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 $1\frac{1}{2}$ 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-