

To appear in:

*Artificial Intelligence*

## **The Emperor's Real Mind**

**Aaron Sloman,**

**School of Computer Science**

**The University of Birmingham**

Review of:

Roger Penrose

*The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*  
(Oxford University Press, Walton Street, Oxford, OX2 6DP, England, 1989); xiii+466  
pages, 20.00

**NB:** All page numbers refer to the hardcover version.

### **Abstract**

*The Emperor's New Mind* by Roger Penrose has received a great deal of both praise and criticism. This review discusses philosophical aspects of the book that form an attack on the "strong" AI thesis. Eight different versions of this thesis are distinguished, and sources of ambiguity diagnosed, including different requirements for relationships between program and behaviour. Excessively strong versions attacked by Penrose (and Searle) are not worth defending or attacking, whereas weaker versions remain problematic. Penrose (like Searle) regards the notion of an *algorithm* as central to AI, whereas it is argued here that for the purpose of explaining mental capabilities the *architecture* of an intelligent system is more important than the concept of an algorithm, using the premise that what makes something intelligent is not *what* it does but *how it does it*. What needs to be explained is also unclear: Penrose thinks we all know what consciousness is and claims that the ability to judge Gödel's formula to be true depends on it. He also suggests that quantum phenomena underly consciousness. This is rebutted by arguing that our existing concept of "consciousness" is too vague and muddled to be of use in science. This and related concepts will gradually be replaced by a more powerful theory-based taxonomy of types of mental states and processes. The central argument offered by Penrose against the strong AI thesis depends on a tempting but unjustified interpretation of Gödel's incompleteness theorem. Some critics are shown to have missed the point of his argument. A stronger criticism is mounted, and the relevance of mathematical Platonism analysed. Architectural requirements for intelligence are discussed and differences between serial and parallel implementations analysed.

## 1. Introduction and overview

Most people working in AI or Cognitive Science will probably have heard of this book. Its title suggests that the objective is to debunk AI. Because many people *want* to believe that AI must fail, it is already something of a cult book; and readers are likely to have seen other reviews, heard radio or television discussions, or perhaps even heard Professor Penrose talk on these topics. The December 1990 issue of the *Behavioral and Brain Sciences* journal [15] includes a full treatment of the book, author's summary, comments by thirty seven reviewers from several disciplines, and an unrepentant "Author's response". At first I was put off the book by expressions of disappointment from other readers and reviewers with AI interests, but when I finally decided to see for myself, I found it well worth reading. Although it has flaws discussed below, it ranges, in a fascinating way, over such varied topics as: philosophy of mind, theoretical computer science, artificial intelligence, tiling theory, the Mandelbrot set, philosophy of mathematics, what makes a "superb" theory, the main ideas of classical physics, quantum physics, cosmology (big bang, black holes and all), the nature of time, and neurophysiology. Often Penrose goes into more detail than his main argument requires, for example in his full exposition of the theory of Turing machines, including tricks for encoding algorithms and data in binary sequences. The detail adds to the interest and entertainment even if it does not contribute to the main thread in the book, which is an attack on the "strong AI" thesis discussed below (first formulated by Searle [17]). Whilst agreeing with several other commentators that the attack is unsuccessful, I believe it raises some important questions concerning computational models of mind and the long term goals of AI. In particular Penrose's critics do not address the question how a finite intelligent agent can think determinate thoughts about infinite sets. I shall explore this issue in connection with Gödel's incompleteness theorem.

Penrose makes use of the theorem in his argument that there are aspects of consciousness that cannot be replicated within any computer model, no matter how sophisticated, as long as the model is based on the standard conception of computation as execution of an algorithm. In order to defend this belief he has to explain what computation is and produce an (alleged) example of what can be achieved by human consciousness that is not amenable to a computational explanation. However, since he is no mystic, he tries to offer at least the germ of an alternative scientific theory according to which the human brain is not a computer, but is a physical system of a type that embodies super-computational mechanisms, i.e. mechanisms that are not subject to the limitations of Turing machines or their equivalents.

His hoped-for mechanisms, unfortunately, can only be understood in terms of as yet unachieved advances in physics concerning quantum gravity theory. These, Penrose claims, will one day link submicroscopic phenomena, cosmological phenomena, and mechanisms of the brain. He says:

I am speculating 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)

I shall argue below that this sort of mechanism has nothing to do with requirements for ordinary mental states, and will challenge both the scientific usefulness of the concept of consciousness

used by Penrose, and his claim that we can “see” Gödel’s undecidable sentence to be true. This undermines his main argument that some super-Turing mechanism, based on quantum mechanics, is needed for intelligence.

I am not competent to comment on the sections of the book concerned with cosmology, quantum physics, or more abstruse mathematics. Instead I shall concentrate on the philosophical topics. Some of these topics may not be of interest to those who regard AI as a branch of engineering, but they are directly relevant to AI as the core discipline of cognitive science.

I shall start (in Sections 2 to 6) by analysing and criticising what Penrose says about the “Strong AI thesis” (compare Searle [17]), showing that there are several versions of the thesis requiring different treatment. The critics (e.g. in the commentaries in [15] and [17]) have mostly failed to notice that Penrose and Searle attack a fairly easily demolished straw man, described in section 4, which I have previously called the “Strong Strong AI thesis” (Sloman [23]). Section 5 shows that there is a lack of clarity afflicting the concepts of “computation” or “algorithm” used by some of those who argue whether the Strong AI thesis is true or false: and section 6 argues that such versions of the thesis are too ill-defined to be worth arguing about. It also introduces the notion that intelligence is related to *how* things are done, not simply *what* is achieved. Later sections claim that besides the extreme obviously false AI thesis and the vague and ambiguous versions, there is an interesting family of theses of different strengths and degrees of plausibility. Sections 7 to 12 consider requirements for human-like intelligence, showing that the idea of a single algorithm for intelligence is not relevant, whereas an architecture consisting of a collection of coexisting mutually interacting processes is. (The processes could be, but need not be, based on neural nets). So AI does not presuppose that any one algorithm can suffice for the production of intelligence. This argument depends in part on the different causal powers of serial and parallel systems, notwithstanding the theorem that parallelism cannot increase the class of functions that can be computed and notwithstanding the fact that some inherently parallel virtual machines may be implemented on a time-shared sequential processor. I also assume that what constitutes having intelligence (or mental processes) is not *what* the system does, but *how* it does it: any behaviour can in principle be produced by an unintelligent system, whereas intelligence requires interaction between coexisting mental states. Section 13 argues that the concept of “consciousness”, as used by Penrose, is too ill-defined to be the subject of serious scientific discussion. This leads into an analysis, in sections 14 to 18, of the keystone in Penrose’s attack on AI, namely the argument based on Gödel’s incompleteness theorem. The argument raises important questions concerning mathematical thinking about infinite sets. Some work in AI has attempted to explain or replicate fragments of human mathematical competence (e.g. Bundy [1]), but Penrose poses a challenge that has not yet been addressed. The question is whether any AI system could explain the kind of reasoning displayed by Gödel in his incompleteness theorem. This depends on the ability to think about infinite sets of formulas and numbers. Gödel thought that we had some way of discovering facts about such infinite sets that could not be derived from any axiomatization. Penrose links this to the philosophical theory known as Platonism, which claims that numbers and other mathematical objects exist independently of our thinking about them.

Some Platonists argue that a non-physical mechanism is required to explain how the minds of mathematicians work, though Penrose does not go this far: instead he postulates new kinds of physical mechanisms based on quantum phenomena. I'll try to show that if we (a) don't exaggerate the human capabilities that need to be explained and (b) design the right sort of explanatory architecture, then such quantum mechanisms will not be required. Section 16 presents an argument which both Penrose and most of his critics appear not to have appreciated, based on the fact that in some models of the formal system Gödel's formula is false. This undermines the claim that anyone can see that it *must* be true, and therefore removes the need to explain *how* they see it to be true. Section 18 criticises Platonism as lacking in content. Section 19 recapitulates, lists eight different versions of the AI thesis, and provisionally settles on a relatively mild version, which retains the intention of the stronger versions without being so easily demolished. On this view a mind is essentially a sophisticated control system, which must satisfy a collection of engineering design requirements, emphasizing architectures rather than algorithms. Section 20 concludes the review.

## 2. The attack on Strong AI

The main target of Penrose's book is what he describes (following Searle) as "the strong AI thesis". This kind of attack is always relished by those who wish to be convinced that they are not "mere" computers and that they can do things that could not be explained by algorithmic processes. Such people have also savoured previous attacks, for example by Weizenbaum [31], Dreyfus [3] and Searle [17], [18]. The new attack is slightly different from earlier versions, but like them involves several muddles that need to be cleared up. In particular Penrose seems to be confused about the nature and objectives of AI, and about how mental processes might be explained, or replicated in a machine. I'll start with his mis-representation of AI, by distinguishing different versions of the Strong AI thesis. Only the less extreme versions (if any) are likely to be believed by the majority of AI researchers.

The thesis is sometimes formulated (by both Searle and Penrose) in such a fashion as to be obviously false, and hardly worth attacking. Penrose writes:

The idea is that mental activity is simply the carrying out of some well-defined sequence of operations, frequently referred to as an *algorithm* (Chapter 1, p 17).

I do not believe the *strong-AI* contention that the mere enaction of an algorithm would evoke consciousness (Chapter 10 p 407).

I shall show that there are more and less extreme interpretations of this claim, and that at least two of the more extreme versions can easily be shown to be false without appealing to such complex matters as Gödel's theorem. If some AI practitioners do believe these very strong claims, that's a reason for criticising *them*, not AI. Penrose frequently refers to "the AI people" as if there were some agreed orthodoxy in the field. Of course, any field has its naive, ill-informed, or over-enthusiastic defenders, and AI is no exception, having attracted many computer science, physics, or mathematics graduates with no training in philosophy or cognitive science. However, the field does have more sophisticated adherents, and, in any case, embodies

an approach that goes beyond the beliefs of the practitioners at any one time, just as 18th century physics carried the seeds of much modern physics that would have been inconceivable to 18th century physicists.

I'll show that on the clearest interpretation of the thesis that mental activity is simply the carrying out of some algorithm, it is so patently false that one should try to come up with a more convincing version before mounting an attack on it. More subtle, less obviously absurd, versions of the thesis will be described below, but I have no evidence that Penrose (or Searle) has thought about them. There are also potential confusions concerning different levels of description of the same system, to which I'll return in section 9.

### **3. What's wrong with the strongest AI thesis?**

There are two sorts of objections to the version of the strong AI thesis discussed by Penrose that are more compelling than his own objections. First of all, in its baldest formulation, I shall try to show that the thesis implies that all sorts of absurd things have minds, like marks on paper or even certain very large numbers, or processes like patterns in leaves blown about by the wind. These things clearly have no mental abilities even if they have a computational structure. It turns out to be a subtle matter to reformulate the AI thesis to avoid such objections. The second, more controversial point is that even after such reformulation the thesis as stated postulates a single stream of processing (the execution of an algorithm) and I shall argue that mental capabilities depend intrinsically on something richer than this: a collection of coexisting mutually interacting capabilities, with causal powers which it is not obvious can be implemented in a single algorithm, in the ordinary sense in which an algorithm controls the creation step by step of a sequence of states. Two implementations with the same input-output mappings may have different causal powers of the required kind. I'll now expand on the first objection, leading to a less extreme formulation of the AI thesis, before going into the second objection.

### **4. Abstractions and static structures can't have minds**

As a reminder that nobody is claiming that the alleged algorithm for intelligence is already known, or will soon be known, or even will ever be discovered by mere human beings, I shall refer to it as "The Undiscovered Algorithm for Intelligence", abbreviated as UAI. There is an extreme and clearly false interpretation of the Strong AI thesis that I suspect lurks behind the attacks by both Penrose and Searle when they use phrases like "mere enaction of an algorithm" (in the quotation from page 407 in section 2) This interpretation, henceforth referred to as **T1**, states that any instantiation of the UAI, even a static one on sheets of paper, will have a mind. Something like **T1** is attacked explicitly by Penrose on pages 21-22, where he discusses the claim that a book describing Einstein's brain would be intelligent; and on page 702 of Penrose [15] he mentions it in response to a comment from Perlis. Penrose is here commenting on chapter 26 of [7], where Hofstadter treats the Einstein book not as a static object but as part of an active process, in which its contents are changed. So if anything is intelligent it would be the book together with whatever is making the changes. Penrose apparently does not see this emphasis on "process" as important for the Strong AI thesis. So he seems to be attacking an

extreme interpretation the Strong AI thesis which claims that any instantiation of the UAI, even a static “trace”, would have mental states. This is what I’ve previously called the Strong Strong AI thesis (Sloman [23]; compare Moor [12]).

One source of pressure towards interpreting the Strong AI thesis in the extreme form of **T1** is this: The only known precise characterization of a computation defines it as an ordered set of structures satisfying certain formal relationships, e.g. the kinds of syntactic relations that hold between conclusions and premisses in formally valid inferences or the formal relationships that hold between states of a machine during correct execution of a program. Formal validity and correctness of execution in this sense are purely syntactic properties of ordered sets of structures. Theorems of computer science concerning complexity and computability assume only that computations are such structured sequences. The theorems are equally applicable whether the sequences are produced in time by some causal mechanism, or are abstract static sequences, or are sequences of patterns produced by leaves blown in the wind. In other words, the only precisely defined concept of computation that we know has no intrinsic connection with the notions of time, mechanism and meaning. (This syntactic conception of computation conflicts with the view of computation as an active, semantically based, process outlined by Brian Smith [28]. However, this alternative conception remains obscure, and certainly has not had a major impact on theoretical or mathematical analysis of computing systems.)

If being an instance of the UAI amounts simply to being a computation in this syntactic sense, then it could not, on its own, be sufficient for the production of mental states, because many static objects and abstract objects that obviously are not minds can be construed as computations in this sense, including sets of marks on paper and, via Gödel-numbering, some very large number encoding the sequence of states in an execution of the UAI. A particularly bizarre example would be regions on a blank wall. There are square patches of wall that have not been marked out in any way. Similarly there are portions of the wall shaped like letters and numbers that have not been marked out, though they *could* be. So the regions constituting those patterns exist and some of the unmarked patterns conform to certain specifications of programs. **T1** would then imply that some such regions, being instantiations of the required program, must be intelligent. This is absurd. This fact that computations can be static or abstract objects is not found obvious by people who forget the precise definition of computation and think only about processes that, besides being computations in the mathematical sense, also satisfy other important properties, some of which I’ll analyse below. Anyhow, if Penrose is attacking an extreme view of AI that implies that even Gödel numbers or static marks on paper can have minds, his attack is justified, but does not require much argument. This target is a straw man, as no sensible AI researcher would defend the extreme thesis. (Penrose claims otherwise when pressed by a critic, in [15].)

## 5. Intelligence requires mechanism and activity

A first attempt to make the Strong AI thesis more plausible, might be to stress the words “carrying out” and “enaction” in the quotations above, in order to avoid the extreme claim that abstract static structures can have minds. Perlis, for example writes: “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” (in [15]). He should have added “not even static printed execution traces.”

But this move still leaves ambiguities: If the sort of computation alleged to be sufficient for mind has to be more than just an abstract ordered sequence of structures, what more is needed? We could revise the thesis thus: **T1a** states that any temporally ordered instantiation of the UAI will have a mind. But adding the temporal ordering requirement is not enough. For suppose that by chance the wind blowing leaves around on a forest floor happened to produce a sequence of patterns instantiating the UAI. A Strong AI thesis implying that this process would be a mind is hardly worth attacking. **T1a** may seem too silly to be worth discussing. Nevertheless Penrose at times seems to think it is worth discussing ([15], page 702), and in any case it clarifies the need for less extreme versions, by showing that temporal patterns of computation or algorithm execution are not enough to rescue **T1** from absurdity. Apparently silly examples typically produced by philosophers reveal the need to make one’s theses more precise. If by “computation” we mean something more than temporally ordered structures conforming to rules, we must be prepared to say what more. Searle would probably claim that he did not intend the Strong AI thesis to be expressed in such an absurdly strong form as to allow static structures and abstract entities, or randomly produced sequences of structures, to be computations, even if his words suggested this. I am not so sure about Penrose: his comments on the book describing Einstein’s brain indicate that he was attacking an absurdly strong thesis.

One tempting way to make the AI thesis less extreme would be to require computations underlying intelligence to be both ordered in time and produced by some *mechanism* in which a representation of the UAI controls behaviour. Indeed, the phrases “carrying out” and “enaction” seem to presuppose a causal connection between an explicit representation of a program and the structures produced. So the milder thesis **T2** states that any process in which the UAI is *enacted* will have a mind, where “enaction” implies some sort of causal connection. The problem now is to specify what sort of causal connection between the UAI and behaviour could be sufficient to produce mental states. Ultimately we’ll have to fall back on engineering design requirements for a sophisticated control system.

Analysing the concept of causation is one of the hardest problems in philosophy. Specifying precisely what sort of causal connection is required between program and process is particularly difficult because there are different sorts of causal connections, with different properties. Searle’s “Chinese room” thought experiment [17] involves a causal link between a printed specification of an AI algorithm (e.g. for understanding Chinese) and the behaviour of a human reader. Is this the right sort of “enaction” to produce mental states (e.g. understanding)? Searle said No, thereby claiming to refute Strong AI. Penrose gives him qualified support (on pages 17-23). Some of Searle’s critics said Yes, in their defence of Strong AI. I suspect they were mistaken:

the causation is too loose. I shall try to explain why, though the questions at issue are too ill-defined for knock-down arguments to be possible.

For engineering purposes we need tight causal links. Many people would be unhappy to fly in a plane whose ‘automatic’ flight landing system consisted of a person blindly following printed instructions without knowing what they were about, like Searle blindly following rules for responding to Chinese sentences. The program would not be properly in control if run not by a machine but by a person who did not know what it was supposed to achieve and who had no explicit desire to use his intelligence to support that objective. The whole process would then be subject to the person’s forgetfulness, whims, preferences, etc. Similarly, with Searle as interpreter the UAI program would not be properly in control, and would lack the causal powers required for mental processes. If either **T1** or **T1a** requiring only (a temporally ordered) instantiation of the UAI were true then control issues would be irrelevant. Instead the milder thesis **T2** requires sufficiently reliable links between program and process to satisfy engineering design requirements for an intelligent control system. Because those links are lacking in the Chinese room, **T2** does not claim that mental processes will occur there. So there is no contradiction between this mild AI thesis and Searle’s claim that there is no understanding in the room.

The Strong AI thesis surely needs such modification. From the design standpoint, a mind is (among other things) a control system, and good control mechanisms cannot be allowed to have excessively sloppy control links, like the link between Searle’s instructions and his behaviour, unless that sloppiness is itself part of a well-engineered collection of feedback mechanisms, as discussed below. Searle and his instructions do not appear to be such a system, though the requirements for **T2** are too vague for the argument to be decisive. Can we be more precise about what causal relation between the UAI and behaviour is required for thesis **T2**? Program and behaviour can be related in many ways, and it is not clear which would be sufficient to produce mental states. The next section discusses ambiguities in concepts of control before returning to Penrose and Searle.

## **6. Types of control by programs**

Consider computer demonstrations produced by running a program and saving a record of the succession of output states produced in a file. Later, for purposes of demonstrating the program, the states are displayed in sequence. That is some kind of *enaction* of the original program, and there is a causal link, but I doubt that this replay fits the requirements for intelligence, even if the original process did. The replay does not recreate the same substates with the same causal powers. AI is not committed to the view that an animated replay of the behaviour of an intelligent system is intelligent. What more is needed then? A tempting suggestion is that **T2** should require that the program structure be in constant and direct control of the process, unlike the replay. This would not be nearly as absurd as **T1** and **T1a**, the more extreme theses, but is still not yet sufficiently precise: what does being in “constant and direct control” mean?

Now consider a computer program that is first compiled and then run. The compiled code will run even if the original version is modified or destroyed after compilation: the control by the original is “ballistic,” not “online”. But it could be argued that a version of original program that is implicit in the compiled version is still in control. Does this meet the requirements of T2? This raises the question whether the transformation produced by a compiler preserves the essence a program. Compilation can introduce many transformations, including optimisations that remove portions of code, linearisation of loops, and so on. If we want to say that the original program remains in control because it exists in the compiled code, there are problems about what sorts of transformations the compiler can perform without changing the program. Suppose the compilation process, in order to trade space for speed, produces a giant lookup table or a giant discrimination tree, mapping a history of inputs to outputs. (In general the combinatorics will make this physically impossible because of the storage space required, but ignore that for now, for the sake of conceptual clarification.) Will the machine running the compiled lookup table be executing the same algorithm as was expressed by the original program?

Most people I have questioned (so far) are reluctant to allow that a computer running a giant lookup table mapping input bit patterns from sensors to output bit patterns for controlling motors is intelligent, even if the resulting behaviour is equivalent to what might have been produced by an intelligent system. The key argument is that whoever produced the table would have had to work out in advance all the appropriate ways to deal with all the possible situations that could arise: the machine consisting of table and interpreter would not be solving any problems or taking any decisions, but only using previously computed solutions and decisions. (This is not to deny that intelligent systems may include some learnt lookup tables.) A machine that, in all situations, acted entirely on instructions that had been pre-computed by someone else would not be as intelligent as one that could create some new strategies or solutions for itself, even if the two produced exactly the same behaviour, viewed externally. Why is the difference important? For trivial programs the difference is not as practically significant as for programs designed for very varied situations: in the latter case the combinatorics may make the pre-computed table so large that no physical machine could possibly store it, or search through it to obtain relevant entries in time. (A decision tree might avoid the latter problem). This argument suggests that intelligence requires certain kinds of underlying mechanisms, with particular re-usable, re-combinable capabilities which are able to produce new solutions to problems as required.

The example helps to show that *behaviour alone is not a sufficient basis for attributing intelligence*, even if it is all we normally have to go on. There are more subtle arguments to do with the internal process architecture required to justify the claim that the machine has internal states corresponding to desires, beliefs, hopes, fears, thoughts, etc., all interacting causally with one another. I’ll return to this later.

If there is a UAI that is capable of producing behaviour in the right way, it does not follow that all compiled versions producing the same external behaviour will also do so in the same way. For example, not all will preserve the important state transitions and causal interactions within high level virtual machines: the (theoretical) possibility of compiling to a giant table

mapping inputs to outputs demonstrates this. So not all causal connections between program and behaviour will do.

Programs that are not compiled but interpreted usually have a stronger causal connection with the behaviour they produce because they have *online* control of behaviour via the interpreter, and any change to the program will produce different behaviour thereafter. I.e. there are many true counterfactual conditional statements (statements about “what would happen if ...”) linking the state of the program to the resulting behaviour, as causal links require. (For a complex and subtle analysis of different kinds of counterfactual conditionals involved in the link between program and behaviour see Maudlin [10].) Nevertheless, even the control of processes by an interpreted program may be limited in a variety of ways, for example by a scheduler that restricts time available, a memory manager that restricts access to some regions of memory, a file system that restricts access to some directories or files, a network manager that restricts communication with other machines, an interrupt handler that grabs control in order to deal with keyboard input or some other device, and so on. Similar, but more subtle, diminution of control by a program occurs when the interpreter has some ability to decide what to do, e.g. to prevent errors, to take short cuts, to report actions to some other program, to produce trace printout, and so on. In such cases we do not have a program that is in *total* control. It is more accurate to view the whole system as made up of many causally interacting components each with partial or “soft” control, each able to monitor, modify or restrict the behaviour of others, where the total system has been carefully designed so that the components form an integrated mutually supportive collection of mechanisms, e.g. an integrated plant control system, with many safety checks, feedback loops, etc., all working together to serve the purposes of the whole system. Something similar will be needed for a mind. It is not clear that all this can be produced in any one algorithm. But if the UAI does exist, then only if relevant causal properties are preserved, will an “enaction” of it of the kind referred to in **T2** produce intelligence.

Specifying criteria for “intensional” identity of processes is difficult. Yet without identity criteria we cannot answer the question whether the required features are present in any particular “enaction” of the UAI program, e.g. a process produced by compiling the UAI into a giant lookup table. The question whether this compiled version has any intelligence also remains fuzzy. There are no “correct” answers to these questions: the concepts used are not precise enough for the questions to have definitive answers. They are in part like the question: “Is it noon on the moon when the sun is at its highest above the horizon, or only when the moon is directly above a portion of the earth at which it is noon locally?” The concept of time of day was designed only to cover a limited class of situations, and there is no *right* way to extend it to all new situations. Similarly with concepts like “intelligent”, “mental process” and “same algorithm”.

Even when we cannot define *precise* boundaries for concepts, we can sometimes identify terrain that is well beyond the boundaries. The lookup table is an obvious example. A process in which Searle interprets the UAI is less obviously beyond the requirements of **T2** but it should be clear that because of the potential for disturbance by Searle’s own thoughts, feelings, etc. the Chinese room does not meet engineering requirements for an integrated control system. So if

there is a UAI some other process of enaction of it will be required.

To summarise so far: being a computation or an instance of an algorithm is a purely structural, or syntactic, property that can be satisfied in all sorts of ways that have nothing to do with causation and control and therefore cannot be relevant to minds or mental processes. So the extreme Strong AI thesis **T1** claiming that there is an algorithm, the UAI, instantiation of which is sufficient for the existence of mental processes, must be false. To attack it is to attack a straw man. The less extreme thesis **T1a** requiring only that the computation be ordered in time is hardly less absurd, since randomly produced patterns could “accidentally” satisfy the condition. **T2** is a still milder thesis requiring the computation to be causally related to an explicit program, but there are many different sorts of causal relations and it is not clear exactly what is required. Moreover, if the intelligent system is to have desires, beliefs, hopes, fears, etc. it looks as if the required design will have to involve many interacting processes, not just the execution of one algorithm. So far nobody has produced a version of the Strong AI thesis stating precisely what sort of computational process is supposed to be sufficient for intelligence. (It is interesting that Smith [28] requires all computational processes to have semantic as well as causal properties. This would make circular any attempt to explain the origins of semantics or understanding in terms of computational mechanisms.) There may be an interesting, defensible, Mild Strong AI thesis, but it is not easy to define and so far nobody has defined it. Hence attacking and defending are both premature.

Many people who believe that Searle and Penrose are wrong will regard all this as a trivial matter of finding an explicit definition of a concept of computation that is already intuitively clear to everyone. However, this intuitive clarity may be a myth, if there is no clearly specifiable sharp distinction between the right sort of control of behaviour by program, and other sorts that Penrose (and Searle) might justifiably attack. If so, the whole debate becomes pointless. (However, showing that it is pointless is not pointless, partly because this helps to clarify the still obscure long term objectives of AI.)

I have hinted that consideration of the engineering requirements for human-like intelligence suggests that any version of the Strong AI theses will remain implausible so long as it suggests that any single algorithm executed on a single Turing machine could suffice for the production of mental processes. I'll try to explain why in the section after next. In doing so I shall probably depart from what most AI theorists believe at present, though not from AI practice. Before continuing with that argument let's consider whether general purpose Turing machines are relevant at all.

## **7. Is Turing machine power relevant to intelligence?**

It is surprising to me that Penrose should regard super-Turing power as relevant to intelligence. Even Turing machine power doesn't seem to be relevant, let alone a super-Turing machine, in discussing intelligence. It used to be important to point out to the misinformed that computers (including Turing machines) had many interesting capabilities not found in previously known mechanisms, and which are analogous to capabilities of a mind (e.g. see Turing [30], Sloman [21]). Continued emphasis on Turing machine power, however, may be a hangover from

the days when people used to think that the ability to do mathematics and play ‘difficult’ games like chess were required for intelligence. We now know that many animals that cannot perform these intellectual feats share aspects of intelligence with us that are much harder to explain: for example they can see, plan, build things, take decisions, learn, control movements in a complex and irregular 3-D environment, and so on. The particular capabilities of Turing machines don’t seem to be at all relevant to the capabilities we commonly find in humans and other animals, except when they are performing very specialised tasks, such as solving mathematical problems. Nevertheless Turing machines might be relevant if they provide a mechanism within which other machines with the right powers can be implemented as ‘virtual machines’. I’ll return to this question later.

Even if it is relevant in that sense, full Turing machine power cannot be *necessary* for mental states. There is no evidence at all that cats and dogs, which clearly have mental states, or even human beings (when unaided by external memories), have the power of a Turing machine. E.g. we quickly get into trouble if we have to parse a deeply nested sentence, whereas this would not bother a Turing machine, or even many computers of lesser power. Even when we use external memory aids analogous to the Turing machine’s tape, to help us with calculations or reasoning, we can still make mistakes of many kinds, that no functioning Turing machine would. When this happens we are still awake, thinking, seeing, feeling etc. Behaviour that’s unlike a Turing machine does not indicate a lack of mind. So it is not *sheer* computational power that is required for mentality. If computational abilities enter into mentality at all, it must have something to do with the particular kinds of computations and the particular kinds of mental capabilities, and it is quite possible that many of these require something less than Turing power, and at the same time something more, which I’ll try to characterise below.

Super-Turing power does not seem to be relevant at all. Interaction with the environment can, in principle, cause computers to produce non-computable outputs (e.g. non-computable infinite binary sequences), if the environment includes non-computable information. The combination of environment and computer would then be a super-Turing machine. But it would not be intelligent. Similarly, in such an environment it might be theoretically possible for (immortal) human beings, interacting with the environment, to produce non-computable sequences. But such super-Turing capabilities would not have anything to do with the ordinary requirements for mind or intelligence, discussed below. All this suggests (but does not prove) that the search for SUPER-Turing mechanisms in the brain to explain consciousness or other mental states may be misguided. It is not computational power in that sense that is needed to explain human and animal capabilities, but the right functional architecture.

## **8. Can a single algorithm suffice for intelligence?**

We’ve seen that there’s vagueness in the causal requirements for an algorithm alleged to generate intelligence. There’s another kind of vagueness concerning the difference between a process produced by one algorithm (e.g. the UAI) and a process involving many different algorithms. Penrose apparently interprets an algorithm as a rule or set of rules specifying permitted sequences of changes of state of some structure. Such rules may be expressed in many

different syntactic forms, including the use of recursion, sub-routines, and other concepts found in high level programming languages. This is somewhat vague, but all attempts to make the notion more precise have so far produced formulations that can be proved to be mathematically equivalent to what can be specified in a Turing machine. Moreover, it is possible to prove that any function computed by a *collection* of Turing machines running (synchronously) in parallel can also be computed by a single Turing machine, by showing how the collection can be modelled on a single machine, e.g. by interleaving their operations. So, from a mathematical point of view, the concept of an algorithm is not extended by allowing parallelism. From an engineering design point of view things are very different. The fact that any function computed by a collection of Turing machines running (synchronously) in parallel can also be computed by a single Turing machine, leaves open the question whether there are other important properties, besides the function computed, that may be different in parallel and serial implementations. Speed differences are relatively uninteresting: they can be overcome in principle by speeding up the machine used for the serial implementation, though there may be physical limits to this. Other differences between parallel and serial implementations are deeper.

Consider the control requirements for a collection of co-existing interacting sub-systems. It is sometimes possible to produce the required interactions on a single time-shared processor, by providing a collection of concurrent *virtual* machines, but virtual parallel processes on a single machine sometimes have slightly different causal powers from processes implemented on a collection of machines, even when they do compute the same input/output function. One obvious causal difference that is important from an engineering point of view, though not a mathematical point of view, is robustness: a bug in the scheduler or memory management system, or even the central processor, can make a single-processor system go irretrievably awry, whereas a multi-processor implementation could include compensatory mechanisms, for instance one processor detecting the error state of another and doing something to change it. This distinction can also be relevant to the difference between a single process and several processes running time-shared on one computer. If two (or more) interacting processes are always ensured a fair share of the time by the scheduler, then if one process has a bug, or gets stuck in a dead-end search, it can be re-directed by another. After all that's exactly the sort of thing that happens in an operating system. So sometimes the advantages of parallelism are to be found even in *virtual* machines.

A less obvious point is that a single-processor system simulating N interacting processors would have to cycle through the changes in those processors in sequence. In doing so it would pass through fragile and meaningless intermediate states that don't occur on a true multi-engine machine where all the processors change concurrently. During these intermediate states the machine with virtual parallelism may be incapable of responding coherently to certain inputs. The risks can be reduced if the inputs from the environment are handled by separate processors that buffer all incoming signals until the main processor is ready to handle them (as happens in time-shared computers) but then we are again dealing with a multi-processor system, even though some of the processors perform only lowly buffering functions.

So the need to interact asynchronously with a complex environment introduces a requirement for real parallelism. In a modern computing system there are many hardware components doing different things asynchronously in parallel (co-processors, disc controllers, memory management units, serial line interfaces, etc.) Some aspects of these mechanisms could not be replicated on a Turing machine without the addition of transducers that could cause either its machine table or its tape to be altered under the influence of external events, such as incoming mail, users typing commands, new programs being developed, and so on. The normal theorems about limits on outputs that Turing machines can produce, if simply given a prepared tape and allowed to run, would no longer apply.

We can sum up the second difference between simulated and true parallelism thus: for a truly parallel machine immediate transitions between remote points in its state space are possible that are not possible for a serial machine: the latter has to traverse a path linking the points. So parallel implementations may have different causal powers from serial implementations of the same collection of algorithms, despite their equivalence at computing input-output mappings. (Tim Read has pointed out in conversation that this feature of simulated parallelism would not matter if the environment were itself a simulated ‘virtual reality’ with clock-steps synchronized with the time required for the intelligent agent to switch between processes.)

The fact that parallel and serial implementations can compute the same input/output relations, yet have importantly different causal properties from a practical point of view is another example of the point made in Section 6 that in describing a complex system, not only *what* behaviour is produced, but also *how* it is produced can be important. Implicit in our notion of intelligence and related mental concepts is the presupposition that we are talking about flexible systems with complex, coexisting, persistent, asynchronously interacting sub-systems with different sorts of capabilities that can be combined and recombined in different ways to deal with novel situations. (This point is developed further in [21], [22], [25], [26], [27].)

We can now return to the revised (milder) Strong AI thesis, **T2**, which claims that enaction of a single algorithm could suffice for production of mental states, or, more precisely, that some causally embedded, program-controlled, temporally ordered, sequence of states would suffice. A variant of the thesis might even claim that a human mind itself can be thought of as going through such a well defined succession of states, like a computer executing an algorithm. (H.A. Simon [19], for example, has sometimes suggested that human intelligence is based on serial processing.) Penrose, like Searle [17], was certainly attacking this version of the AI thesis, in addition to the more absurd versions. Is it worth defending? Not if a mind requires more than ‘enaction of a single algorithm.’

## **9. Towards requirements for a mind**

Human mental life is much richer than a succession of momentary states each following its predecessor according to some rules. A human mind has many enduring interacting states with different histories and different durations, some remaining static while others change. (Minsky makes similar points in [11].) These processes have different functions, such as detecting information, interpreting it, storing it, reasoning, generating and analysing motives, forming

plans, controlling actions, monitoring actions, learning, and many more, to do with feelings and emotions. The coexisting perceptual states, beliefs, desires, intentions, plans, attitudes, moods, emotions, sensations, and other states of which we are unaware, interact with one another concurrently and asynchronously. There are also many sensory transducers constantly reacting to aspects of the environment (including the agent's body) and a host of processes analysing and interpreting the information they provide. Visual perception alone requires simultaneous processing, in real time, of many diverse locations in the visual field (some requirements for visual perception are analysed in Sloman [25]). Intermediate level perceptual processes that buffer some of their interpretations of sensory data may underly the experience of sensory contents referred to by some philosophers by the term "qualia" (often supposed by unimaginative philosophers to be resistant to computational explanation), while coexisting higher level processes correspond to perceptual judgements. All this concurrency is justifiable from the engineering standpoint, which views a mind as a sophisticated control system for a very complex and fragile mechanism in a fast moving, rich and potentially dangerous environment.

In addition to all these concurrent processes any human mind also has a rich collection of enduring, hierarchically organised dispositional states that are capable of influencing processes of many kinds, but need not actually do so at any one time, and which may or may not be known to the person concerned. For example, few people know much about their own grasp of phonetics, and most lack full knowledge about their own attitudes and personalities. Most of these internal states manifest themselves only very indirectly in particular thoughts, experiences, decisions and actions, and their causal powers endure even when not activated. (For more on all this see Sloman [27]).

Replicating human mentality therefore requires the design of an architecture that is capable of supporting all these coexisting states and processes with appropriate causal relations between them. The description is clearly reminiscent of the description of a general purpose multi-user time-shared computing system. It is very misleading to describe either a mind or such a computing system as merely composed of a *single* sequence of states ordered in time, as implied by talk of one algorithm being in control.

The notion of "algorithm" could be re-defined so that all the processes in one system constitute one algorithm, but that makes the thesis that one algorithm suffices trivial and uninformative. It is true that in a uniprocessor computer, time-shared or not, there is something like one algorithm, at the microcode level, but that algorithm is unchanged no matter whether the machine is running AI programs or just doing number-crunching. That algorithm is merely concerned with getting instructions from memory and executing them. Knowing it gives us no insight into the very diverse high level user processes and system processes that it implements. The existence of one algorithm at one level of description may be compatible with a host of unrelated algorithms interacting at another level of description. So it cannot explain the *particular* properties of the application programs (including the operating system) that happen to be running at that time. For example, it could obscure important engineering design features such as the monitoring of one process by another to increase robustness. Treating all this as one algorithm would be partly analogous to trying to explain how a computer works by saying that it

uses only matter composed of carbon, copper, iron, silicon, etc. Although true, this would explain nothing about the specific properties of the computer that distinguish it from other physical systems. Similarly, saying that everything that a brain does can be implemented by executing a single low level algorithm, even if true, would not necessarily explain any interesting properties of brains that distinguish them from other computing systems, the vast majority of which are totally unintelligent. All the features of an intelligent computing system that explain its mental capabilities may be independent of the existence of that one low level microcoded algorithm, since the system could be implemented in many other ways. Moreover the low level algorithm does not suffice for intelligence.

It might be replied that even though an intelligent agent requires many independent interacting states and processes, at least when described at a high level, nevertheless, all those processes could be implemented in a single low level process generated by a single algorithm, and this would be the UAI. Whether any one algorithm generating a low level sequence of states would suffice will depend on whether that sequence implements the right sorts of higher level interacting processes with the right causal powers, such as the ability of one process to modify another. We have already seen that a giant lookup table might implement an algorithm in the sense of producing the same mappings from input to output, without doing so in the required manner. We have also seen that truly parallel implementations can have different causal powers from serial implementations of simulated parallelism. So it is an open question whether a serial UAI could preserve the important properties of the original design. (If it cannot, that's no objection to AI as a research enterprise. It would be a *discovery* of AI: AI has no commitment to a UAI.)

To summarise: The causal powers of the architecture required for a mind imply many counterfactual conditional statements about 'what would happen if ...' at different levels of abstraction. For example, if X has an enduring attitude of prejudice against people of type P, this implies many statements about decisions that X would have made about another individual Y, if X had thought that Y was of type P and had thought that Y wanted something. This attitude towards people of type P might itself be changed if X had certain new experiences involving them. So descriptions of any one mental state imply a complex set of counterfactual conditionals about other mental states. The truth of some counterfactuals concerning relatively "dormant" states persists during interactions between other states, e.g. truths about your long term ambitions persist during normal perception and actions that have nothing to do with those ambitions. The persisting causal powers need not manifest themselves in any way over a time interval. This is somewhat like the persistence of properties of an operating system that guard against violations of access restrictions, or wait for interrupt signals. Similarly, X's grammatical knowledge, perceptual abilities, problem-solving skills, etc. can all persist while other things are going on. Further investigation is needed to show whether or not such a system can be implemented properly, with the right causal powers and 'subjunctive' properties, in a single serial process. But even if it cannot, that does not undermine the AI research program. Many of "the AI people", especially those who have tried to build working robots with visual sensors and controllable motors, wouldn't dream of trying to make the whole system simply go step by step through any

one algorithm, except in the trivial sense mentioned above: when processes are time-shared on one processor, there may be a single “fetch-execute” algorithm. This suggests that discussions of what any one algorithm might or might not be able to do has little relevance to the objectives of AI. The UAI is a red herring!

### **10. Is one processor enough?**

Even if it is agreed that the UAI is a myth and there is no one algorithm whose execution could suffice for the production of anything remotely like human mental states, that still leaves open the question whether all the different processes required could run on a single time-shared processor. Could a suitable multi-processing architecture be implemented in a collection of interacting virtual machines supported by a single physical machine? A positive answer would take the form of an even milder version of the Strong AI thesis than **T2**. The new version, **T3**, states that instead of a single (as yet unknown) algorithm there is some design involving multiple interacting computational processes, possibly involving many distinct algorithms, such that any instance of that design, with the right causal powers, including a time-sharing implementation on a single machine, would have mental processes. A full assessment of **T3** will require further research to specify exactly what sorts of causal interactions are necessary between internal states and processes in an intelligent system. We could then ask whether causal interactions between virtual states and processes on a single processor could satisfy these requirements or not. The analysis has not yet been done.

I’ve already shown (section 9) that simulated parallelism can have different causal powers from the real thing. The intervention of the processor switching contexts can disrupt the causal relations between coexisting states, by making the links too indirect. But this does not prove that such simulated parallelism could not produce mental processes. Some philosophers would argue that a uniprocessor implementation will not do because they believe that it is impossible for causal relationships to hold between “supervenient” states and processes, like the relations between states of a high level virtual machine. But this would imply that many of the things we currently regard as causal connections are not really so, because they are really supervenient virtual processes implemented in low level mechanisms studied in advanced physics. For example pressing a button would not really cause a light to go on: the “real” causes would not involve buttons and lights, but something far more esoteric. However, this philosophical position implies that there is a unique “bottom level” layer of reality at which “real” causal relations hold. I see no reason to believe this. (Taylor [29] explains how causal relations can hold in many different sorts of domains.) So, I do not believe there is any general philosophical argument about the nature of causation, that rules out a uniprocessor implementation of the multi-processing architecture needed for intelligence. There may be engineering objections, however, concerned with reliability and the need for asynchronicity, as discussed above. If so we’ll need to retreat yet another step, to a further weakened thesis: **T4** states that there is a collection of computational processes such that if they run on some distributed collection of processors they will produce mental states. (This formulation leaves open the question whether a uni-processor implementation could suffice.)

Of course, if we concern ourselves only with input-output behaviour, there can be nothing special about a multi-processor implementation. The input-output behaviour produced over any finite time period by any system composed of a network of interacting computers will be finite and can therefore be produced by a single processor running one program. In fact, any given sequence can be produced in infinitely many ways. But not all such implementations will be correctly describable as having the same collection of causally interacting internal states, and so not all will be intelligent even if they look intelligent. (They may, of course, be intelligent in the same sense as a gadget can be described as “clever” if it instantiates a clever design, but not in the same sense as the designer is clever.) This is just another example of the point made in section 6 that how behaviour is produced is what makes it intelligent, not what the behaviour is. Of course, this means that we cannot simply be relying on observed behaviour when we attribute intelligence to humans and other animals. But that’s partly because we have vague intuitive theories about how they work, and partly because there’s reason to believe that no physical system can in fact produce their behaviour unless it employs the kind of architecture required for intelligence: for instance pre-computed lookup tables of the required size could not possibly fit into their brains. Even if true, this still leaves open the question about what is possible in principle, e.g. in another type of universe.

## **11. Open questions about architectural requirements for minds**

All this still leaves open which versions of the Strong AI thesis are worth defending. I have tried to undermine the notion that AI is essentially committed to uni-processor theories. I have not proved that real as opposed to simulated parallelism is required for intelligence: the discussion merely shows that there are open questions that cannot be answered until we have a precise and detailed theory about the causal powers required in the substates and processes that interact in an intelligent agent. The answer may be that nothing definite is “required”, but that different designs may be more or less like human minds in various respects. So I leave open the question whether a physical multi-processor mechanism is required for mental states like ours. So defending uni-processor versions of AI is premature, and attacking them is not an attack on AI in general.

Certainly a major goal of AI is to achieve a general understanding of the nature of various kinds of intelligent systems (human, animal and artificial) and to use this general knowledge both to help us understand the human mind and to help us solve various practical problems. This sort of goal does not *presuppose* that it is possible for a uniprocessor mind to exist, and it is just perverse to attribute such a presupposition to the whole AI research enterprise, even if some people wrongly assume it.

From this point of view, the characterisation of Strong-AI given by Penrose is a crude oversimplification, e.g. (page 17)

For any significant kind of mental activity of a human brain, the algorithm would have to be something vastly more complicated but, according to the strong-AI view, an algorithm nevertheless ... all mental qualities -- thinking, feeling, intelligence, understanding, consciousness -- are to be regarded, according to this view, merely as

aspects of this complicated functioning: that is to say, they are features merely of the algorithm being carried out by the brain.

I have never met any AI person who believes the brain carries out only one algorithm, though I suspect that many believe that it implements a complex multi-processor architecture supporting both fine-grained and coarse-grained parallelism, of the kind loosely sketched above. (This coarse-grained parallelism is not to be confused with the fine-grained parallelism of connectionism, though there could be connectionist implementations of course-grained parallelism, in which distinct sub-nets interact with one another.) And there may be some who believe that the multi-processing could be implemented on a single processor, as **T3** allows, though most of them have not done the analysis required to check that the required causal powers of mental states would be preserved.

Very little work in AI has so far attempted to identify a complete architecture that might suffice for an intelligent human-like system (though some over-enthusiastic people have prematurely described their AI programs as seeing, learning, understanding, planning, deciding, etc., a type of fallacy now being repeated by some members of the neural net community). For reasons concerned with the enormous difficulty of designing complete agents, most AI work so far has been concerned with tiny fragments of intelligent mechanisms required for simple well-defined tasks, like playing chess, proving theorems, interpreting images of blocks, taking in simple stories and answering simple questions; though there have been more ambitious robot projects, with limited success. Even those who do think about more general architectures (Minsky [11], Moravec [13], Sloman [21, 24]) either do so only in the context of a subset of human capabilities, usually a narrowly circumscribed set of cognitive abilities, or if they do attempt to survey a broad range they risk shallowness much of the time. My own explorations analysing architectural requirements for intelligent systems with human-like motive processing capabilities, reveal ways in which such requirements lead to mechanisms that are capable of getting into states that have many of the characteristics of human emotions. Even so, only a tiny fragment of the phenomenological richness of human emotions is accounted for.

Section 10 discussed engineering differences between true parallelism and simulated parallelism, which led to the weakened AI thesis **T4**. Even this may not be weak enough, for it could turn out that in order to model animal brains we need not only many asynchronous concurrent computations, but also additional non-computational mechanisms, e.g. chemical processes (for global control?). Such mechanisms do not fit neatly into existing concepts of computation. If these mechanisms are simply regarded as computational no matter what their nature, then the AI thesis risks becoming trivially true by definition. Accommodating the extra properties while avoiding triviality requires still further weakening of the Strong AI thesis, perhaps in the form of thesis **T5** which would be similar to **T4**, but would allow that the design of an intelligent agent requires a subset of components of a type that would not normally be described as computational. Unless **T5** can be further refined so as to say which computational and which non-computational components are required, it remains very vague, and so weak as to be almost uninteresting.

However, if it turns out that there are deep reasons why a variety of different sorts of low level mechanisms, including chemical mechanisms, are required for intelligent agents, discovery of those reasons would be an achievement of AI, not an objection to it. All scientific disciplines and all long term research programmes undergo evolution of their basic explanatory concepts.

## 12. Is there a division between things with and without minds?

I conjecture that there is a true, but mild, version of the Strong AI thesis, something like **T4**, which claims that a certain kind of architecture, as yet unknown, perhaps essentially composed of interacting computational components (perhaps alongside some non-computational mechanisms), would necessarily have mental states, that might be more or less like ours, depending on the architecture (just as some animals probably have mental states less like ours, some more like ours). At this stage it is not clear how many different kinds of interacting components and what sorts of interactions are required, though it is clear that what is needed is a very complex and changing architecture supporting processes simultaneously serving many different purposes, concerned with perception, memory, motivation, affect, reasoning, planning, controlling actions, and different sorts of learning. That's a very different view of AI from one that looks for an *algorithm* of sufficient richness to produce mental states. The right functional architecture may include a changing collection of very varied algorithms performing many different tasks concurrently.

Even this mild AI thesis would be too naive if it stated that there is some definite dividing line between things with and things without minds, or between things with and things without consciousness. That assumes that our ordinary concepts like "consciousness" have sufficient generality and precision to enable us cleanly to divide the space of possible mechanisms (including all those not yet conceived of) into some with and some without. What is more likely is that exploring alternative designs will reveal many inadequacies in present day concepts, and as our theories about possible systems become more general and more precise we'll come up with new, and better, concepts for classifying the capabilities of different organisms and machines, just as theories about the structure of matter generated new improved concepts for classifying kinds of stuff. The question "Which things do and which do not have consciousness?" will then be replaced by a much larger collection of questions about which organisms have which combinations of (precisely defined) capabilities, and which of them can be replicated by various sorts of machines. Similarly, nobody would now want to divide all complex substances into mixtures and compounds: we have a much richer system of categories. Note that I am not claiming that there is a continuum of cases: rather there are many discontinuities in design space, still waiting to be discovered and analysed.

I am not saying that there's any magic underlying intelligence, or that a mechanistic explanation of the human mind is impossible; nor am I suggesting, like Searle [17], that digital computers cannot be used to replicate mentality. The point is more subtle: there are good engineering reasons why no *one* computational process, consisting of the execution of *a single* algorithm on a single machine can in principle have the properties required to generate and explain the host of co-existing processes and persistent counterfactuals required in an intelligent

control mechanism like the human mind. For a physically embedded agent this is partly because interaction with the environment requires multiple asynchronous transducers, as already explained. But there is also the deeper point that even a disembodied or disconnected intelligence, concerned only with exploring mathematical structures, would need not just one process but many interacting processes, to produce persistent internal states (beliefs, desires, skills, plans, etc.) with the right causal powers. The wrong sort of design could produce the same “trace” of behaviour and even report the same mathematical discoveries, but without using any intelligence. It’s not just what is done that matters, but how it is done.

Nothing said so far rules out the possibility that the architecture required for human-like mental processes might be embedded in a very complex network of computers, or even very fast Turing machines, interacting asynchronously with one another and with the environment, as suggested in thesis **T4**. A set of such interacting computers could be modelled on a single Turing machine if they were all driven by a single digital clock (however small the clocking frequency), but not if the time intervals between events on different machines vary continuously and the time intervals are significant, e.g. in controlling behaviour of some physical mechanism. This topic will be resumed later. Meanwhile let us attend to the nature of consciousness.

### **13. Is the nature of consciousness self-evident?**

Before we can begin to discuss the truth or falsity of any particular thesis about the architecture and sub-mechanisms that may be required for mental states, we need a much clearer idea of what we mean by “mental” states, and whether there are different kinds that need different sorts of architectures. Penrose thinks that we know what we mean by “consciousness”, and, moreover, that it refers to some thing or entity “that is, on the one hand, evoked by the material world, and, on the other, can influence it” (page 405). Philosophers use the phrase “reification fallacy” to label the assumption that a well understood noun or noun phrase necessarily refers to some *thing*. Many people fall into the fallacy over words and phrases referring to mental phenomena, e.g. “imagination”, “emotion”, “intelligence” and “consciousness”. If consciousness were a thing (like the appendix, or the ability to see) then we could ask why it evolved, or what “selective advantage” it confers (page 405), or whether its operation could be explained by quantum mechanisms (see page 399). The problem is that there is every reason to believe that there is no such unique thing, and that the concept of “consciousness” is full of muddle and confusion.

Dreams provide one illustration of the incoherence of the concept: does a person who experiences fear or pain in a dream have consciousness or not? He surely must be conscious, for how can there be experiences without consciousness? But surely the person is asleep, and therefore unconscious, i.e. lacking consciousness? Are we conscious when acting under hypnosis? Many animals can be asleep (unconscious) and then wake up (regain consciousness). Are we to assume that in the latter state they all have the same property of consciousness? I believe these and other intrinsically unanswerable questions can be used to show that the ordinary concept of consciousness, far from providing the basis of any rigorous argument about the nature of mind, is actually incoherent! Compare Dennett’s attack (in [2]) on the concept of

“pain”.

Penrose alludes briefly to such puzzles (page 406) then moves on to suggest that he can identify well enough what he is talking about by relying on “our subjective impressions and intuitive common sense as to what the term means and when this property of consciousness is likely to be present” (page 406). This is naively optimistic. He is saying, in effect, “You all know what mental states are because you’ve got them.” People have vast amounts of grammatical knowledge without knowing anything about linguistic theory. It’s just a myth that we have direct knowledge of the nature or contents of our minds.

Many people feel, like Penrose, that the nature of mental states is somehow self evident to those who have them. But this is just an illusion. The illusion probably arises out of the fact that, for good biological (or engineering) reasons our brains include (limited) self-monitoring mechanisms, which give us *some* information about our internal states and processes (just as modern computer operating systems have limited self-monitoring capabilities). But this internal perception was not designed to give us full and detailed information of a kind needed for scientific explanatory purposes, any more than our eyes give us full and detailed information about the constitution of material objects in the environment. Perceptual mechanisms, whether internal or external, evolved to serve limited practical needs. They can simplify or even distort reality, so long as they serve those needs. (Exactly what purposes are served by our self-awareness is still not clear. They probably include: being able to inform others about our states, some high level control functions, and perhaps certain kinds of learning through explicit self-modification. None of this requires perfect self-awareness.)

Penrose assumes that consciousness is also intimately involved in the ability to form judgements as to truth or falsity. This is extremely unclear, but if it has any content I believe it to be wrong, for several reasons, including the fact that our perceptual systems are able to take in complex and ambiguous information and produce correct judgements about what is out there many of which never reach consciousness. For example unconscious visual mechanisms are involved in posture control and the detailed guidance of hand movements, and unconscious mechanisms driven by auditory input produce decisions about morphological and syntactic ambiguities in speech, of which we are totally unaware (until we study linguistics). Attempts to analyse and model these processes suggests that they have an internal richness that is partly analogous to explicit, conscious, reasoning, including formulating hypotheses and testing them: as for example when perceptual cues are ambiguous and suggest alternative hypotheses which are unconsciously tested and pruned (perhaps using competitive processes in neural nets). If this is so, then judgements of truth and falsity do not require consciousness, unless the claim is restricted, trivially, to *conscious* judgements of truth and falsity. Penrose tries to avoid this circularity by restricting his claim to a subset of judgements: (p. 411) “Somehow, consciousness is needed in order to handle situations where we have to form new judgements, and where the rules have not been laid down beforehand.” It is not at all clear what this means. There are many animals that can learn to solve new problems. Human and animal visual systems constantly cope extraordinarily well with all sorts of novel configurations of objects producing novel retinal images. Are these all cases where the rules have been “laid down beforehand”?

Work on computer vision suggests, on the contrary, that seeing inherently involves some problem-solving, whether we are conscious of it or not.

Using unconvincing analogies between the alleged “oneness” of consciousness and the superposition of many quantum states (page 399), does not help to make a muddled concept any less muddled. When we understand the kind of architecture (i.e. division into separate interacting mechanisms) required for human mental capabilities, I suspect we’ll discover that there are *very many* sub-mechanisms that are concerned with different kinds of internal monitoring and control, and that we have only a dim and confused awareness of some of this internal richness that leads people to think they know what they mean by “consciousness”. In the long run, instead of being explained by new quantum mechanisms, as Penrose suggests, this incoherent concept will follow the same path to obsolescence as the concept of a continuously enduring point of space: Here too it is easy to fool yourself into thinking you know what you are talking about, simply by attending to it. This leads to a pre-relativistic model of space. The notion of a point of space as an indivisible entity with intrinsic identity that is preserved indefinitely has evolved into the conception of a point of space as a family of spatial relationships some of which may change while others are preserved. Similarly, early concepts of kinds of stuff evolved into a far richer theory of the varieties of chemical elements and compounds. Concepts of mental states and processes will also have to evolve if they are to be used in deep explanatory theories.

#### **14. Does consciousness entail the ability to understand Gödel?**

We come now to the core argument in the book. Penrose not only insists that consciousness is necessary for the formation of judgements of truth and falsity but even argues that super-Turing capabilities are presupposed. He bases this on the ability of human mathematicians to understand Gödel’s proof. This is a very odd argument from the point of view of anyone who believes (like Penrose) that mental states and processes exist in many non-human animals who are quite incapable of such abstract mathematical reasoning, to say nothing of human infants and the vast majority of human beings who, alas, appear to be unable to comprehend meta-mathematical arguments. His defence is that there is no other way to make his attack on Strong AI mathematically watertight.

Surely any argument about necessary requirements for consciousness should focus on capabilities that are common to a wide range of agents, and not just the capabilities of the select few. But nowhere does Penrose give even a hint of an argument that the consciousness of cats or chimpanzees or human toddlers requires super-Turing capabilities. So the very most he can claim to have shown is that mathematicians cannot be computers, or cannot be modelled on computers. No doubt there are mathematicians who would like to believe that they have a higher form of mentality than the rest of mankind, and who might relish the thought that this is because their brains use some kind of ill-understood quantum capability to give them super-Turing powers. But they can’t expect the rest of us to take this seriously.

Despite all this, it is worth looking at the argument, because it poses long term questions for AI, concerned with how it is possible to think about infinite sets. Penrose claims that Gödel's incompleteness theorem shows that we do this using non-computational mechanisms. I'll argue instead that the theorem shows that our thinking about infinite sets is radically indeterminate. So we cannot "see" the truths Penrose thinks we can see, and there is no need to invoke quantum phenomena to explain how we see them!

### 15. Are mathematicians super-Turing machines?

What exactly happens when a mathematician understands and accepts Gödel's proof? Penrose purports to show that some mechanism more powerful than a Turing machine is required to account for this thought process. I shall attempt to expose and challenge assumptions that most critics appear to have missed. His arguments actually take several different forms. I shall concentrate on what seem to be the core ideas.

For any formal system  $F$  rich enough to express the arithmetic of natural numbers, there is a construction, using Gödel-numbering, of an arithmetical formula  $P_k(w)$ , as follows.  $P_k$  is an arithmetical predicate on integers, recursively defined so that it is true of the integer  $w$  if and only if for no integer  $N$  is the  $N$ 'th possible proof in  $F$  a proof of the formula for which  $w$  is the Gödel number. (Notice that this involves quantifying over infinite sets of formulas and numbers.) This predicate is used to construct a formula  $P_k(k)$  where  $k$  is the Gödel number of the formula  $P_k(k)$ . The formula thus constructed from  $F$  will henceforth be referred to as  $G(F)$ . (The actual construction is too intricate to be described here. For more details see Penrose p. 105-8, Nagel and Newman [14], or Hofstadter [6]. The latter includes much discussion relevant to this paper). Gödel demonstrated that if  $F$  is consistent (strictly, omega-consistent) there can be no derivation in  $F$  of  $G(F)$  or of its negation, so that  $G(F)$  is undecidable, and  $F$  is incomplete. I shall try to present the rest of Penrose's argument in as strong a form as I can, paraphrasing freely.

The argument is roughly this: What  $G(F)$  *asserts* (or appears to assert - see below), is that  $G(F)$  is not provable in  $F$ . Therefore what it asserts must be *true* if  $F$  is consistent, since Gödel proved just this. Insofar as  $F$  is little more than a specification of the properties of the natural number series and operations on numbers, mathematicians can see that it is consistent (and omega-consistent?) because the natural numbers provide a model for it. So what  $G(F)$  asserts must be unconditionally true. So Penrose can apparently see something to be true which cannot be derived in  $F$  even if  $F$  is meant to be the formal system defining how Penrose thinks about arithmetic! Suppose  $F$  is a more general system supposed to specify how Penrose thinks (a version of the UAI). Then the same sort of argument applies: if  $F$  really describes Penrose's thinking then  $F$  describes something that exists and must be consistent, and therefore the corresponding Gödel formula must be neither provable nor refutable in  $F$ , and, since that is what the formula says, it must be true. Penrose can see this truth, even though it is not derivable in  $F$ . So no formal system like  $F$  can define how he works, and there is no algorithmic explanation of his thinking. He exults on page 108: "Somehow we have managed to *see* that  $P_k(k)$  is true despite the fact that it is not formally provable within the system."

Of course, given any  $F$ , with its corresponding  $G(F)$ , it is possible to extend it with additional axioms to produce a new system  $F'$ , in which  $G(F)$  will then be derivable. But exactly the same construction can be used for  $F'$  to produce a new formula  $G(F')$  which is not provable in  $F'$ , and which says that it is not provable in  $F'$ , which Penrose can see is true; and this can be continued indefinitely. After discussing this, he says, on page 110:

We *see* the validity of the Gödel proposition  $P_k(k)$  though we cannot derive it from the axioms. The type of 'seeing' that is involved in a reflection principle requires a mathematical insight that is not the result of the purely algorithmic operations that could be coded into some mathematical formal system.

From all this he concludes (p. 417):

If the workings of the mathematician's mind are entirely algorithmic, then the algorithm (or formal system) that he actually uses to form his judgements is not capable of dealing with the proposition  $P_k(k)$  constructed from his personal algorithm. Nevertheless, *we* can (in principle) see that  $P_k(k)$  is actually *true!* This would seem to provide *him* with a contradiction, since *he* ought to be able to see that also. Perhaps this indicates that the mathematician was *not* using an algorithm at all.

There are several standard objections made by critics of Penrose (and Lucas, whose arguments are similar), all found in the 1990 BBS commentaries [15]. One is that Penrose hasn't seen that  $G(F)$  is true, because proving this requires proving that  $F$  is consistent. I don't fully understand Penrose's answer, but he could reply that  $F$  is clearly consistent because it has the natural number sequence as a model, a fact that we can somehow directly establish. Another common objection is to agree that  $G(F)$  is seen to be true, but only by using fallible and incomplete procedures such as would enable a suitably designed artificial mathematician also to see it to be true, in the same way as we do. Penrose could object that there's nothing fallible in the derivation that enables us to see that  $G(F)$  is true. Unlike these critics, I'll argue below that we don't see that it is true.

Another popular reply says that even if  $G(F)$  cannot be proved in  $F$ , there is some more encompassing formal system, meta- $F$ , in which the formula is provable. I.e. we may be able to prove something like

Provable( $G(F)$ , meta- $F$ )

But even if this is so it cannot account for the (alleged) discovery that  $G(F)$  is true, because truth is a semantic property, and establishing provability in meta- $F$  merely establishes a new syntactic property of  $G(F)$ , namely that it is derivable from the axioms and rules of meta- $F$ . We still need some way of knowing that the axioms of meta- $F$  are true. Penrose comments on page 108: "The way that a strict formalist might try to get around this would perhaps be not to talk about the concept of truth at all, but merely to refer to *provability* within some fixed formal system." That doesn't explain how a mathematician sees that something is true. The proof in the new formal system merely demonstrates the syntactic derivability of the formula in the new system (trivially, if it was added as an axiom to create that system). The proof says nothing about how it is

possible for us to see that the original formula has the semantic property of truth, or more precisely that  $G(F)$  has a semantic relationship with  $F$ , i.e. saying something true about what is not derivable in  $F$ . This sort of semantic property cannot be reduced to any syntactic property of a formula in any enriched formal system. Most of the tempting objections to Penrose's use of Gödel's theorem fail to address this argument, because they fail to distinguish the syntactic property of derivability from the semantic property of truth.

## 16. Has the Gödel sentence been proved to be true?

When I first learnt about the theorem I too thought I had grasped something that could not be formalized. I shall try to explain why, like Penrose and many others, I was wrong. The conclusion that  $G(F)$  is true feels very compelling when one has been through all the steps of Gödel's argument. This depends crucially on the impression that there is a clear proposition expressed by the formula  $G(F)$ , and that that proposition can be expressed in English as something like "this formula is not provable or refutable in  $F$ ." In part this depends on the fact that besides the *syntactic* properties of well-formedness, derivability etc, there are *semantic* properties of reference, predication, truth and falsity. I am not disputing this. I am challenging only the suggestion that the semantic properties of  $F$  somehow uniquely determine which (infinite) model is being talked about. Unless they do, they cannot uniquely determine the truth-values of all the formulas expressible in  $F$ . My argument that  $G(F)$  does not have the meaning it is commonly taken to have depends on the fact that because neither  $G(F)$  nor its negation can be derived in  $F$ ,  $F$  will have some models in which  $G(F)$  is true and some in which it is false (the latter sometimes called "non-standard" models). Therefore  $F$  can be "seen" to be true only if there is some means, other than  $F$ , of specifying which model is in question.

$G(F)$  certainly has a meaning that can be expressed in English by saying that a certain very large number  $k$  has a very complex arithmetical property expressed by the predicate  $P_k(k)$ . This assertion could be true or could be false in relation to any particular number. Gödel proves that whichever it is it is not derivable in  $F$  if  $F$  is consistent. But why are people so convinced that what it says is *true*? This conviction depends crucially on the mapping that is set up by Gödel's numbering, which tempts us to say that "k" denotes not the number  $k$  but the corresponding formula, and that the predicate  $P_k$  expresses not just a property of numbers but a syntactic property of formulas in  $F$ , i.e. the property of not being derivable or refutable in  $F$ . So we are led to believe that  $P_k(k)$  asserts that the formula corresponding to the number  $k$ , i.e.  $P_k(k)$  (i.e.  $G(F)$ ), is not decidable in  $F$ . Then having learnt that the formula has the *syntactic* property of not being decidable in  $F$  (if  $F$  is consistent) we are tempted to say: "but that is what the formula asserts, therefore what it asserts is TRUE (if  $F$  is consistent)." I.e. we infer the *semantic* property of being true. Several critics of Penrose in [15] argued that  $G(F)$  has not been proved to be true because  $F$  has not been proved to be consistent. But I think that is missing the point. Even if we had a proof that the system  $F$  in question is consistent (or omega-consistent) there would still be a deep flaw in the argument. This is because  $G(F)$  does not assert what it seems to assert.

There are two crucial flaws in the argument put forward by Penrose (and others before him). The deeper flaw, (a) below, is ignored by most of his critics:

- (a) Gödel's argument makes people think they have grasped some semantic relation and had some insight into the truth of a certain proposition, about provability of a formula, but this is an illusion.  $k$  is, after all, just a numeral: it denotes a number, not a formula. Similarly,  $P_k$  is a complex arithmetical predicate about numbers, not a predicate concerned with derivability of formulas in  $F$ . More importantly, because Gödel proved that  $G(F)$  is neither refutable nor derivable in  $F$ , it follows that a consistent system is obtained by adding  $G(F)$  to  $F$  and also by adding its negation to  $F$ . So there will be models of  $F$  in which  $G(F)$  is true and models in which its negation is true. So Penrose can't have "seen" that it *must* be true. It also follows that the derivation of  $G(F)$  from  $F$  does not have the certainty that Penrose claims in [15], where he shifts from talk about "seeing the truth" of  $G(F)$  to saying (p. 694):

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.

He apparently has not noticed that the theorem implies that there are some models of  $F$  in which  $G(F)$  is false.

What has been missed is that  $G(F)$  does not express some definite true proposition about formulas in  $F$ : it is merely an assertion about numbers, an assertion that has not been proved. (Perhaps it will turn out in some of the "non-standard" models that make  $G(F)$  false, that the assumed mapping between complex arithmetical expressions and metalinguistic statements about  $F$  goes awry?)

- (b) It remains conceivable that if AI research ever creates an autonomous intelligent agent, it will be tempted by exactly the same mistaken thoughts when it encounters Gödel's proof. In fact it is very likely that any intelligent system will be misled in the same way because it too will have a strong tendency to confuse structural mappings with semantic relations, since structural mappings are very often used as a basis for semantic relations.

Some people will be tempted to argue against (a) that the models that make  $G(F)$  false are irrelevant, because  $G(F)$  is true in the "intended" model of  $F$ , the model based on our intuitive grasp of the natural number series, containing 0 and all its successors. But how are we supposed to grasp which model we are talking about? How can we unambiguously identify this infinite set either for ourselves or for purposes of communication with others? The answer cannot be that we identify it using a formal axiom system, because Gödel's theorem shows us that no such system will uniquely identify any model: it will always be incomplete, like  $F$ . If the axiom system is rich enough to include arithmetic there will always be models in which  $G(F)$  turns out false and other counter-intuitive results appear. If we are able somehow to specify precisely which infinite set we mean, but only by using a non-formal method, i.e. something that cannot be expressed as an axiom system or an algorithm that uniquely generates just the required set of integers, then Penrose has won this particular argument. (I return to this possibility later.)

Perhaps the feeling that we can completely and unambiguously identify an infinite mathematical set is just wrong: however much we think it is determinate, our conception of the set will always include indeterminate aspects and will never distinguish “standard” from “non-standard” models. And if *we* can’t do this, then perhaps there is no need for artificial intelligent agents to do so either? If that is correct, there is no need for an AI theory to explain the non-existent capabilities, and no need for Super-Turing quantum mechanisms.

We have touched on several questions in the area of overlap between AI and philosophy: What is the semantic property of being true? What is involved in grasping that a formula expresses a proposition with certain truth conditions? What is involved in “seeing” that such a property is true? These are hard questions, and so far I don’t think ANY work in AI has said anything useful about them. AI, to my mind, hasn’t yet even given us a convincing theory of what it is for a machine to understand “Block A is on Block B” as humans do, never mind understanding sophisticated meta-mathematical statements like Gödel’s formula. Understanding should be related to states like believing, desiring, imagining, perceiving, etc., though it is not exactly a precisely defined concept [22]. Without a good theory of what it is to understand symbols and make judgements of truth and falsity in general, it may be premature to argue about Gödel’s formula. Nevertheless it is worth digging a little deeper.

### **17. How can we think about infinite sets?**

There remains the question whether we have a non-logical, non-formal, way of specifying an infinite series? It may be that we do have ways of representing information that cannot be modelled with full precision on Turing machines or equivalent computational systems. Various authors (e.g. Sloman [20], Funt [4]) have suggested that there are methods of representing and manipulating information using pictures, maps, models and other formalisms that are distinct from “Fregean” or “applicative” formalisms to which the limit theorems of logic and computer science apply. Is there some way of representing the notion of the infinite set of integers that is different from the use of a formal system of the kind considered by Gödel? Could we use a physical mechanism for generating numerals, i.e. representations of numbers, indefinitely? Is it possible that there is some way of perceiving properties of such a concrete numeral-generating *mechanism* that is different from a formal derivation or a digital computation? Some mathematicians seem to think that the mental operation of “adding 1” is the basis of the grasp of the whole infinite series of natural numbers. Could this be based on something like perception of some kind of iterative mechanism and its properties? Perhaps such formalisms and mechanisms would have to play a role in the thought processes of a robot mathematician with human capabilities.

Until it is demonstrated that we do have some way of completely specifying exactly which infinite set we are talking about as a model of  $F$  then it is not the case that we can claim to have seen that  $G(F)$  is true: for it will actually be false in some models of  $F$  and we have no basis for saying that our grasp of the ‘intended’ model rules this out. At present it is totally unclear what such a method of determining the ‘intended’ model could be like, except that Gödel’s theorem shows that no axiom system or algorithmic method suffices, as Penrose correctly points out.

If the argument (a), above, is incorrect, and there really is a way of seeing that  $G(F)$  is true despite its formal undecidability, then perhaps we have to accept Penrose's conclusion that there is something mathematicians do that does not correspond simply to deriving formulas in a formal system, and cannot be modelled by any algorithm. That should not be too surprising if the arguments given above against the UAI hypothesis are accepted. Moreover, if (a) is incorrect, that does not constitute an objection to AI as a research programme seeking computational techniques for producing intelligent agents. In particular, the difference between the semantic property of being true and the syntactic property of being derivable in some formal system is obvious without the paraphernalia of Gödel's theorem. Even if some mild version of the AI thesis is correct in claiming that all mental states and processes depend ultimately on low level processes that can be implemented as syntactic operations on structures, Penrose is still right in saying that the phenomenon of mathematical insight, like the other mental phenomena discussed above, needs something more than an algorithm for syntactic operations: at higher levels of description we need a functional architecture supporting a highly differentiated collection of interacting mental states and processes, as explained above. These causal interactions between syntactic processes are not themselves syntactic processes. (For more on requirements for semantic capabilities see [22], [22a]. Rapaport [16] argues that syntactic processing is enough for the production of mental processes, though perhaps that is because he does not acknowledge the importance of these different levels of explanation.)

## **18. Mathematical Platonism**

Penrose makes much of mathematical Platonism, claiming that certain mathematical entities, such as the natural number series and the Mandelbrot set, exist independently of us, and that we can somehow discover truths about them. His Platonism has exasperated some critics who regard it as metaphysical or mystical nonsense. For Penrose, however, it plays a crucial role in explaining how we discover facts like the truth of  $G(F)$ . He thinks that we have some kind of direct contact with these entities, which enables us to grasp statements about them. Some Platonists (e.g. Plato?) claim that if mathematical entities exist in a special non-physical realm then the discovery of mathematical truths must employ special spiritual mechanisms that enable such entities to be explored. This would pose a real threat to the long term aims of AI as a discipline committed to the use of mechanisms wholly embedded in the physical world. However, Penrose does not construe Platonism in this extreme anti-physicalist form. In particular, he does not believe that the brains of mathematicians depend on anything that is in principle beyond the reach of physics. All he is claiming is that mathematical truths and concepts exist independently of mathematicians, and that they are discovered not invented. This, I believe, deprives Platonism of any content, and certainly leaves it as no threat to AI.

Despite the effort Penrose puts into his defence of mathematical Platonism, and the strong counter-claims of others that it is a mystical, or anti-scientific doctrine, such disagreements are really empty. It makes not a whit of difference to anything whether the Mandelbrot set, or the natural number series, does or does not exist prior to our discovering them. The dispute, like so many in philosophy, depends on the mistaken assumption that there is a clearly defined concept

(in this case “existence of mathematical objects”) that can be used to formulate a question with a definite answer. We all know what it means to say that a unicorn (defined as a horse with a single horn) exists, and we know how to investigate whether that is true or false. Quite different procedures are involved in checking the equally intelligible question whether there exists a prime number between two given integers  $N_1$  and  $N_2$ . But there is no reason to assume that any clear content is expressed by the question whether *all* the integers do or do not “really” exist, or exist independently of whether we study them or not. For example, this cannot make any difference to the design requirements for mathematical intelligence.

The practice of mathematics, the process of exploration and conjecture, the nature of proof, the devastating effect of counter-examples, would all be the same no matter whether entities exist in advance of discovery or not. Intuitionists have argued that because mathematical objects have no independent existence certain methods of proof, e.g. those using

$$\sim \sim p \rightarrow p$$

are not valid. But other mathematicians have happily gone on ignoring this stricture without any disastrous consequences. Mathematics is a subject in which different classes of things can be studied and different methods of reasoning can be explored. Once a method is well specified we can then find out what can and what cannot be done with it. Arguing that one is *right* and another *wrong* because certain things do or do not *exist* is pointless when the relevant notion of existence in question is so ill-defined. I conclude that the question whether Platonism is true is just one of those essentially empty philosophical questions that have an aura of profundity, like “Where exactly is the Universe?” or “How fast does time really flow?” An intelligent machine, like many intelligent human beings, may be tempted to misconstrue such questions as having significance, but they provide no basis for doubting the possibility of intelligent machines.

Penrose, unfortunately, not only believes that there is a real question whether mathematical objects exist independent of our thinking of them, but also sometimes adopts turns of phrase that do not appear to be consistent with his physicalist philosophy. He writes, on page 428: “I imagine that whenever the mind perceives a mathematical idea, it makes contact with Plato’s world of mathematical concepts”, and later on the same page “...communication [between mathematicians] is possible because each is directly in contact with the *same* externally existing Platonic world!” He seems to be quite insensitive to the fact that “makes contact” and “directly in contact” are very obscure metaphors in this sort of context. We know what contact between physical objects is, and what adjacency in the number series is, but what is contact between a mind and a number? These are empty words.

There is, however, a real problem here, which can be put in much more straightforward language by stating that many mathematicians and others claim to be able to think about and communicate about types of abstract objects that cannot easily be specified by giving examples, and which need some kind of indirect means of identification. One of the oldest examples is the infinite set of natural numbers, or even the infinite set of numerals denoting them. We can easily present examples of increasingly large subsets, but never the whole thing. Immanuel Kant [8] claimed many years ago that such infinite totalities can only be grasped via rules that generate them, and this is widely believed. But, as I’ve explained above, if the rules are specified in

something like a formal system, Gödel's incompleteness theorem shows that there is a problem about whether the "intended" set can be specified completely. Unless there is some important undiscovered mechanism for such thinking, all those mathematicians are fooling themselves, not about the particular theorems they prove, but about the nature of their understanding. If so that suggests that artificial intelligences may fall into the same trap. If *every* method of specifying infinite sets is equivalent to the use of a formal axiom system, then Gödel's theorem, far from proving the impossibility of artificial intelligence, is a pointer to some limitations of intelligence in general.

## 19. Recapitulation

I have shown that there are several "strong" AI theses, the strongest of which (**T1**) states that there is some form of computation, the UAI, all of whose instances will be mental processes, where computations are simply defined mathematically as ordered sets of structures satisfying the relationships specified in the algorithm. This extreme thesis, like the slightly revised version, **T1a**, requiring temporally ordered structures, turns out to be obviously false. A slightly milder Strong AI thesis **T2**, which states that there is some algorithm (the UAI) whose *execution* on a computer or a Turing machine would suffice for the production of mental states has been shown to be extremely unclear because there are different modes of execution some of which clearly cannot support mental processes, and no clear specification has been given of the intended modes of execution that rules out these and other bizarre cases including "replays" of traces, processes based on precomputed tables, and the Chinese room. Moreover, even if it turns out that we can define the required causal relation between program and process, thesis **T2** will still be misleading because (human-like) mental states and processes require not the execution of any one algorithm producing a particular succession of states in one machine, but a multi-(virtual)-machine architecture in which several enduring states with their own histories interact causally with each other and with the environment. So thesis **T3** claims that there is a collection of parallel, enduring, interacting states and processes that suffice to produce a mind, without assuming that these can be produced by a single algorithm. However, it does not rule out implementation as virtual parallel machines on a single time-shared computer. **T3** was implicitly defended in my 1978 book [21], and is probably the version of Strong AI that most thoughtful AI practitioners would agree with, though it has a number of conceptual problems to do with different kinds of causal interaction and control (discussed above and in Sloman [23]).

It is possible that further analysis will show that the required causal connections between sub-processes (including reliability of control) cannot be achieved on a single serial implementation of the multi-processing architecture. I argued above that an intelligent agent embedded in the physical world would need more than one processor, because of the need to cope with asynchronous interactions with the environment. Perhaps, as suggested above, all such interactions can be handled by relatively unintelligent processors that do little more than buffering of input and output signals between transducers and a time-shared central processor. Whether some collection of virtual parallel processes on the central processor could or could not suffice for the production of human-like mental states is a question that will remain unclear until

we have done more analysis of the functional requirements for different kinds of intelligent systems. This requires a systematic exploration of the space of possible designs for systems with abilities close to ours (Sloman [26]).

The mildest version of the Strong AI thesis, **T4**, would claim that some collection of programs running on a network of computers and other digital devices (or a collection of very fast Turing machines able to modify one another's tapes and be modified by the environment) would suffice for the production of familiar mental processes. I don't know if this is true, but I suspect it is (though, for reasons given above, the meaning of the question is still unclear because our concepts of mental processes are still unclear). Further research might show that even if it would be true in principle if there were no limits to digital processing speeds, actual physical limits imply that intelligent, mobile, robots cannot in practice be implemented in this world without the aid of some analogue physical or chemical mechanisms required for speed. For instance, we still don't know how to build machines that can cope with (representations of) complex shapes in real time, as squirrels, birds and monkeys appear to be able to do while moving quickly through tree-tops. Perhaps that's partly because *general* purpose computational mechanisms can never be fast enough, including Turing machines. This leads to consideration of **T5**, which allows that the design of an intelligent agent might require some components (e.g. chemical processes) of a type that would not normally be described as computational, even if the bulk of the implementation were computational. **T5** is clearly very weak and very vague.

Another possibility, **T6**, is that a complete intelligent agent with mental processes exactly like ours in all relevant respects could be implemented in virtual machines embedded in a suitably slowed down *computer simulation* of the total environment. This depends on the possibility of simulating the physical world on a computer, which cannot be done if the world includes non-Turing computable information as conjectured above. This is perhaps the mildest interesting version of the Strong AI thesis. It assumes that there are no essentially continuous processes at the basis of our interaction with the environment. If there are, then they cannot be approximated discretely if they are essentially chaotic (see Gleick [5] or Penrose pp. 173-183). If they are chaotic, then the discrepancies inherent in simulation models, however minute initially, can lead to arbitrarily large deviations between model and reality, in any given time interval. There seems to be plenty of evidence for chaotic mechanisms in the environment, which rules out the possibility of an accurate computer model of an agent in a simulated environment.

Even if it were granted that, because of the physical richness of the environment, actual physically embodied intelligent robots or accurate simulations of such robots require something more than computational mechanisms, so that all the above theses apart from **T5** are false, a further question remains. Is there a slightly more interesting true thesis **T7** stating that a certain important *subset* of mental functioning is implementable on one serial but time-shared computer? A slightly weaker version of this, **T8**, would claim that even if a uniprocessor implementation could never possess the causal powers required for any interesting subset of mental processes, nevertheless an implementation on a network of computers, could. Both **T7** and **T8** remain exceedingly vague until more is said about *which* subset of mental processes is in question. It could, for example, be the mental processes of a disembodied mathematician, with goals, desires,

thoughts, plans, etc. concerned only with mathematical structures, problems, and proofs. This kind of individual wouldn't see fields and forests, nor experience tingles or itches, thirst or lust, nor ever smile or wince, but it might be pleased at solving certain problems and extremely disappointed on discovering Gödel's incompleteness theorem and other limit theorems.

There may be other varieties of intelligence, that differ from human intelligence in interesting ways, waiting to be discovered as we explore the space of possible designs for minds.

The task remains of providing a more detailed specification of this family of theses, and investigating whether any of them is true. By comparison, the issue of how to get a machine to "see" that Gödel's formula  $G(F)$  says something true is clearly not very important or central to AI, or cognitive science, since most intelligent animals cannot do that, and I have argued that even sophisticated mathematicians are deluding themselves when they think they can. Still, a complete theory of intelligence would have to account for mathematical thinking, including the fact that certain kinds of intelligent mathematicians fall into this trap. This would include explaining how they are able to think about infinite sets and why they have a strong tendency to believe that their own thoughts about infinite sets are more semantically determinate than they really are. If it turned out that this capability depended on some kind of inner inspection of the properties of an iterative mechanism for generating numerals, that might support the "intuitionist" philosophy of mathematics. Similarly, a complete AI theory should explain why certain kinds of intelligences are strongly tempted to espouse, and others strongly tempted to refute, mathematical Platonism and other philosophical vacuities.

## **20. Conclusion**

Penrose has written a very stimulating book. As an introduction to various aspects of mathematics, physics and computer science, and as a presentation of the world-view of a distinguished scientist it merits attention, and there are many sections I am sure would repay re-reading. However, as a contribution to the profound questions in philosophy of mind about the nature of consciousness it is flawed both by the assumption that we really have a clear idea of what consciousness is, and a failure to distinguish obviously false extreme versions of the strong AI thesis from weaker more interesting milder versions. His attempt to use Gödel's theorem as a basis for criticising the long term AI research programme fails because it takes too seriously unjustified intuitions about what has been proved, though the theorem does point to some important unanswered questions about how intelligent mechanisms can think about infinite totalities. As a contribution to the study of the mechanisms of mind, his speculations about the relevance of quantum mechanics are totally unconvincing, and he seems to feel the need for them only because he has not explored some of the more sophisticated types of designs that might, one day, emerge from work in AI. In short, he has not (yet) seriously tried doing AI.

Nevertheless, in reporting on his experiences as a creative mathematician, I believe he has provided some empirical evidence concerning the nature of at least some real human minds (e.g. a mathematical emperor?), and the sorts of experiences they can have. Any mechanism that explains how human minds work must account for these phenomena as well as the more mundane phenomena patiently catalogued in the laboratories of psychologists, the notebooks of

philosophers, and the many poems, plays and novels that reflect the human condition.

## Acknowledgements

I am grateful for useful critical comments on earlier drafts by Alan Bundy, Dave Chalmers, Dan Dennett, Claudio Gutierrez, Steve Knight, Mark Madsen, Don Perlis, Robin Popplestone, Tim Read, Peter Ross, Nigel Seel, Alan Sexton, Ben Sloman, Luc Steels, and Richard Yee. The journal's review editors Mark Stefik and Steve Smoliar, were extraordinarily patient and helpful in providing constructive criticism, and allowing the author to have the last word!

## References

- [1] Alan Bundy *The Computer Modelling of Mathematical Reasoning*, London: Academic Press, 1983.
- [2] D.C. Dennett, *Brainstorms*, Bradford Books and Harvester Press, 1978.
- [3] Hubert L. Dreyfus, *What Computers Can't Do*, Harper and Row, (revised edition), 1979.
- [4] Brian V Funt, Problem-solving with diagrammatic representations, *Artificial Intelligence*, vol 13 no 3, 201-230, 1980 (reprinted in R.J. Brachman and H.J. Levesque (eds) *Readings in Knowledge Representation*, Morgan Kaufmann, 1985)
- [5] James Gleick, *Chaos*, Harmondsworth: Penguin Books 1988
- [6] Douglas R Hofstadter, *Godel, Escher, Bach: an Eternal Golden Braid*, Hassocks: The Harvester Press, 1979.
- [7] D. R. Hofstadter and D.C. Dennett, *The Mind's I: Fantasies and Reflections on Self and Soul*, (Penguin Books, London, 1981)
- [8] Immanuel Kant, *Critique of Pure Reason*, 1781. Translated by Norman Kemp Smith, London: Macmillan 1929.
- [9] J.R. Lucas, Minds, machines and Gödel, *Philosophy* 36, pp 112-27, 1961
- [10] Tim Maudlin Computation and consciousness, in *The Journal of Philosophy*, 1989, pp 407-432.
- [11] M.L. Minsky, *The Society of Mind*, London: William Heinemann Ltd, 1987.
- [12] James H. Moor, The pseudorealization fallacy and the Chinese room argument, in James H. Fetzer (ed) *Aspects of Artificial Intelligence*, pp 35-53, Kluwer Academic Publishers, 1988.
- [13] Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press (Cambridge, Mass; London, England), 1988.
- [14] E. Nagel and J.R. Newman *Gödel's Proof*, Routledge and Kegan Paul Ltd, 1958.
- [15] Roger Penrose, Precis of *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, in *The Behavioral and Brain Sciences*, 13,4 pp 643-705, 1990.
- [16] William J. Rapaport, Syntactic semantics: foundations of computational natural-language

understanding, in James H. Fetzer (ed) *Aspects of Artificial Intelligence*, pp 81-131, Kluwer Academic Publishers, 1988.

- [17] John R Searle, 'Minds Brains and Programs' in *The Behavioral and Brain Sciences*, 3,3, 1980.
- [18] John R Searle, *Minds Brains and Science*, (The Reith lectures) BBC Publications, 1984.
- [19] H.A. Simon, : 'Motivational and Emotional Controls of Cognition' 1967, reprinted in *Models of Thought*, Yale University Press, pp 29-38, 1979.
- [20] Aaron Sloman 'Interactions between Philosophy and A.I.', *Proceedings 2nd International Joint Conference on Artificial Intelligence*, London 1971. Reprinted in *Artificial Intelligence*, 1971, and in J.M. Nicholas (ed), *Images, Perception, and Knowledge* Dordrecht-Holland: Reidel 1977.
- [21] Aaron Sloman *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind*, Harvester Press, and Humanities Press, 1978.
- [22] Aaron Sloman, 'What enables a machine to understand?' in *Proceedings 9th International Joint Conference on AI*, pp 995-1001, Los Angeles, 1985. (Also University of Sussex Cognitive Science Research Paper 053)
- [22a] Aaron Sloman 'Reference without causal links' in *Proceedings 7th European Conference on Artificial Intelligence*, Brighton, July 1986. Re-printed as J.B.H. du Boulay, D.Hogg, L.Steels (eds) *Advances in Artificial Intelligence - II* North Holland, pp 369-381, 1987. (Also Sussex University Cognitive Science Research Paper 047)
- [23] Aaron Sloman, 'Did Searle attack strong strong or weak strong AI' in A.G. Cohn and J.R. Thomas (eds) *Artificial Intelligence and Its Applications*, John Wiley and Sons 1986.
- [24] Aaron Sloman, 'Motives Mechanisms Emotions' in *Emotion and Cognition* 1,3, pp 217-234 1987, reprinted in M.A. Boden (ed) *The Philosophy of Artificial Intelligence* "Oxford Readings in Philosophy" Series Oxford University Press, 1990.
- [25] Aaron Sloman 'On designing a visual system: Towards a Gibsonian computational model of vision' *Journal of Experimental and Theoretical AI* 1,4, 289-337 1989
- [26] Aaron Sloman 'Beyond Turing Equivalence' in P. Millican and A. Clark (eds) *Proceedings Turing90 Colloquium* to appear 1992, Oxford University Press.
- [27] Aaron Sloman Prolegomena to a theory of communication and affect in A. Ortony, J. Slack, and O. Stock, (eds.) *A.I. and Cognitive Science: Perspectives on Communication*. Heidelberg, Germany: Springer, forthcoming 1992. (Also available as Cognitive Science Research Paper No 194, University of Sussex.)
- [28] Brian Cantwell Smith, The semantics of clocks, in James H. Fetzer (ed) *Aspects of Artificial Intelligence*, pp 3-31, Kluwer Academic Publishers, 1988.
- [29] C.N.Taylor, *A Formal Logical Analysis of Causal Relations*, draft D.Phil Thesis, School of Cognitive and Computing Sciences, Sussex University, 1992.

- [30] A.M. Turing, Computing machinery and intelligence in E.A. Feigenbaum and J. Feldman (eds) *Computers and Thought* (McGraw-Hill, New York, 1963) 11-35, (Originally in *MIND* vol 59, pp 433-460, 1950).
- [31] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, W.H.Freeman 1976