Minds, Machines, And Mathematics
A Review of Shadows of the Mind by Roger Penrose

David J. Chalmers
Department of Philosophy
Washington University
St. Louis, MO 63130
U.S.A.

dave@twinearth.wustl.edu

PSYCHE, 2(9), June, 1995
http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html

KEYWORDS: Gödel's theorem, Penrose, artificial intelligence, computation, consciousness, quantum mechanics

REVIEW OF: Roger Penrose (1994) *Shadows of the Mind*. New York: Oxford University Press. 457 pp. Price: $25.00. ISBN 0-19-853978-9.

1. Introduction

1.1 In his stimulating book *Shadows of the Mind*, Roger Penrose presents arguments, based on Gödel's theorem, for the conclusion that human thought is uncomputable. There are actually two separate arguments in Penrose's book. The second has been widely ignored, but seems to me to be much more interesting and novel than the first. I will address both forms of the argument in some detail. Toward the end, I will also comment on Penrose's proposals for a "new science of consciousness".

2. The First Argument

2.1 The best way to address Gödelian arguments against artificial intelligence is to ask: what would we *expect*, given the truth of Gödel's theorem, if our reasoning powers could be captured by some formal system F? One possibility is that F is essentially unsound, so that Gödel's theorem does not apply. But what if F is sound? Then we would expect that:
(a) F could not prove its Gödel sentence G(F);
(b) F could prove the conditional "If F is consistent, then G(F) is true";
(c) F could not prove that F is consistent.
2.2 If our reasoning powers are capturable by some sound formal system F, then, we should expect that we will be unable to see that F is consistent. This does not seem too surprising, on the face of it. After all, F is likely to be some extremely complex system,

perhaps as complex as the human brain itself, and there is no reason to believe that we can determine the consistency of arbitrary formal systems when those systems are presented to us.

2.3 There does not seem to be anything especially paradoxical about this situation. Many arguments from Gödel's theorem, such as that given by Lucas, founder at just this point: they offer us no reason to believe that we can see the truth of our own Gödel sentence, as we may be unable to see the consistency of the associated formal system. How does Penrose's argument fare?

2.4 Penrose is much more cautious in his phrasing. In Chapter 2, he argues carefully for the conclusion that our reasoning powers cannot be captured by a "knowably sound" formal system. This seems to be correct, and indeed mirrors the analysis above. If we are a sound formal system F, we will not be able to determine that F is sound. So far, this offers no threat to the prospects of artificial intelligence. The real burden of Penrose's argument is carried by Chapter 3, then, where he argues that the position that we are a formal system that is not "knowably sound" is untenable.

2.5 One position that an advocate of AI might take is to argue that our reasoning is fundamentally unsound, even in an idealization. I will not take this path, however. For a start, I have some sympathy with Penrose's idea that we have an underlying sound competence, even if our performance sometimes goes astray. But further, it seems to me that to hold that this is the only problem in Penrose's argument would be to concede too much power to the argument. It would follow, for example, that there are parts of our arithmetical competence that no *sound* formal system could ever duplicate; it would seem that our unsoundness would be essential to our capacity to see the truth of Gödel sentences, for example. This would be a remarkably strong conclusion, and does not seem at all plausible to me. So I think that the deepest problems with Penrose's argument must lie elsewhere.

2.6 I will concede to Penrose that we are fundamentally sound, then. As before, the natural position for an advocate of AI is that our powers are captured by some sound formal system F that cannot demonstrate that F is sound. What is Penrose's argument against this position? He has two sub-arguments here, depending on whether we can know that F is the formal system that captures our reasoning.

2.7 If we *could* know that F captures our reasoning, Penrose's argument would be very straightforward:
(1) We know that we are sound;
(2) We know that F captures our reasoning;
so (3) We know that F is sound.
One might question premise (1) -- I will raise some problems with it later -- but it does have a certain plausible quality. Certainly, it seems antecedently more plausible than the much stronger position that we know that F is sound. But all this is irrelevant, as premise (2) is so implausible. There is very little reason to believe that if our reasoning is captured by F, then we could know that fact.

2.8 It might *seem* plausible that we could know that F underlies our processing -- why couldn't we just investigate our underlying brain processes? But to do this would be to change the game. It is of no help to Penrose if we can know *using external resources* (such as perceptual inputs) that F captures our reasoning. For to use external resources would be to go beyond the resources provided by F itself. And there would be no contradiction in the supposition that F could know, using external resources, that F is consistent, and therefore that G(F) is true. A contradiction would only arise if F could know this wholly under its own steam.

2.9 For this argument to be at all relevant, then, we would need to know that F captures our reasoning powers wholly using our internal resources -- that is, the resources that F itself provides. But there is not the slightest reason to believe that we could do this. If we are a formal system, we certainly cannot determine which formal system we are on the basis of introspection! So again, the advocate of artificial intelligence is in no danger. She need simply hold the unsurprising position that we are a formal system F, but that we can't tell through introspection that we are F.

2.10 To make his case, Penrose needs to argue that if we are a sound formal system F, then we could determine that F is sound, *independently* of any knowledge that we are F. That is, he needs to make the case that if F is presented to us, we could determine that it is sound through an analysis of F alone. This is the burden that Penrose tries to meet in section 3.3. It is this section that effectively carries all the crucial weight; if it does not succeed, then this line of Penrose's argument simply fails.

2.11 How does Penrose argue that we could see that F is sound? He argues in 3.3 that we can see F as a system of axioms and inference rules. Clearly, we can see that each of the axioms is true: if F can see their truth, so can we. Further, Penrose argues, we must be able to see that each of the basic inference rules is valid, as it is extremely implausible that our reasoning could rely on inference rules that we regard as "fundamentally dubious". And if we know that the axioms are true and that the inference rules are valid, then we know that F is sound.

2.12 But why should we accept that F consists of a set of axioms and inference rules? F, after all, is supposed to potentially correspond to *any* sort of computational system -- it might be a simulation of the whole of the human brain, for example. This will not look anything like a neat logical system: we will not be able to decompose it into an underlying set of "axioms" and "rules of procedure". Rather, it will be a big computational system that churns away on a given statement as input, and eventually outputs "yes" or "no".

2.13 It is true that for any Turing machine that accepts a certain class of statements, we can find a corresponding axiom-plus-rules system that accepts the same class (or at least the closure of that class under logical consequence). There is a lemma by Craig to this effect; without it applications of Gödel's theorem to draw conclusions about Turing machines would not even get off the ground. But the "axiom-plus-rules" system that we

end up with may be extraordinarily complex. In particular, the "inference rules" may be just about as complex as the original system -- perhaps equivalent to a complex connectionist procedure for generating further theorems. And as before, there is no reason why we should be able to see that this sort of "rule" should be valid, any more than we could see from an analysis that an overall computational brain process is sound. This is not to say that we think we are relying on "fundamentally dubious" procedures -- it is just that the procedures that govern the dynamics of our brain are too complex for us to analyse them as sound or otherwise.

2.14 In this section, Penrose seems to assume that the relevant class of computational systems are all something akin to theorem-provers in first-order logic, but of course there is no reason to make such an assumption. For his argument to have its full generality, proving that our physical processes could not even be simulated computationally, it must apply to any sort of computational process. Even within the realm of existing AI research, there are many computational procedures, such as connectionist networks, which are not decomposable into axioms and rules of inference.

2.15 (I suspect that even an advocate of logic-based AI might have a response to make here. It might be held, for example, that we may occasionally use certain complex inference rules (when we generate Gödel sentences by transfinite counting, for example), whose validity is not obvious to us on analysis, without this in any way impugning the reliability of our reasoning. We might soundly "use" a procedure despite its resistance to our analysis. This indeed is just what we might expect around the "outer limits" of Gödelization, which after all is really where Penrose's argument gains its force. There is no difficulty in the idea that the reasoning methods we use in *everyday* mathematics can be seen to be sound -- Penrose's arguments really apply at the level of our unusual "Gödelizing" procedures, which rely on our ability to count transfinite ordinals. But to be able to see that some Gödelizing rule is valid would be akin to making that last step in a Gödelization procedure, the one that is just complex enough to be beyond us. But I leave these difficult issues aside for now.)

2.16 It is section 3.3 that carries the burden of this strand of Penrose's argument, but unfortunately it seems to be one of the least convincing sections in the book. By his assumption that the relevant class of computational systems are all straightforward axiom-and-rules system, Penrose is not taking AI seriously, and certainly is not doing enough to establish his conclusion that physics is uncomputable. I conclude that none of Penrose's argument up to this point put a dent in the natural AI position: that our reasoning powers may be captured by a sound formal system F, where we cannot determine that F is sound.

3. Penrose's Second Argument

3.1 Hiding at the back of Chapter 3, however, Penrose has a new argument that escapes many of these problems. It is unfortunate that this argument was so deeply buried; most commentators seem to have missed it. Unlike the previous argument, this argument does *not* depend on the claim that we if we are a sound formal system F, we would be able to

see that F is sound. Because of this, it is a more novel and interesting argument, and more worthy of attention.

3.2 The argument is developed in a roundabout way (which may have led some readers astray), but is summarized in the fantasy dialogue with a robot mathematician in 3.23. The argument is given in a somewhat indirect form, involving complex procedures by which a given formal system might have evolved, but its basic structure is very simple. In a simplified and somewhat loose form, the argument goes as follows:
(1) Assume my reasoning powers are captured by some formal system F (to put this more briefly, "I am F"). Consider the class of statements I can know to be true, *given* this assumption.

(2) Given that I know that I am F, I know that F is sound (as I know that I am sound). Indeed, I know that the larger system F' is sound, where F' is F supplemented by the further assumption "I am F". (Supplementing a sound system with a true statement yields a sound system.)

(3) So I know that G(F') is true, where this is the Gödel sentence of the system F'.

(4) But F' could not see that G(F') is true (by Gödel's theorem).

(5) By assumption, however, I am now effectively equivalent to F'. After all, I am F supplemented by the knowledge that I am F.

(6) This is a contradiction, so the initial assumption must be false, and F must not have captured my powers of reasoning after all.

(7) The conclusion generalizes: my reasoning powers cannot be captured by any formal system.
3.3 Strictly speaking, the conclusion that must be drawn is that I cannot *know* that I am identical to a formal system F; in showing that I can see the truth of G(F'), we assumed not just that I am F but that I know I am F. But this is still a strong conclusion. For example, it would rule out even the possibility that we could empirically discover that we were identical to some system F -- if we were to "discover" this, the reasoning would lead us to a contradiction. So even this would be threatening to the prospects of AI.

3.4 The power of this argument stems from the fact that is does not depend on one's ability to determine that a system F is sound, or to determine that we are F. Rather, it relies on the *assumption* that one is F to reach the relevant conclusions, thus contradicting the assumption. On the face of it one might have thought that making such an assumption would show only that the larger system F' could prove the Gödel sentence of the smaller system F, but the insight of the argument is that things can be bootstrapped into a situation where F' sees its own Gödel sentence, leading to trouble.

3.5 As far as I can determine, this argument is free of the obvious flaws that plague other Gödelian arguments, such as Lucas's argument and Penrose's earlier arguments. If it is

flawed, the flaws lie deeper. It is true that the argument has a feeling of achieving its conclusion as if by magic. One is tempted to say: "why couldn't F itself engage in just the same reasoning?". But although there are various directions in which one might try to attack the argument, no knockdown refutation immediately presents itself. For this reason, the argument is quite challenging. Compared to previous versions, this argument is much more worthy of attention from supporters of AI.

3.6 On reflection, I have come to believe that the greatest vulnerability in this argument lies in the assumption that we know (unassailably) that we are consistent. This assumption seems relatively innocuous, compared to the previous strong claim that we could determine that F is consistent; on the face of it, it does not seem vastly stronger than the assumption that we are consistent. But I think that in fact, it is this assumption, and not the assumption that we know we are F, that carries the central responsibility for generating the contradiction. (I have largely become convinced of this through discussions with Daryl McCullough, and the central argument below, an adaptation of a result of Löb's, was suggested by him.)

3.7 The best way to see this is to show that the assumption that we know we are consistent *already* leads to a contradiction in its own right, even without the further assumption that we know we are F. Specifically, we can argue that any system that "unassailably" believes in its own consistency will in fact be led to a contradiction (under certain plausible further assumptions). This can be done as follows.

3.8 In these matters, we are concerned with a system's reasoning about its own beliefs, as well as about mathematics. So we can assume it has a symbol B, representing belief, where B(n) corresponds to the statement that it believes the statement with Gödel number n. (Below, I abbreviate by writing "B(A)" instead of "B(`A')", where `A' is the Gödel number of A.) And let us write "|- A" if the system has the power to "unassailably" assert A. (By using this notation I do not intend to beg the question about whether the system is computational!) Then the following assumptions are reasonable (suppressing universal qualifiers):

(1) If |- A, then |- B(A).
(2) |- B(A_1) & B(A_1 -> A2) -> B(A2)
(3) |- B(A) -> B(B(A))

3.9 (1) says that if the system has the power to assert A, it has the power to assert B(A). (2) says essentially that the system knows it has the power to reason by modus ponens. (3) says, in effect, that the system knows (1). All of these assumptions seem unproblematic. To these we add the key assumption:
(4) |- not B(false)
which says that the system asserts that it is not inconsistent. It turns out that these assumptions, along with the assumption that the system has the resources to do Peano arithmetic, lead to a contradiction.

3.10 To see this, we simply construct a sentence G such that

(5) |- G -> not B(G).
This is a standard diagonal construction, and does not rely on any assumptions about the system's computability. We define the function "diag" in Peano arithmetic so that diag(`C(x)') is `C(`C(x)')' for any predicate C. (For clarity, I reintroduce the `' notation for Gödel numbering.) Then let G be the sentence $not B(diag(`not B(diag(x))'))$. It is straightforward to show that G -> not B(`G'). As long as the system has at least the capacities of Peano arithmetic, it can replicate this reasoning, so that |- G -> not B(`G').

3.11 G is effectively a sentence that says "I do not believe G", much like a standard Gödelian construction, but without any assumptions about computability. It is not hard to see how the contradiction arises. The system knows that if it believes G, it is unsound; so it knows that if it is sound, it does not believe G. But this is to say that it knows that if it is sound, G is true. By assumption, it knows that it is sound, so it knows that G is true. So now it must be unsound, as it has fallen into a contradiction. This reasoning is easily formalized:
(6) |- B(G) -> B(not B(G)) [from (5), (1), (2)]
(7) |- B(G) -> B(B(G)) [from (3)]
(8) |- B(G) -> B(false) [from (6), (7), (2)]
(9) |- B(false) -> B(G) [from (2), along with |- B(false -> G)]
(10) |- G -> not B(false) [from (5), (8), (9)]
(11) |- B(G) [from (10), (4), (1)]
(12) |- B(false) [from (12), (9)]
3.12 We can see, then, that the assumption that we know we are sound leads to a contradiction. One might try to pin the blame on one of the other assumptions, but all these seem quite straightforward. Indeed, these include the sort of implicit assumptions that Penrose appeals to in his arguments all the time. Indeed, one could make the case that all of premises (1)-(4) are implicitly appealed to in Penrose's main argument. For the purposes of the argument against Penrose, it does not really matter which we blame for the contradiction, but I think it is fairly clear that it is the assumption that the system knows that it is sound that causes most of the damage. It is this assumption, then, that should be withdrawn.

3.13 Penrose has therefore pointed to a false culprit. When the contradiction is reached, he pins the blame on the assumption that our reasoning powers are captured by a formal system F. But the argument above shows that this assumption is inessential in reaching the contradiction: A similar contradiction, via a not dissimilar sort of argument, can be reached even in the absence of that assumption. It follows that the responsibility for the contradiction lies elsewhere than in the assumption of computability. It is the assumption about knowledge of soundness that should be withdrawn.

3.14 Still, Penrose's argument has succeeded in clarifying some issues. In a sense, it shows where the *deepest* flaw in Gödelian arguments lies. One might have thought that the deepest flaw lay in the unjustified claim that one can see the soundness of certain formal systems that underlie our own reasoning. But in fact, if the above analysis is correct, the deepest flaw lies in the assumption that we know that we are sound. All

Gödelian arguments appeal to this premise somewhere, but in fact the premise generates a contradiction. Perhaps we are sound, but we cannot know unassailably that we are sound.

4. The Missing Science Of Consciousness?

4.1 A reader who is not convinced by Penrose's Gödelian arguments is left with little reason to accept his claims that physics is noncomputable and that quantum processes are essential to cognition, although these speculations are interesting in their own right. But even if one accepts that human behavior can be accounted for computationally, there is still the question of human consciousness, which after all is Penrose's ultimate target.

4.2 Penrose is clear that the puzzle of consciousness is one of his central motivations. Indeed, one reason for his skepticism about AI is that it is so hard to see how the mere enaction of a computation should give rise to an inner subjective life. Why couldn't all the computation go in the dark, without consciousness? So Penrose postulates that we to appeal to physics instead, and suggests that the locus of consciousness may be a quantum gravity process in microtubules. But this seems to suffer from exactly the same problem. Why should quantum processes in microtubules give rise to consciousness, any more than computational processes should? Neither suggestion seems appreciably better off than the other.

4.3 Although Penrose's quantum-gravity proposal might at least *conceivably* help explain certain elements of human behavior (if behavior turned out to be uncomputable, for example), it simply seems to be the wrong sort of thing to explain human consciousness. Indeed, Penrose nowhere claims that it does, and by the end of the book the "Missing Science of Consciousness" seems as far off as it ever was. As things stand, even by the end of Penrose's book, we seem to be left in Penrose's position D: these physical theories leave consciousness entirely unexplained.

4.4 This might seem odd, given that Penrose says he embraces position C, but in fact C and D are quite compatible. This is because Penrose's four positions run together a number of separate issues. For convenience, I repeat the positions here:
A: All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.

B: Awareness if a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness.

C: Appropriate physical action evokes awareness, but this physical action cannot even be properly simulated computationally.

D: Awareness cannot be explained by physical, computational, or any other scientific terms.

4.5 Note that A, B, and C all concern how awareness is *evoked* , but D concerns how awareness is *explained*. These are two very different issues. To see the contrast, note that almost everybody would accept that the brain *evokes* awareness -- if we were to construct a duplicate brain, there would be conscious experience associated with it. But it is far from clear that a physical description of the brain can *explain* awareness -- many people have argued that given any physical account of brain processes, the question of *how* those processes evoke conscious experience will be unanswered by the physical account.

4.6 To really clarify the positions in the vicinity, we have to distinguish three questions:
(1) What does it take to simulate our physical *action* ?
(2) What does it take to *evoke* conscious awareness?
(3) What does it take to *explain* conscious awareness?

4.7 In answer to each question, one might say that (a) Computation alone is enough, (b) Physics is enough, but physical features beyond computation are required, or (c) Not even physics is enough. Call these positions C, P, and N. So we have a total of 27 positions, that one might label CCC, CPN, and so on.

4.8 Question (1) is the question Penrose is concerned with for most of the book, and the issue that separates B and C above. He argues for position P-- over C--. Descartes might have argued for N--, but few would embrace such a position these days.

4.9 Question (2) is the issue at the heart of Searle's Chinese room argument, and the issue that separates A from B and C above. Searle argues for -P- over -C-, and Penrose is clearly sympathetic with this position. Almost everyone would accept that a physical duplicate of me would "evoke" consciousness, so position -N- is not central here.

4.10 Question (3) is the central question about the *explanation* of consciousness (a question that much of my own work is concerned with). Penrose's positions A, B, and C are neutral on this question, but D is solely concerned with it; so in a sense, D is independent of the rest. Many advocates of AI might hold --C, some neurobiologists might hold --P, whereas my own position is --N.

4.11 The four positions Penrose describes come down to CC- (A), CP- (B), PP- (C), and --N (D). Penrose seems to think that in arguing for position C (PP-) he is arguing against position D (--N), but it is clear from this analysis that this is not so. In the end, nothing in Penrose's book bears on question (3), which is a pity, though it is certainly understandable. It would be very interesting to hear Penrose's position on just how physical theories might or might not explain human consciousness.

4.12 Indeed, one might even combine positions A and D, as I do, embracing CCN. On this position, human-like behavior can be produced computationally, and indeed enacting the right computation will give rise to consciousness, but neither a computational account nor a physical account alone will explain consciousness. It might seem odd that computation should evoke but not explain consciousness, but this is no more odd than the corresponding position that neurophysiology might evoke but not explain consciousness.

In either case, consciousness emerges from some underlying basis, but we need a further element in the theory to explain just how and why it emerges.

4.13 One can have a lot of fun cataloging positions (Dennett is CCC; Searle may be CPP; Eccles is NNN; Penrose is PPP; I am CCN; some philosophers and neuroscientists are CPN or PPN; note that all these are "non-decreasing" in C->P->N, as we might expect), but this is enough for now. The main point is that Penrose's treatment runs together question (3) with questions (1) and (2), so that in the end the question of how consciousness might be explained is left to one side.

4.14 A true science of consciousness will have to address all of these questions, and especially question (3). Penrose has produced an enormously enjoyable and challenging book, but it seems to me that for all his hard work, the science of consciousness is still missing.<1>

Notes

<1> This review is an elaboration of my review of Penrose's book in *Scientific American*, June 1995, pp. 117-18.