continuous systems subject to differential equations as being "not intrinsically restricted to Turing equivalent computation." In fact, a significant portion of *Emperor* was devoted to this very question. Not a great deal of a rigorous mathematical nature seems to be known about it (except for the seminal work of Pour-El & Richards, and there is the new theoretical framework of Blum et al. – both of which I refer to in *Emperor* and my *BBS* Précis). As I tried to argue in *Emperor*, there does not seem to be anything usefully nonalgorithmic in continuous action according to standard physical laws, which is one of the reasons I believe we must look to CQG for the needed usefully nonalgorithmic action. The case is no doubt arguable, but Breuel does not even mention it!

Breuel, Doyle and **Manaster-Ramer, Savitch & Zadrozny** refer to Turing's "oracle" machines and "relative computability" (and "infinite nets"). These ideas certainly do go outside what is meant by algorithmic. (An oracle is a putative device that can perform nonalgorithmic actions of the nature of giving correct yes/no answers to, for example, all propositions of type [Q] above.) The trouble is, of course, that no one knows how to build an oracle; moreover, the very possibility of constructing such a device would be denied by all those (including strong-AI supporters) who believe that we live in an algorithmic universe. Those who rest their hopes on the

possibility of constructing such oracles should be pursuing my general line of endeavour and trying to locate nonalgorithmic ingredients in actual physical laws! My own proposals were, in this respect, more modest, since (despite what **McDermott** seems to suggest) I nowhere insisted that an oracle must actually be constructable but merely that nonalgorithmic ingredients be present in physical laws!

Glymour and Kelly present an argument purporting to show that even if human mental output is nonalgorithmic, we shall never be able to ascertain this fact. There seem to be two parts to the argument. First, the "we" in the above claim might actually be Turing machines; and I am certainly prepared to accept Glymour & Kelly's argument that Turing machines cannot "tell" whether or not another object is a Turing machine. The claim tells us nothing of relevance to my own arguments, however, since I am claiming that we are not Turing machines! Second, Glymour & Kelly seem to be claiming that, on some rather obscure limiting interpretation ("Δ_2 property of the Borel hierarchy"), the argument still goes through even if the "we" are not Turing machines (i.e., we are instead "noncomputable scientists"). The argument seems to boil down to something like the following: Can an oracle recognize when some other object is merely a Turing machine? An oracle would be an example of an "noncomputable scientist." We may assume that the Turing machine the oracle is examining is one that considers the truth of statements of the form [Q] one after the other, and correctly asserts *yes* or *no* in each case, except that sometimes it will get stuck and run on forever without ever giving any answer at all. In those cases the poor oracle will just have to sit it out forever (itself knowing the correct answer all the time), but at no stage will it be sure that the Turing machine isn't just a very slow oracle that hasn't yet come out with the answer!

Technically **Glymour & Kelly** are right, if I can take it that this is the kind of thing they do mean, but this is adopting an absurdly narrow view of what science can (and has been able to) achieve. We could likewise argue that the confirmation of *any* scientific theory depends on only a finite number of data points, and on a finite number of bits of information. There are infinitely many different physical theories that could fit those data, and no "scientist," in Glymour & Kelly's sense, could ever be sure that the appropriate theory had been confirmed. Yet scientific progress has undoubtedly been made – and superbly impressive progress at that. A good measure of common sense (or appropriate judgements with regard to one's statistical analysis, as **Gigerenzer** very usefully points out) is always required in the (albeit temporary) confirmation of any physical theory. Precisely the same considerations will become relevant when CGQ (or whatever turns out to the appropriate nonalgorithmic theory) finally emerges. Considerations such as those of Glymour & Kelly will not be (or at least should not be) any bar to progress!

The relevance of Gödel's theorem to cognition and AI

Several critics have argued that Gödel's theorem is *irrelevant* to the important issues of mind and AI. I completely disagree. The role of Gödel's theorem is *crucial*, though I am certainly not trying to suggest that one has to be able to appreciate Gödel's theorem to be conscious! (I can hardly be suggesting that, since in *Emperor*, I say that I believe that at least certain nonhuman animals – dolphins, chimpanzees, dogs, . . . ? – have some degree of consciousness.) Clearly one does not have to be a mathematician (or a mathematician actually doing mathematics) to be conscious. But the central role for Gödel's theorem is that it is only *here* (or with mathematical reflection principles generally) that one can demonstrate, on anything like rigorous mathematical terms, that our conscious understanding *must* be nonalgorithmic. It is, indeed remarkable that any such clear general statement about the nature of our thinking can be made at all! Without Gödel's theorem (or even with it, before the full impact of that theorem is properly appreciated) one would have to resort to much vaguer and less conclusive arguments about "semantics," "understanding," "insight," and perhaps "inspiration."

In *Emperor*, I use such arguments also – and I am criticized by **Boyle, Breuel, Eagleson, Perlis, Stanovich, Taylor, Tsotsos, Waltz & Pustejovsky** for "simply" using inconclusive and anecdotal arguments from "introspection" to support my case for nonalgorithmic action. These reviewers have simply ignored the powerful argument from Gödel's theorem, however (and others, such as Wilensky, seem to underestimate it). I never intended that the case for a nonalgorithmic ingredient to our conscious thinking should rest on anecdotal arguments of this kind. These provide only a supplement to the powerful Gödelian case, providing some additional tentative clarification as to what must really be going on. These arguments are for those who are already prepared to entertain the probability that some ingredients of our thinking are nonalgorithmic, and who are interested in gaining some insights into the possible nature of those nonalgorithmic ingredients. Like **Higginbotham,** I certainly do not believe that the nonalgorithmic quality of human thought is something restricted to the appreciation of sophisticated mathematical thinking. If it were, it would never have arisen by natural selection! It must have been useful for such things as the conception of mammoth traps and the like – or for apes perceiving the value of using tools! It is the elusive quality of *understanding* (and things related to it) that has this nonalgorithmic character – and this is what Gödel's theorem demonstrates, albeit (of necessity) in a particularly sophisticated context.

I have taken the line that consciousness is a quality that *must* be present for a nonalgorithmic (Gödelian) activity of the brain to come into play. Here I would agree with **Butterfield** and others that my case is not so strong, resting, as it does, on the fact that consciousness is (as I believe) essential for actual understanding (and hence for the appreciation of the truth of Gödel's theorem). From the different point of view of his "Chinese Room," John Searle has also used "understanding" (and related concepts such as "semantics" and "intentionality") as a fundamental quality that cannot be evoked merely by computation. These arguments do not directly address other aspects of consciousness, but I feel sure that they must ultimately have relevance for them. [See Searle: "Consciousness, Explanatory Inversion, and Cognitive Sci-

ence" *BBS* 13(4) 1990.] **Butterfield, McDermott, Niall, Roskies,** and **Stanovich** take me to task for not distinguishing various *different* aspects of consciousness and for not being precise enough about what I do mean by the term. Perhaps I should have been more explicit, but I am not sure that this would have been helpful. I agree that any direct connection between the Gödelian insight and "red" or "pain" or "pride" or "hope" would be pretty hard to discern; but it seems to me that there is some sort of unified concept of consciousness, albeit a difficult one to pin down, and any small progress toward understanding some part of this concept will lead ultimately to progress toward understanding the rest of it.

McDermott objects to my "oneness" of consciousness, but I think he totally misunderstands what I mean. I know that consciousness can take multifarious forms, and also that the mind (where this term includes the unconscious mind) can indulge in many activities simultaneously. Yet, especially considering the incredible amount of unconscious activity that goes on at once in the brain, it is quite remarkable that *consciously* we feel like a single person at all. I feel sure that Freud and James would agree with me that this aspect of consciousness is quite unlike the action of a parallel computer – which was my point!

Libet raises the point that much of what is *un*conscious in mental activity could well be nonalgorithmic also (I think **Gigerenzer** and perhaps **Roeper** are making a similar point). On reflection, I think I might well agree with them. I am certainly taking the line that CQG (or whatever turns out to be the appropriate physics spanning the two quantum processes **U** and **R**) is something that can (and does) take place independently of consciousness, and I am also taking the line that this physics should turn out to be nonalgorithmic. From this I ought indeed to be inferring that important nonalgorithmic actions are taking place at much lower levels than the level of consciousness. I think, however, that consciousness itself must have some *special* role to play in relation to question of noncomputability (something to do with the level at which nonalgorithmic aesthetic truth-judgements begin to be important?), but the understanding of these matters must lie a long way beyond even CQG!

Be that as it may, AI researchers will ignore the arguments from Gödel's theorem at their peril. Mathematical thinking, although a tiny minority activity, is thinking, after all, and if that is demonstrably nonalgorithmic, then nonalgorithmic thought is shown to be possible. That is all we need from the argument. I believe that most AI people are grossly underrating the importance of a possibility of nonalgorithmic control (perhaps as exhibited by conscious action?) as a separate quality from computation – that could indeed (**Wilensky** take note) even be relevant to one's walking down the hallway!

My inadequate treatment of AI

Several critics (**Perlis, Smithers, Tsotsos, Waltz & Pustejovsky**) have chastised me for being out of date, misleading, or inadequate in my treatment of artificial intelligence. As I confessed in *Emperor,* I am aware of only a small part of the current activity in this subject, and I fully admit that a somewhat more comprehensive treatment of this activity would have been helpful. It is in the nature of this activity, however, that although there is much of it going on, I have been unable to perceive any fundamentally new principles that have emerged, of relevance to what I had to say. A survey would have amounted largely to a list of things that people in different parts of the world are doing, and that would not have been very helpful.

Many (e.g., **Perlis**) rebuke me, naturally enough, for not having referred to their own work, while **Smithers** takes me to task particularly for not referring to the "foundational" 1976 paper of Newell & Simon, where a form of "symbolic manipulation" that he claims is more relevant than Turing machine algorithmic action is described. I cannot understand why he thinks there is anything new in principle here. Newell & Simon's system, for all its virtues with regard to practicality, and so on, is nevertheless still subsumed by Turing's simple yet comprehensive original scheme. A trouble, of course, is that coming to the subject from the outside, I do not see clearly which publications AI people think I should be concentrating on. For example, no critics other than Smithers tell me that I should have referred to Newell & Simon's (1976) paper, although it receives a passing mention by **MacLennan.** I am quite prepared to accept that the paper may be important to theoretical AI, but I still do not see how it affects the arguments that I have been putting forward in any substantive way.

Tsotsos tells me that my references are out of date (though I did describe what must surely be the most impressive achievements of artificial intelligence to date: the development of such chess-playing machines as "Deep Thought," with its success in co-winning an international chess tournament in late 1988 and defeating a grand master). He thinks I would be upset if, in a description of a *physical* theory, all the references were old ones; yet the most recent reference to electromagnetic theory that I gave in my own book was to Maxwell, 1865. The important thing is to be right (like Maxwell), not necessarily to be recent! I have yet to be convinced that any substantial new advance in the theory of AI has occurred in very recent years. Chess-playing machines apart, the early things I described still seem to be the most striking achievements of AI.

Smithers maintains that I am wrong in attributing to AI workers the desire to simulate feelings of any kind, such as pleasure and pain (cf. pp. 14–17 of *Emperor*). But surely these *are* among the intentions of the supporters of *strong* AI, since *all* mental qualities are supposed to arise from computation. Weak-AI practitioners need certainly have no such intentions, and I apologize to them if they feel that they have been misrepresented. Nevertheless, the very name "artificial intelligence" implies that even the practitioners of weak AI must be attempting to simulate genuine intelligence, and actual intelligence is something that (in my view of what the word means at least, and despite what **Stanovich** says) implicitly involves some degree of awareness. Simulated pleasure and pain are certainly aspects of simulated awareness – and **Doyle** claims that a limited simulation of actual consciousness has *already* been achieved (though how he knows this is hard for me to fathom!).

Pointers toward a new physics

Comparatively few commentators (**Butterfield, Davis, Kentridge, Lutz, MacLennan, Manaster-Ramer, Savitch & Zadrozny, Madsen, Tsotsos, Varela, Zytkow**) address the actual physical ideas I put forward in *Emperor,* despite the fact that the physics is the central theme of the work. Perhaps it is unfair to expect much discussion of such matters in a journal mainly devoted to issues of psychology, neurophysiology, and philosophy, but I am nevertheless disappointed at the overall response here because of the importance of the issues raised. **Varela** refers to my "wild speculations" (which he appears to place on a par with even the fantastical physical suggestions of Moravec 1988), whereas **Madsen** refers to my "extremely tall and particularly shaky edifice" to be accepted "on faith alone." They are unable to point to any actual flaws or omissions relevant to my chain of reasoning, however (although I consider that Madsen is being complimentary when he says that my book "raises more questions than it answers"!). Madsen asserts that I ignore Mott's work on α-particles (citing no reference – but presumably he means a paper I actually refer to in my reference list in *Emperor* – Mott 1983), appearing to claim that Mott showed that large quantum systems will necessarily behave like classical ones – though Mott actually did nothing of the sort; nor was he trying to. Madsen is upset that I give short shrift to the still somewhat fashionable inflationary models (though my placing them in the TENTATIVE category is surely unexceptionable). Such models usually do depend on GUT theories; but the fact that some may not is neither here nor there. They do not in any way alter the arguments that I give in my book (see Penrose 1989 for a critical discussion of the relevant points). Madsen closes with some very odd ideas about "particles . . . using Turing's methods to compute their trajectories in phase space," although such ideas certainly played no role in my own discussions – nor would a theory of the (deliberately?) absurd type suggested by **McDermott** at the beginning of his critique.

Various criticisms refer disparagingly to my search for ideas that might lead to an improved quantum mechanical theory that would be more satisfactory with regard to the "measurement problem" ("vague thought," says **McDermott,** "purest speculation" according to **Madsen** – though these critics offer no counterarguments, apparently forming their judgements on the basis of my own apparently over-modest reference to my suggestions being "only a germ of an idea," p. 371 of *Emperor*). **Tsotsos** and **Zytkow** appear to have totally misunderstood my "one-graviton criterion," incorrectly thinking that it refers to effects at 10^{-33} cm or else to *real* graviton emission or absorption. No doubt I was not clear enough: Only virtual, or *longitudinal* gravitons are needed (cf. Note 5 on p. 372). These can occur at a much tinier level of energy-difference than real gravitons, and the relevant scales of distance can be as large as you like (e.g., ten metres or so, as in the Aspect experiment; cf. Aspect & Grangier 1986). Moreover, I am referring to *all* wavefunction collapse, not just the "very special circumstances of quantum gravity" **Zytkow**; and that would be taking place in the brain all the time.

I should have thought I had made it clear that in my view, wavefunction collapse (**R**) occurs spontaneously, quite independently of the presence or absence of conscious observers. I do *not* hold to the view that it is the "interaction between minds and the rest of the universe" (**MacLennan**) that leads to **R**; nor do I believe in a "subjective view of the state vector involving conscious observers" (**Boyle**). Moreover, I did not say that I believe the brain to be a quantum computer, in the presently understood sense – though there may well be aspects of that conception in brain function; and despite what MacLennan says in a footnote, there are very fundamental distinctions between quantum and classical linearity.

I also find it remarkable how many AI people appear to regard the many-worlds interpretation as "obviously true" (**MacLennan, McDermott**), despite its numerous problems and (deserved) unpopularity in the physics community as a whole – and the fact that it does not (in the absence of further nonstandard ingredients) allow one to derive the correct quantum probabilities. (The "observer being part of the physical universe," which I believe in as much as does MacLennan, is no explanation of why a many-worlds universe looks like one world!) I trust that their reasons for believing in the validity of the AI programme are more soundly based.

Kentridge's remarks about the possible effects of quantum processes in computation are well thought through and helpful (if perhaps somewhat inconclusive), but I do not think that **Mortensen** or **Roeper** have yet grasped the point that quantum mechanics is not just a "fuzziness" in our classical descriptions, but a theory of the utmost precision. Quantum processes can lead to effects that are not possible to achieve by classical means. I should point out to **Lutz,** however, that Deutsch's (1985) analysis of quantum computation is that it does not, in itself, lead to nonalgorithmic behaviour – which is one reason I believe that more (e.g., CQG) is needed.

Butterfield refers to my speculations, in relation to the timing of consciousness, that putative causality-violating effects might sometimes *not* lead to inconsistency (because of the timelessness of the Platonic world). His criticisms are indeed pertinent, and I agree with him that any too blatant causality violation by a "Platonic realization" could lead to serious consistency problems. My remarks were supposed to be exploratory only, however, and were intended to indicate that there might be more scope for temporal nonlocality in consciousness than one might otherwise have thought. Whether or not any kind of a consistent scheme can be worked out remains to be seen.

I had in mind, with these considerations, the curious time delays involved in the experiments of Kornhuber and of Libet (Libet et al. 1979). I am grateful for **Libet**'s reassurance that I had not misrepresented him, and also for his comments in relation to his own more recent work in following up the Kornhuber-type experiments. (I had become vaguely aware of these more recent experiments just before I completed my book, but I had not had the opportunity to follow them up.) It seems to me that there are still some very puzzling aspects of the combined force of all these important experiments (despite what **McDermott** says), but I think I had better reserve my judge-

ments until I have had the opportunity to examine these newer results a little more closely.

Perhaps I may comment, at this point, that I am also grateful to **Eccles** for his remarks, and that I am very much in agreement with his belief that scientific methods – and particularly neurophysiological experiments, such as those he and **Libet** have pioneered, in addition to a deeper probing of physical laws – must hold the key to our eventual understanding of the mind/brain issue. I shall try to follow up his comment concerning the "elimination" of the brainstem centre theory of consciousness (and also his recent work referred to in his commentary). Experimental verification of all these ideas (my own included) is always the final test. The one-graviton suggestion for the onset of **R**, for example, should certainly be experimentally testable, and certain tentative ideas along these lines have already been mooted.

If the physical suggestions I have been putting forward can be formulated as a fully fledged nonalgorithmic CQG, then that theory (including its very nonalgorithmic nature – in spite of **Glymour & Kelly**'s remarks) will eventually also be put to the experimental test. If nonalgorithmically behaving objects are ever shown to exist in nature, then that will alter the whole focus of what we actually mean by a "machine" or, rather, by an analogue machine. (According to the terminology of **Chalmers** and **Manaster-Ramer, Savitch & Zadrozny,** the Church(-Turing) thesis would then be false – though I prefer the original terminology according to which that thesis refers to a *mathematical* idea, now almost universally accepted as true.)

Platonism, mysticism, and dualism

Several commentators (**Garnham, Higginbotham, MacLennan, Varela, Waltz & Pustejovsky**) express discomfort (or worse, **Madsen**) with the idea of mathematical *Platonism* – some seeming to identify it with mysticism. Yet the evidence from the majority of mathematicians (who appear to be Platonists of some degree, cf. Davis & Hersch 1982) should not be ignored. It is the mathematicians, after all, who know most about their subject. Though Platonism does have certain difficulties (mainly to do with deciding which enormous sets should be considered actually to exist, or the problem of deciding when to accept a "proof" as a proof), the various alternative viewpoints (formalism, intuitionism, finitism) all have much more severe difficulties (cf. **Garnham**). **Higginbotham** writes that the mathematicians have "had to learn to live with incomplete theories," but it is only the formalists whose mathematics has been rendered incomplete by Gödel. The Platonists are not so constrained. Moreover, the evidence for a Platonic existence for such things as the Mandelbrot set seems to me to be overwhelming; nor does a pragmatist view (as expressed by **Garnham**) come to terms with the extraordinary agreement between mathematicians as to whether or not a mathematical result is true, or with the very fact that they settle their issues by abstract reasoning. It would have been helpful if some of the opponents of Platonism had provided more convincing arguments against such powerful cases for a Platonist view. The facts that certain mathematical concepts may take many years to take shape and that nonstandard analysis can coexist alongside standard analysis (just as nonEuclidean geometry can coexist with Euclidean geometry) are, despite what **MacLennan** seems to be claiming, perfectly consistent with Platonism.

As for mysticism, I have never myself made that association. Platonism is, to me, totally consistent with (and even a concomitant of) a completely scientific viewpoint. But it if is insisted that Platonism entails mysticism, then I shall insist that it is the only form of mysticism I can accept. What about dualism? **Waltz & Pustejovsky,** not to mention **MacLennan**, seem to be insisting that I am a dualist, presumably because of my Platonistic opinions. If I am, then it can only be in the sense that the strong-AI people are also dualists – because they believe that Platonically existing algorithms provide the substance of awareness. (**Higginbotham** has trouble with the apparently temporal aspect of my argument: "Strong-AI supporters must be Platonists, because if algorithms exist only in minds, pre-existing minds are needed for algorithms and pre-existing algorithms for minds!" But I intended this argument to be taken logically, not temporally, and then the circularity is surely clear.) If my Platonism (with regard to the nonalgorithmic procedures I believe must underlie consciousness) makes me a dualist, then, again, that would be the only form of dualism I could accept. Unlike the AI supporters, however, I have not made the claim that it is the mere enaction (or existence) of some mathematical process that evokes consciousness.

Perlis writes about the "dynamic action of algorithms" as being what AI people are concerned with – and presumably what strong-AI people would claim evokes awareness – rather than the mere static existence of an algorithm. This is totally unclear, however. "Dynamic action" presumably means that what counts is the moving around of bits of matter in accordance with the algorithm. The (strong) AI position seems to be that it does not make any difference what the bits of matter that are being moved around actually are. If the nature of the matter is irrelevant, why is the matter itself relevant? What about showing a film of the pages in a book in which the working of the algorithm is written out – or just running one's finger (or a little window in a piece of paper) down the page? The viewpoint is unclear at best.

ACKNOWLEDGMENT

I am grateful to Angus MacIntyre for very useful comments and for his support; also thanks go to Rob Baston for his patience and use of his E-mail address.

NOTES

1. Ironically, a truly committed formalist would accept G(F) *only* if it is false! For only then is it actually a theorem of F.

2. Examples of such Q(n)s would be: "6n is divisible by 3"; "if n is prime, then 2n + 1 is prime"; "n is the sum of four squares"; "2n + 4 is the sum of two primes". For the corresponding universally quantified proposition [Q], we have, in each case: The first is obviously true, the second false, the third nonobviously true (by a theorem of Lagrange), and the fourth (Goldbach's conjecture) unknown. Although I suppose that a case can be made (according to a fairly extreme form of intuitionism) that it is not even meaningful to say whether statements of the form [Q] are either true or false unless they have been "demonstrated" to be, none of my critics have followed that unreason-

able (to my mind) stance (except for a brief allusion to it by **Mortensen**) and I shall not pursue it.

3. If we allowed an infinite set of theorems here, we would need some procedure for algorithmically generating them. The argument given in the text still applies (cf. pp. 109–10 of *Emperor*). **Chalmers** please note.

4. In fact, one can strengthen the argument I gave, since it applies to any "stopping-tester" H(m;n) which is allowed even to make mistakes or to run on forever – provided that it does not ever assert that $T_m(n)$ runs on forever when in fact $T_m(n)$ stops (the only directly falsifiable case!). (Thus, H(m;n) = 0, 1 or □, if $T_m(n) = \square$; and H(m;n) = 1 or □, if $T_m \neq \square$. Here, in the notation of my book, '□' means "does not stop" and H = 1 asserts that the action of $T_m(n)$ stops while H = 0 asserts that it does not.) We construct k so that $1 + T_n(n) \times H(n;n) = T_k(n)$ and consider the action $T_k(k)$. We can see that this action does not stop, but H either gives no answer or gives the wrong answer. It would not be too hard to construct a completely explicit Turing machine T_q that gives k in terms of the Turing number of H (with q probably less than half as long again as the number u that I gave in my book for T_u to be a universal Turing machine). One might think that since there is an algorithm for producing the number k that defeats H, we could put the whole (Turing-)Gödelization process on a machine, but this is not so, for the same reason as in the argument in the main text.

5. According to a *reflection principle*, an understanding of the *meaning* underlying the rules of a formal system F can provide new mathematical truths not accessible within F (see p. 110 of *Emperor* or, more extensively but with a somewhat different emphasis, Rucker 1984).

References

Agree, P. (1988) The dynamic structure of everyday life, Artificial Intelligence Laboratory Technical Report 1085, Massachusetts Institute of Technology. [FJV]

Albus, James S. (1981) *Brains, behavior, and robotics*. BYTE Books. [RE]

Allport, D. A. (1980) Attention and performance. In: *Cognitive psychology: New directions*, ed. G. Claxton. Routledge & Kegan Paul. [KES]

Anscombe, G. E. M. (1981). *Metaphysics and the philosophy of mind*. Basil Blackwell. [KKN]

Armstrong, D. M. & Malcom, N. (1984) *Consciousness & causality*. Basil Blackwell. [KES]

Arnhart, L. (in press). Aristotle, chimpanzees, and other political animals. *Social Science Information*.

Aspect, A. & Grangier, P. (1986) Experiments on Einstein-Poddsky-Rosen-type correlations with pairs of visible photons. In: *Quantum concepts in space and time*, ed. R. Penrose & C. J. Isham. Oxford University Press. [rRP]

Baerends, G. P. & Kruijt, J. P. (1973) Stimulus selection. In: *Constraints on Learning*, ed. R. A. Hinde & J. Stevenson-Hinde. Academic Press. [DH]

Barrow, J. O. & Tieler, F. J. (1986) *The anthropic cosmological principle*. Oxford University Press. [MSM]

Bell, John L. (1986) A new approach to quantum logic. *British Journal of the Philosophy of Science* 37:83–99. [RE]

Bell, J. S. (1987) *Speakable and unspeakable in quantum mechanics*. Cambridge University Press. [aRP]

Benacerraf, P. (1967) God, the Devil, and Gödel. *The Monist* 51(1):9–32. [JH, AR]

Berliner, H. (1988) New hitech computer chess success. *AI Magazine* 9(2):133. [TS]

Bishop, E. (1967) *Foundations of constructive analysis*. McGraw-Hill. [BM]

Blum, L., Shub, M. & Smale, S. (1989) On a theory of computation and complexity over the real numbers: NP completeness, recursive functions and universal machines, *Bulletin of the American Mathematics Society* 21:1–46. [aRP]

Bohm, D. (1980) *Wholeness and the implicate order*. Routledge & Kegan Paul. [DH]

Brooks, R. (1986a) A robust layered control system for a mobile robot, *IEEE Journal of Robotics and Automation*, RA-2, April, 14–23. [DW]

(1986b) Achieving artificial intelligence through building robots, Artificial Intelligence Laboratory Memo 899. Massachusetts Institute of Technology. [FJV]

Caines, P., Greiner, R. & Wang, S. (1989) Classical and logic-based dynamic observers for finite automata. Technical Report, McGill University, May 12. Draft copy. [RE]

Cariani, P. (1989) On the design of devices with emergent semantic functions, unpublished Ph.D. thesis. State University of New York at Binghamton. [DW]

Casti, J. L. (1989) *Alternate realities: Mathematical models of nature and man*. Wiley. [DLG]

Chisholm, R. M. (1957) *Perceiving*. Cornell University Press. [KKN]

Chomsky, N. (1986) *Knowledge of language*. Praeger. [TR]

Churchland, P. M. (1979) *Scientific realism and the plasticity of mind*. Cambridge University Press. [KES]

(1988) *Matter and consciousness*, 2nd ed. MIT Press. [KES]

Churchland, P. M. & Churchland, P. S. (1990) Could a machine think? *Scientific American* 262(1):32–7. [KES]

Churchland, P. S. (1983) Consciousness: The transmutation of a concept. *Pacific Philosophical Quarterly* 64:80–95. [KES]

(1986) *Neurophilosophy: Toward a unified science of the mind/brain*. MIT Press. [KES]

Cutting, J. E. (1986) *Perception with an eye to motion*. MIT Press. [KKN]

Davis, M. (1977) A relativity principle in quantum mechanics. *International Journal of Theoretical Physics* 16:867–74. [MD]

Davis, M., Matijasevic, Y. & Robinson, J. (1976) Hilbert's tenth problem: Diophantine equations; positive aspects of a negative solution. *Proceedings of Symposia in Pure Mathematics* 28:323–78. [MD]

Davis, P. J. & Hersh, R. (1982) *The mathematical experience*. Harvester Press. [rRP]

(1986) *The blind watchmaker*. Longman Scientific and Technical. [DH]

Deecke, L., Grozinger, B. & Kornhuber, H. H. (1976) Voluntary finger movements in man: Cerebral potentials and theory. *Biological Cybernetics* 23:99–119. [BL]

Delbrueck, M. (1986) *Mind from matter?* Blackwell Scientific Publications. [DH]

Dennett, D. C. (1965) What numbers could not be. *Philosophical Review* 74:47–73. [AR]

(1969) *Content and consciousness*. Routledge and Kegan Paul. [KES]

(1978a) The abilities of men and machines. In: *Brainstorms*. MIT Press. [AR]

(1978b). *Brainstorms*. MIT Press. [KES]

(1987) *The international stance*. MIT Press. [DP, KES]

(1988) When philosophers encounter artificial intelligence. *Daedalus* 117:284–95. [KES]

(1989) Murmurs in the cathedral. *Times Literary Supplement*, The London Times, Sept. 29–Oct. 5. [DCD, DH]

Deutsch, D. (1985) Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London* A400:97–117. [arRP, RL]

Dewitt, B. S. (1973) *The many-worlds interpretation of quantum mechanics*. Princeton University Press. [aRP, DM]

Dewitt, B. S. & Graham, R. D. (eds.) (1973) *The many-worlds interpretation of quantum mechanics*. Princeton University Press. [aRP, DM]

Dirac, P. A. M. (1938) Classical theory of radiating electrons. *Proceedings of the Royal Society of London* A167:148. [aRP]

Dreyfus, H. (1979) *What computers can't do*. Harper & Row. [FJV]

Eccles, J. C. (1986) Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proceedings of the Royal Society of London* B 227:411–28. [DH, JCE]

(1990) A unitary hypothesis of mind-brain interaction in the cerebral cortex. Proceedings of the Royal society of London B. [JCE]

Einstein, A., Podolsky, P. & Rosen, N. (1935) Can quantum-mechanical descriptions of physical reality be considered complete? *Physics Review* 47:777–80.

Elgot-Drapkin, J. (1988) Step-logic: Reasoning situated in time. Ph.D. dissertation, Computer Science Department, University of Maryland. [DP]

Epstein, R., Kirshnit, C. E., Lanza, R. P. & Rubin, L. C. (1984) "Insight" in the pigeon: Antecedents and determinants of an intelligent performance. *Nature* 308:61–62. [DH]

Everett, H. (1957) "Relative state" formulation of quantum mechanics. *Review of Modern Physics* 29:454–62. [aRP, BM]

Fodor, J. A. & Pylyshyn, Z. W. (1981) How direct is visual perception? Some reflections on Gibson's "ecological approach." *Cognition* 9:139–96. [KKN]

Fodor, J. (1983) *The modularity of mind*. MIT Press. [DW]

Fontana, W. & Schuster, P. (1987) A computer model of evolutionary optimization. *Biophysical Chemistry* 26:123–47. [DH]

Geroch, R. & J. B. Hartle (1986) Computability and Physical Theories. *Foundations of Physics* 16:533–50. [AM-R]

Ghirardi, G. C., Rimini, A. & Weber, T. (1986) Unified dynamics for microscopic and macroscopic systems. *Physical Review* D34:470. [aRP]

Gibson, J. J. (1979) *The ecological approach to visual perception.* Houghton Mifflin. [KKN, DLG, FJV]

Gibson, W. (1984) *Neuromancer.* Gollancz. [CM]

Gigerenzer, G. & Murray, D. J. (1987) Cognition as intuitive statistics. Erlbaum. [GG]

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Kruger, L. (1989) *The empire of chance. How probability changed science and everyday life.* Cambridge University Press. [GG]

Ginsberg, M. (1987) *Readings in non-monotonic reasoning.* Morgan Kaufmann. [DP]

Gödel, K. (1964) What is Cantor's continuum problem? In: *Philosophy of mathematics,* ed. P. Benacerraf & H. Putnam. Prentice-Hall. [KKN]

(1986/1990) *Kurt Gödel, collected works,* vol. I (orig. pub. 1929–1936) & vol. II (orig. pub. 1938–1974), ed. S. Feferman, J. W. Dawson Jr., S. C. Kleene, G. H. Moore, R. M. Solovay, J. van Heijenoort. Oxford University Press. [rRP]

Goldberg, D. S. (1989) *Genetic algorithms.* Addison-Wesley. [DH]

Halmos, P. (1974) *Measure theory.* Springer. [CG]

Harnad, S. (1989) Minds, machines, and Searle. *Journal of Experimental and Theoretical AI* 1(1):5–25. [TS]

(1990) The symbol grounding problem. *Physica D* 42:335–46. [DW, TS]

Harth, E. (1982) *Windows on the mind.* Harvester Press. [BL]

Haugeland, J. (1985) *Artificial intelligence: The very idea.* MIT Press. [DM]

Hebb, D. O. (1954) The problem of consciousness and introspection. In: *Brain mechanisms and consciousness,* ed. J. F. Delafresnaye. Blackwell. [aRP]

Hilbert, D. & Ackermann, W. (1928) *Grundzüge der Theoretischen Logik.* Springer. [MD]

Hillis, D. (1988) Intelligence as an emergent behavior, or, The songs of Eden. *Daedalus* 117(1):175–90. [DW]

Hinman, P. (1978) *Recursion-theoretic hierarchies.* Springer. [CG]

Hodges, A. P. (1983) Alan Turing: *The enigma.* Simon and Schuster. [rRP]

Hofstadter, D. R. (1979) *Gödel, Escher, Bach: An eternal golden braid.* Basic Books. [RWK]

(1985) *Metamagical themas.* Basic Books. [KES]

Holland, J. (1975) *Adaptation in natural and artificial systems.* University of Michigan Press. [DH]

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America* 79:2554–58. [RWK]

Jackson, F. (1982) Epiphenomenal qualia. *Philosophical Quarterly.* [JB]

Johnson-Laird, P. N. (1983) *Mental models.* Cambridge University Press. [AG]

Kahneman, D., Slovic, P. & Tversky, A. (1982) *Judgement under uncertainty: Heuristics and biases.* Cambridge University Press. [RWK]

Kelly, K. (1990) Characterizations of inductive inference on arbitrary sets of sequences. Laboratory for Computational Linguistics, Department of Philosophy, Carnegie Mellon University. [CG]

Kihlstrom, J. (1987) The cognitive unconscious. *Science* 237:1445–52. [KES]

Kleene, S. C. (1950) *Introduction to metamethametics.* Van Nostrand. [AM-R]

Lakatos, I. (1976) *Proofs and refutations: The logic of mathematical discovery.* Cambridge University Press. [BM]

(1976) *Proofs and refutations.* Cambridge University Press. [CM]

Lewis, D. (1969) Lucas against mechanism. *Philosophy* 44:231–33. [JB]

(1979) Lucas against mechanism II. *Canadian Journal of Philosophy* 9:373–76. [JB]

(1990) What experience teaches. In: *Mind and cognition: A reader,* ed. W. Lycan. Cambridge University Press. [JB]

Libet, B. (1965) Cortical activation in conscious and unconscious experience. *Perspectives in Biology and Medicine* 9:77–86. [BL]

(1980) Mental phenomena and behavior. *Behavioral and Brain Sciences* 3:434. [BL]

(1982) Brain stimulation in the study of neuronal functions for conscious sensory experiences. *Human Neurobiology* 1:235–42. [BL]

(1985) Unconscious cerebral intiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529–66. [BL]

(1987) Consciousness: Conscious subjective experience. In: *Encyclopedia of neuroscience,* vol. I, ed. G. Adelman. Birkhauser. [BL]

(1989) Conscious subjective experience vs. unconscious mental functions: A theory of the cerebral processes involved. In: *Models of brain function,* ed. R. M. J. Cotterill. Cambridge University Press. [BL]

Libet, B., Gleason, C. A., Wright, E. W. Jr. & Pearl, D. K. (1983) Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential); the unconscious initiation of a freely voluntary act. *Brain* 106:623–42. [BL]

Libet, B., Wright, E. W. Jr., Feinstein, B. & Pearl, D. K. (1979) Subjective referral of the timing of a conscious sensory experience. A functional role for the somatosensory specific projection system in man. *Brain* 102:193–224. [rRP, BL]

Libet, B., Wright, E. W. Jr., & Gleason, C. A. (1982) Readiness potentials preceding unrestricted "spontanious" vs. pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology* 54:322–35. [BL]

Lucas, J. R. (1961) Minds, machines and Gödel. *Philosophy* 36:112–27. [arRP, JB, RWK, DJC]

Lyons, W. (1986) *The disappearance of introspection.* MIT Press. [KES]

Mace, W. M. (1977) James J. Gibson's strategy for perceiving: Ask not what's inside your head, but what your head's inside of. In: *Perceiving, acting, and knowing,* ed. R. Shaw and J. Bransford. Lawrence Erlbaum Associates. [DLG]

MacLennan, B. J. (1987) Technology-independent design of neurocomputers: The universal field computer. *Proceedings of the IEEE First Annual International Conference on Neural Networks* III:39–49. [BM]

Margenau, H. (1984) *The miracle of existence.* Ox Bow Press. [DH]

Marr, D. (1977) Artificial intelligence – a personal view. *Artificial Intelligence* 9:37–48. [DH]

Martin, R. (ed.) (1984) *Recent essays on truth and the liar paradox.* Oxford University Press. [JH]

Maturana, H. & Varela, F. (1987) *The tree of knowledge: The biological roots of human understanding.* New Science Library. [FJV]

McAlister, L. L. (1976) *The philosophy of Brentano.* Duckworth. [KKN]

McCammon, J. A. & Harvey, S. C. (1987) *Dynamics of proteins and nucleic acids.* Cambridge University Press. [DH]

McCarthy, J. (1979) Ascribing mental qualities to machines. In: *Philosophical perspectives in artificial intelligence,* ed. M. Ringle. Humanities Press. [DP]

McClelland, J. L. & Rumelhart, D. E. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition.* MIT Press. [KES]

McGinn, C. (1987) Could a machine be conscious. In: *Mindwaves,* ed. C. Blakemore & S. Greenfield. Blackwells. [DH]

Mitchell, M. & Hofstadter, D. R. (in press) The emergence of understanding in a computer model of concepts and analogy-making. *Physica D.* [DJC]

Minsky, M. (1986) *Society of mind.* Simon and Schuster. [DW, DM, KES]

Monod, J. (1972) *Chance and necessity.* Random Paperback Knopf. [DH]

Moravec, H. (1988) *Mind children: The future of robot and human intelligence.* Harvard University Press. [rRP, FJV]

Mortensen, C. (1989) Mental images. In: *Computers, brains, and minds,* ed. P. Slezack & W. Albury, Kluwer. [CM]

Mott, N. F. (1983) The wave mechanics of α-ray tracks. In: *Quantum theory and measurement,* ed. J. A. Wheeler & W. H. Zurek. Princeton University Press A126:79–84 (orig. pub. 1929, Proceedings of the Royal Society of London). [rRP]

Nagel, E. & Newman, J. R. (1958) *Gödel's proof.* Routledge & Kegan Paul Ltd. [rRP]

Nagel, T. (1979) *Mortal questions.* Cambridge University Press. [DH]

Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83. [RE]

(1982) The knowledge level. *Artificial Intelligence* 18:87–127. [RE]

Newell, A. & Simon, H. A. (1976) Computer science as empirical enquiry: Symbols and search. *Communications of the ACM* 19:113–26. [rRP. BM, TS]

Nisbett, L. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment.* Prentice-Hall. [KES]

Nisbett, L. & Wilson, T. (1977) Telling more than we know: Verbal reports on mental processes. *Psychological Review* 84:231–59. [KES]

Pearle, P. (1989) Combining stochastic dynamical state-vector reduction with spontaneous localization. *Physical Review* A39:2277–89. [aRP]

Penrose, R. (1989) *The emperor's new mind: Concerning computers, minds, and the laws of physics.* Oxford University Press.

Penrose, R. (1989) Difficulties with inflationary cosmology, Proceedings of the 14th Texas Symposium on Relativistic Astrophysics, ed. E. Fenves. New York Academy of Science. [rRP]

(1990) Matter over mind. *The New York Review of Books,* Feb. 1, 37(1):3–5. [FJV]

Perlis, D. (1987) How can a program mean? *Proceedings, International Joint Conference on Artificial Intelligence,* Milan. Morgan Kaufmann Publishers. [DP]

(1989) Some brief essays on mind. Technical Report 302, Computer Science Department, University of Rochester. [DP]

Popper, K. R. & Eccles, J. C. (1977) *The self and its brain*. Springer Verlag. [JCE]
Posner, M. I. (1989). *Foundations of cognitive science*. MIT Press. [KES]
Pour-El, M. B. and Richards, I. (1981) The wave equation with computable initial data such that its unique solution is not computable. *Advances in Mathematics* 39:215–239. [aRP]
Putnam, H. (1960) Minds and machines. In: *Dimensions of mind*, ed. S. Hook. New York University Press. [CM]
(1988) *Representation and reality*. MIT Press. [KKN]
Pylyshyn, Z. W. (1980) The "causal power" of machines. *Behavioral and Brain Sciences* 3:442–44. [KES]
(1984) *Computation and cognition*. MIT Press. [RE]
Rayner, K. & Pollatsek, A. (1989) *The psychology of reading*. Prentice-Hall. [KES]
Robinson, A. (1966) *Non-standard analysis*. North-Holland. [BM]
Rogers, H. (1968) *Theory of recursive functions and effective computability*. McGraw-Hill. [CG]
Rollman, G. B. & Nachmias, J. (1972) Simultaneous detection and recognition of chromatic flashes. *Perception & Psychophysics* 12:309–14. [KES]
Rorty, R. (1979) *Philosophy and the mirror of nature*. Princeton University Press. [KES]
Rosen, R. (1978) *Fundamentals of measurement and the representation of natural systems*. North Holland. [DW]
(1987) On the scope of syntactics in mathematics and science: The machine metaphor. In: *Real brains, artificial minds*, ed. J. Casti & A. Karlqvist. North Holland. [DW]
Rucker, R. (1984) *Infinity and the mind: The science and philosophy of the infinite*. Paladin Books/Granada Publishing Ltd. [rRP]
Ryle, G. (1949) *The concept of mind*. Barnes & Noble. [KES]
Saltzman, E. (1979) Levels of sensorimotor representation. *Journal of Mathematical Psychology* 20:91–163. [RE]
Schank, R. & Abelson, R. (1977) *Scripts, plans, goals, and understanding*. Erlbaum. [JKT]
Schrödinger, E. (1935) Die gegenwärtige situation in der Quantenmechanik, Naturwissenschaftenp 23:807–12, 823–28, 844–49. [aRP]
Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417–50. [aRP, DP, DM, RW]
(1983) *Intentionality*. Cambridge University Press. [KKN]
(1984) *Minds, brains and science*. Harvard University Press. [GG, FJV]
(1990) Is the brain's mind a computer program? *Scientific American* 262:26–31. [DW]
Seidenberg, M. S. & McClelland, J. L. (1989) A distributed, developmental model of word recognition and naming. *Psychological Review* 96:523–68. [KES]
Sejnowski, T. J. & Rosenberg, C. R. (1988) Learning and representation in connectionist models. In: *Perspectives in memory*, ed. M. S. Gazzaniga, MIT Press. [KES]
Shankar, N. (1986) Proof checking metamathematics. Ph.D. dissertation, Department of Computer Science, University of Texas at Austin. [RW]
Sherrington, C. S. (1940) *Man on his nature*. Cambridge University Press. [JCE]
Simon, H. (1973) The organization of complex systems. In: *Hierarchy theory*, ed. H. H. Pattee. George Braziller. [RE]
(1981) *Sciences of the artificial*. MIT Press. [DW]
Smith, P. & Jones, O. R. (1986) *The philosophy of mind*. Cambridge University Press. [KES]
Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23. [DJC]
Springer, S. & Deutsch, G. (1985) *Left brain, right brain*. W. H. Freeman. [KES]
Stanfill, C. & Waltz, D. (1986) Toward memory-based reasoning. *Communications of the ACM* 29(12):1213–28. [DW]
Stanovich, K. E. (1986) *How to think straight about psychology*. Scott, Foresman. [KES]
(1989) Implicit philosophies of mind: The dualism scale and its relationships with religiosity and belief in extrasensory perception. *Journal of Psychology* 123:5–23. [KES]
Sternberg, R. J. & Detterman, D. K. (1986) *What is intelligence?* Ablex. [KES]
Stich, S. (1983) *From folk psychology to cognitive science*. MIT Press. [KES]
Sumida, B. H., Houston, A. I., McNamara, J. M. & Hamilton, W. D. (in press) Genetic algorithms and evolution. *Journal of Theoretical Biology*. [DH]
Tranel, D. & Damasio, A. (1985) Knowledge without awareness: An autonomic index of facial recognition by prospagnosics. *Science* 228:1453–54. [KES]
Tsotsos, J. K. (in press) Analyzing vision at the complexity level. *Behavioral and Brain Sciences* 13(3). [JKT]
Waltz, D. (1988) The prospects for truly intelligent systems. *Daedalus* 117(1):191–212. [DW]
Whiteman, M. (1967). *Philosophy of space and time*. Allen & Unwin. [DH]
Wigner, E. P. (1961) Remarks on the mind-body problem. In: *The scientist speculates*, ed. I. J. Good. Heinemann. [aRP]
Wilkes, K. V. (1984) Is consciousness important? *British Journal of Philosophy of Science* 35:223–43. [KES]
(1988) *Real people: Personal identity without thought experiments*. Oxford University Press. [KES]
Winograd, T. & Flores, F. (1987) *Understanding computers and cognition*. Addison-Wesley. [FJV]
Wittgenstein, L. (1967) *Remarks on the foundations of mathematics*. Blackwell. [AG]
Wolf, F. A. (1989) On the quantum physical theory of subjective antedating. *Journal of Theoretical Biology* 136:13–19. [DH]
Yolton, J.W. (1984) *Perceptual acquaintance from Descartes to Reid*. University of Minnesota Press. [KKN]
Zohar, D. (1990) *The quantum self*. Bloomsbury. [DH]

***Abstract:** *The emperor's new mind* (hereafter *Emperor*) is an attempt to put forward a scientific alternative to the viewpoint of "strong AI," according to which mental activity is merely the acting out of some algorithmic procedure. John Searle and other thinkers have likewise argued that mere calculation does not, of itself, evoke conscious mental attributes, such as understanding or intentionality, but they are still prepared to accept the action the brain, like that of any other physical object, could in principle be simulated by a computer. In *Emperor* I go further than this and suggest that the outward manifestations of conscious mental activity cannot even be properly simulated by calculation. To support this view, I use various arguments to show that the results of mathematical insight, in particular, do not seem to be obtained algorithmically. The main thrust of this work, however, is to present an overview of the present state of physical understanding and to show that an important gap exists at the point where quantum and classical physics meet, as well as to speculate on how the conscious brain might be taking advantage of whatever new physics is needed to fill this gap to achieve its nonalgorithmic effects.
*From page 643.

HUMANE INNOVATIONS AND ALTERNATIVES IN ANIMAL EXPERIMENTATION

A NOTEBOOK

IN THIS ISSUE

PSYETA

PSYCHOLOGISTS FOR THE ETHICAL TREATMENT OF ANIMALS
Kenneth J. Shapiro, Ph.D., Executive Director
P.O. Box 87, New Gloucester, ME 04260 (207) 926-4817

https://doi.org/10.1017/S0140525X00080778