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.