Can Humans Escape Gödel?
A Review of *Shadows of the Mind* by Roger Penrose

Daryl McCullough
ORA Corporation
301A Harris B. Dates Drive
Ithaca, NY 14850-1313.

daryl@oracorp.com

PSYCHE, 2(4), April 1995
http://psyche.cs.monash.edu.au/v2/psyche-2-04-mccullough.html

KEYWORDS: belief, consistency, Gödel, knowledge, Penrose, self- reference, truth.

REVIEW OF: Roger Penrose (1994) *Shadows of the Mind*. New York: Oxford University Press. 457 pp. Price: $25 hbk. ISBN 0-19-853978-9.

1. Gödel's Theorem And The Mind

1.1 In the first part of *Shadows of the Mind*, Penrose gives an argument that human reasoning must go beyond what is computable. Therefore, no computer program can ever hope to be as intelligent as a human being. Penrose doesn't give a direct argument for his thesis. He doesn't for instance, show that there is some task that humans can perform which no computer can. (Although he suggests without offering a proof that certain kinds of geometric visualization may allow us to deduce facts in an inherently noncomputable way.) Instead, Penrose uses an indirect proof-he assumes that there exists a computer program that is every bit as intelligent as a human, and shows that that leads to a contradiction.

2. Penrose's Argument

2.1 The basic contradiction for Penrose is this: Assume that the reasoning powers of some mathematician, say Penrose himself, are completely described by some formal system F. What this means is that for every mathematical statement S in the language of F that Penrose finds to be "unassailably true", S is a theorem of F, and vice-versa. We will further assume that Penrose knows that F describes his reasoning.

2.2 According to Penrose, the belief that F describes his own reasoning entails a belief in the soundness of F. (Penrose justifies this, saying "It would be an unreasonable

mathematical standpoint that allows for a disbelief in the very basis of its own unassailable belief system.")

2.3 By Gödel's theorem, since F is sound, then G(F), the Gödel statement for F, must be true, but not a theorem of F. Therefore, since Penrose believes that F is sound, he must conclude that G(F) is "unassailably true". So there is something (namely, G(F)) that Penrose finds unassailably true, but which is not a theorem of F. This contradicts the assumption that F completely describes the reasoning powers of Penrose (including his knowledge that F has this property.)

3. Loose Ends

3.1 This pretty much proves Penrose's conclusion, except for a few loose ends. First of all, there is a slight ambiguity in the meaning of F that needs to be addressed. One possible interpretation of F is that it represents the inherent reasoning ability of the mathematician. An alternative interpretation is that it represents a "snapshot" of the state of the mathematician's brain at one instant, and so includes both inherent reasoning ability and also empirical knowledge acquired during the mathematician's lifetime. A third possibility is that F represents the limits of what could ever be known by the mathematician, no matter whether through reasoning or through empirical knowledge. The differences among these alternatives become important when it comes to the question of whether the mathematician knows that his reasoning is described by F. It may, for instance, be that the mathematician learns the role of F through empirical means, and so this additional knowledge is not reflected in F. Penrose, in section 3.16 gets around this problem by considering a new system F', which includes F plus everything that follows from the information that the mathematician's reasoning powers were described by F (immediately prior to learning this knowledge). Then the same argument can go through using F' instead of F.

3.2 Other loose ends: In order for Penrose's argument to go through, he needs to make the following assumptions about human mathematical reasoning:

- Human mathematical reasoning is sound. That is, every statement that a competent human mathematician considers to be "unassailably true" actually is true.
- The fact that human mathematical reasoning is sound is itself considered to be "unassailably true".

3.3 So, the Gödel argument doesn't prove that human reasoning must be noncomputable - it only proves that if human reasoning is computable, then it must either be unsound, or it must be inherently impossible for us to know both what our own reasoning powers are and to also know that they are sound. Penrose dismisses the possibility that we know our reasoning powers but don't know that they are sound in the discussion in section 3.2 of *Shadows*. Penrose claims that if we knew that some particular computer program F was

equivalent to human reasoning, then we would be forced to conclude that F was sound. But it is this point that I take issue with.

3.4 To me, this is more a statement of psychology than of mathematics. Penrose considers certain of his beliefs about mathematics to be "unassailably true", and he cannot even consider the possibility that some of these beliefs might be wrong. Given that he holds this conviction, it doesn't follow that Penrose's reasoning is not computable, it only follows that Penrose can never be convinced that it is. For people (such as me) who have a more relaxed attitude towards the possibility that their reasoning might be unsound, Penrose's argument doesn't carry as much weight.

3.5 In the next sections, I will discuss two additional questions which I think were not discussed adequately by Penrose: (1) Does the assumption that human reasoning is noncomputable save us from the Gödel-style paradoxes? (2) If our reasoning is inconsistent, then where could the inconsistency come from? How could a careful, intelligent mathematician make the sorts of mistakes that could lead to an inconsistency?

4. Can Noncomputable Theories Escape Gödel's Theorem?

4.1 Even among mathematical experts, there is a widespread misconception that Gödel's theorem only applies to computable theories. I believe that the reason for this belief is that Gödel's theorem fails to apply to the only well-known noncomputable theory, namely the complete theory of arithmetic. However, it is not difficult to show a stronger form of Gödel's incompleteness theorem:

4.2 Given any theory (collection of statements close under logical deduction) T, the theory is either unsound or incomplete if the following conditions hold:

- The formulas of T can be encoded as terms of T so that syntactic operations such as substitution can all be defined in T.
- A "theoremhood" predicate is definable in T. That is, there is a formula P(x) expressing the proposition that x is a code for a theorem of T.

4.3 To see that these conditions lead to the incompleteness of T, let us first define a substitution function. If A is some formula (possibly having free variables) in the language of T, and i is its code, then let sub(i,j) be the code of A', which is the result of substituting j for each free variable occurring in A. We have assumed that T is expressive enough to define such syntactic functions. Now define G_0 to be a formula expressing $not P(sub(x,x))$. (Strictly speaking, G_0 may not actually be $not P(sub(x,x))$, since there may be no symbol "sub" in the language. However, the definability of substitution implies that there is a formula expressing essentially the same meaning.)

4.4 Let G be the formula constructed from G_0 by replacing all free variables by n, where n is the code for G_0. G thus expresses the statement $not P(sub(n,n))$. It is clear that G holds if and only if the term sub(n,n) is not the code of a theorem of T. But on the other hand, G is the result of substituting n for each free variable occurring in the statement (namely, G_0) whose code is n. Therefore, by definition of the substitution function, the code of G is sub(n,n). So G holds if and only if G is not a theorem of T.

4.5 It is clear that if G were a theorem of T, then G would be false (since it "says" that it is not a theorem). Therefore, if G is a theorem, then T is unsound. Turning that around, it follows that if T is sound, then G is not a theorem (and therefore, true). So if T is sound, it must be incomplete (there is a true sentence, G, which is not a theorem.)

4.6 With a few more mild conditions on the theoremhood predicate (due to the logician Lob), it is possible to prove a stronger statement: G is true (and unprovable) if and only if T is consistent. This is a much more useful result, since consistency is definable within T, while soundness is not. It follows that if T is consistent, then T cannot prove its consistency.

4.7 So, the incompleteness theorem does not rely on a theory being axiomatizable; it only relies on the theory possessing a theoremhood predicate. In the case of computable theories (at least those extending Peano Arithmetic), a theoremhood predicate is always definable. However, for noncomputable theories, a theoremhood predicate may or may not be definable. In the case of true arithmetic, Tarski proved in effect that there is no theoremhood predicate. There is no formula P(x) in the language of arithmetic expressing the fact that x is a code for a true statement of arithmetic. However, there is such a formula in the language of set theory (ZFC). Therefore, the theory ZFC+, whose axioms are (1) all true statements of arithmetic, and (2) all axioms of ZFC is an example of a noncomputable theory that nevertheless has a theoremhood predicate. Gödel's theorem applies to this noncomputable theory, so there is a "Gödel statement" which is true but ZFC+ cannot prove it. Also, just like computability theories, ZFC+ cannot prove its own consistency.

5. Does Gödel's Theorem Apply To Humans?

5.1 Penrose's arguments depend on the ability of mathematicians to grasp certain "unassailable truths". While it may be the case that some truths are so difficult that they can never be considered unassailably true, it should be the case that nothing false can be unassailably true. However, it can be shown that, even though it might be the case that nothing false is ever judged to be unassailably true, this fact cannot be an unassailable truth.

5.2 The "quick and dirty" way to show this is to use an explicitly self-referential sentence. Let G be the following sentence:
This sentence is not an unassailable belief of Roger Penrose.
If we suppose that G is one of Roger Penrose's unassailable truths, then we immediately conclude that it must be false. Therefore, Roger Penrose's unassailable beliefs include at least one false statement. Turning that around, if Roger Penrose's beliefs are sound (they do not include any false statements), then it must be that G cannot be one of his unassailable beliefs. But since G says that it is not one of his unassailable beliefs, it follows that G must be true. So, we conclude:
If Roger Penrose is sound, then G is true.
Now, since Roger Penrose is capable of seeing the truth of the above implication, it follows that if he believes himself sound, then he will believe G. But, by definition of G, if Penrose believes G (unassailably), then G must be false. So, if Roger Penrose believes he is sound, then G is false and yet Roger Penrose believes that it is true. Therefore, we conclude:
If Roger Penrose believes he is sound, then he is, in fact, unsound.
5.3 A slightly more mathematical argument uses definition paradoxes such as Richard's paradox ("The smallest number that can not be described in fewer than thirteen words.") Here is a related paradox:

Let F(x) be a function from integers to integers defined as follows:
Interpret the binary expansion of x as a sequence of bytes, or characters. If x unambiguously defines a total function G from integers to integers, then the value of F(x) is G(x) + 1. Otherwise, the value of F(x) is 0.
Now, let N be the binary number coding the bytes in the above description and consider the expression F(N). To evaluate this expression, we need first to determine whether N codes an unambiguous definition of a total function. Well, N is just the definition of F, which at least appears to be well-defined. But then the definition of F would then require the value of F(N) to be F(N) + 1, which is impossible. This contradicts the assumption that F is a total function; it can't possibly be defined for N. However, if we know that N does not define a total function, then the above definition seems to give a definite result: F(N) is specified to be 0.

5.4 The resulting paradox seems to me to show that the notion of "unambiguous definition" cannot itself be unambiguous. Similarly, the notion of "unassailable truth" cannot itself be unassailable.

5.5 Such self-referential arguments may seem perhaps too "cute" to be believed. We know from the Liar paradox to be suspicious of explicitly self-referential sentences. However, we can eliminate the explicit self-reference and still reach the same conclusion. All that is necessary is the construction of a sentence G such that G holds if and only if it is not an unassailable truth.

5.6 Since Penrose rejects the idea that human reasoning is beyond science, he seems to be committed to the belief that one day we might have a mathematical theory of how the human brain works. Therefore, using that mathematical theory, it will be possible (principle) to formulate a mathematical formula P(x) which holds only if an (idealized, error-free) human brain would find the statement coded by x to be "unassailably true". Whether or not P(x) is computable, we can use this formula to construct a "Gödel statement" for humans: a statement G which, if our reasoning is consistent, would be true but not believed to be "unassailably true" by us. Using Penrose's principle that we are forced to believe in our own soundness, it follows that we would be forced to conclude that G must be true. But this contradicts the definition of G as true but not believed to be true by us!

5.7 The resulting contradiction shows that either Penrose is wrong, and we can't be unassailably convinced of our own soundness, or else Penrose is wrong, and the human brain can never be described by mathematics (and thus not by science, according to the current view of science). Therefore, if Penrose's arguments support any conclusion about the human mind, it would seem to me to support the position that the mind is forever beyond science (philosophical position D in the discussion of mind in section 1.3 of *Shadows*), rather than simply that it is beyond what is computable. There is nothing in Penrose's argument that couldn't just as well rule out any mathematical theory of the mind, not just computable theories.

6. How Could Inconsistency Creep Into Human Reasoning?

6.1 As I discussed in the last section, Penrose's arguments, if taken to their logical conclusion, show us not that the human mind is noncomputable, but that either the human mind is beyond all mathematics, or else we cannot be sure that it is consistent. If we reject the "mysterian" position that mind is beyond science, we are left with the conclusion that we can't know that we are consistent. This seems very counter-intuitive. If we are very careful, and only reason in justified steps, why can't we be certain that we are being consistent?

6.2 Let me illustrate with a thought experiment. Suppose that an experimental subject is given two buttons, marked "yes" and "no", and is asked by the experimenter to push the appropriate button in response to a series of yes-no questions. What happens if the experimenter, on a lark, asks the question "Will you push the 'no' button?". It is clear that whatever answer the subject gives will be wrong. So, if the subject is committed to answering truthfully, then he can never hit the "no" button, even though "no" would be the correct answer. There is an intrinsic incompleteness in the subject's answers, in the sense that there are questions that he cannot truthfully answer.

6.3 Now, there is no real paradox in this thought experiment. The subject knows that the answer to the experimenter's question is "no", but he cannot convey this knowledge. Thus there is a split between the public and private knowledge of the subject. But now, let's extend the thought experiment.

6.4 Someday, as science marches on, we will understand the brain well enough that we can dispense with the "yes" and "no" buttons (which are susceptible to lying on the part of the subject). Instead of these buttons, we assume that the experimenter implants probes directly into the subject's brain, and we assume that these probes are capable of directly reading the beliefs of this subject. If the probes detect that the subject's brain is in the "yes" belief state, it flashes a light labeled "yes", and if it detects a "no" belief state, it flashes a light labeled "no". Now, in this improved experiment, the subject is asked the question "Will the 'no' light flash?"

6.5 In this improved set-up, there is no possibility of the subject having knowledge that he can't convey; the probe immediately conveys any belief the subject has. If the subject believes the "no" light will flash, then the answer to the question would be "yes", and the subject's beliefs would be wrong. Therefore, if the subject's beliefs are sound then the answer to the question is "no". Therefore, since the subject cannot correctly believe the answer to be "no", he similarly cannot correctly believe that he is sound. If the subject reasons from the assumption of his own soundness, he is led into making an error.

6.6 As can be seen from this thought experiment, the inability to be certain of one's own soundness is not a deficiency of intelligence. There is no way that the subject in the experiment can correctly answer the question by just "thinking harder" about it.

7. How Can Inconsistency Creep Into Mathematics?

7.1 Penrose in *Shadows of the Mind* was not concerned with beliefs in general, but only with beliefs about mathematics. In the pristine world of mathematics, is there a way to be careful, and make sure that our reasoning is consistent? It is understandable that if we start playing around with axioms we don't understand, such as the large - cardinal axioms of set theory, we might run into an inconsistency. However, suppose we stick to more concrete, understandable mathematics. For instance, Peano's theory of arithmetic. Surely, we can be certain that elementary arithmetic is consistent? Its axioms are only statements about plus and times which are obviously true to anyone who understands the simplest facts about numbers.

7.2 Let's try to imagine a mathematician who is trying to figure out the limits of what the "unassailable truths" of arithmetic are. If the mathematician starts proving facts about arithmetic one at a time, using standard arithmetical methods (such as proof by induction) he can be pretty sure that he will never make a mistake. After a while, he might realize

that everything he is doing can actually be automated - he can formalize the rules of arithmetic as axioms and rules of inference, and he could, in principle, write a computer program that could, given enough time, prove every possible theorem that can be obtained using those rules. Since the mathematician is confident that he set up the axioms correctly, he can, following Gödel, conclude that the resulting theory is consistent, so the Gödel statement for that theory is true but unprovable.

7.3 The mathematician could then construct a second theory, which used as axioms all the axioms of the first theory, plus the Gödel statement for that first theory. This theory should be as sound as the first was. In a similar way, the mathematician could construct a third, more powerful theory, and a fourth, etc. All of them would seemingly be as sound as the first.

7.4 Sooner or later, the mathematician might take a step back from his theory-building, and think: "You know, I think this process of building theories could itself be automated. I could build a new theory, I will call it the Omega theory, which will be the union of all theories that are obtainable by a finite number of steps in my original sequence of theories."

7.5 Once the mathematician sets up the Omega theory, he can again use Gödel's theorem to get a theory more powerful than that, and another even more powerful. Eventually, he would get around to building a second Omega theory infinitely more powerful than the first Omega theory, and then a third Omega theory infinitely more powerful than the second. Then, the mathematician might get the idea of building an Omega-squared theory, which would be the union of all the Omega theories. He can go on forming more and more powerful theories, corresponding to bigger and bigger ordinals.

7.6 Now, all of the mathematician's theories seem to only use the two obviously sound principles: A statement is considered to be "unassailably true" under the following circumstances: (1) It is a theorem of PA, or (2) It is a statement of the form G(T), where T is a theory consisting of only "unassailably true" facts. Surely, there is absolutely no way that an inconsistency could ever arise in the collection of "unassailably true" facts. So, why can't we conclude that the collection of unassailably true facts (those obtained using only these two rules of inference) is consistent?

7.7 The problem is that in order to use Gödel's theorem to get ever more powerful mathematical theories, our mathematician needs to formalize more and more of his own reasoning, and then make the "leap" to the conclusion that that formalization is itself consistent (and therefore, the corresponding Gödel statement is true.) However, if the mathematician formalizes too much of his own reasoning, including the "leaps", then the resulting theory will be able to formalize itself, and make the leap to the conclusion that its own Gödel statement is true. But this conclusion leads immediately to a contradiction.

7.8 So, either (1) the mathematician at some point stops short of formalizing all of his reasoning (in which case, the collection of all facts he can prove will be an axiomatizable theory), or else (2) he formalizes all of his reasoning, and the resulting theory is inconsistent (it would be able to prove its own consistency).

8. Conclusion

8.1 Penrose's arguments that our reasoning can't be formalized is in some sense correct. There is no way to formalize our own reasoning and be absolutely certain that the resulting theory is sound and consistent. However, this turns out not to be a limitation on what computers or formal systems can accomplish relative to humans. Instead, it is an intrinsic limitation in our abilities to reason about our own reasoning process. To the extent that we understand our own reasoning, we can't be certain that it is sound, and to the extent that we know we are sound, we don't understand our reasoning well enough to formalize it. This limitation is not due to lack of intelligence on our part, but is inherent in any reasoning system that is capable of reasoning about itself.