Why Godel's Theorem Cannot Refute Computationalism

Geoffery LaForte, Patrick J. Hayes & Kenneth M. Ford
Institute for Human & Machine Cognition
University of West Florida
Pensacola, Florida 32514

ABSTRACT

Gödel's theorem is consistent with the computationalist hypothesis. Roger Penrose, however, claims to prove that Gödel's theorem implies that human thought cannot be mechanized. We review his arguments and show how they are flawed. Penrose's arguments depend crucially on ambiguities between precise and imprecise senses of key terms. We show that these ambiguities cause the Gödel/Turing diagonalization argument to lead from apparently intuitive claims about human abilities to paradoxical or highly idiosyncratic conclusions and conclude that any similar argument will also fail in the same ways.

1. INTRODUCTION

How do we think? This is one of the oldest questions, and only recently has any progress been made towards answering it. While nobody has a fully plausible account, one apparently useful methodology can be summed up as *the computationalist hypothesis* — roughly, the hypothesis that the human brain can be understood as a kind of biological computer, and mental activity as a kind of computational process. The corresponding technology is called *artificial intelligence*, the effort to program computers to think.

This project produces in some a reaction of revulsion, even horror. Indeed, the spectre of artificial intelligence has inspired many of the world's scholars to engage in rituals of exorcism — their goal being the banishing of this spirit to the mechanistic netherworld from which it came. It is quite a piece of luck for these thinkers when an actual scientist becomes a champion claiming to draw a magic circle which pens the evil spirit of computationalism and keeps it from stealing mankind's glories. It is especially lucky if this scientist turns out to be some sort of mathematician. Mathematicians prove things, and what better way could there be to lay a ghost to rest than to prove that it never existed in the first place? There can, therefore, be little surprise at the acclaim and joy greeting the appearance of two recent books claiming to do just this by the distinguished mathematical physicist Roger Penrose. This is especially true since Penrose uses one of the most celebrated mathematical discoveries of the last hundred years, Gödel's incompleteness theorem, as the basis for his attack.

Our purpose here is not to give arguments for the computationalist hypothesis, or to defend it against the many other attacks which have been made upon it, but only to discuss this one particular recent argument against it, and to show why this, or any similar argument based on Gödel's incompleteness theorem, cannot succeed. Part of our case will involve showing that some apparently intuitive claims about consistency and self-knowledge lead after deeper examination to some very unintuitive conclusions.

2. THE PENROSE ARGUMENTS

Penrose's key idea is essentially the same as that of the philosopher J.R. Lucas in [6], an article which has attracted responses from several writers over the last thirty-five years (for example, Benacerraf in [1], and Lewis in [4, 5]). In brief, Lucas argued as follows. Gödel's incompleteness theorem shows that, given any formal system, there is a true sentence which the formal system cannot prove to be true. But since the truth of this unprovable sentence is proved as part of the incompleteness theorem, humans *can* prove the sentence in question. Hence, human abilities cannot be captured by formal systems.

Penrose has elaborated and extended this argument, defended it against many criticisms, and produced various forms and versions of it; indeed, his second book was written as a response to criticisms of the arguments in his first book. In subsequent discussions he has produced yet a third version of the argument, claiming that it avoids the criticisms of the second; but this suffers from essentially the same basic problems. All these arguments involve a kind of intellectual shell game, in which a precisely defined notion (such as *proof*, *computable* or *satisfiable*) to which a mathematical result applies, is switched for a vaguer notion (such as *argument*, *computational* or *sound*) which is used to carry the philosophical burden. It is ironic that the theorems being cited here — notably, Gödel's incompleteness results — arose ultimately out of the foundational crisis that finally forced philosophy of mathematics to come to terms with the unpalatable fact that untutored intuition (which Penrose boasts of using [9]) is not able to cope with the complexities arising from puzzles of self-reference. It is clear that Penrose, like many critics of the computationalist hypothesis, is not reporting a discovery but searching for a weapon to help him refute a view he finds offensive. When one club breaks, he picks up a new one. Hilary Putnam [10] and others [2] have already pointed out some of Penrose's technical errors, and he has conceded several of the technical points, but there does not seem to be a complete statement of the fundamental problems with this kind of argument.

Benacerraf in [1] made two points against the original Lucas argument, both of which we will develop more precisely here. The first focuses on the fact that in order to know the truth of the unprovable “Gödel sentence”, one has to know that the formal system is consistent. The thesis that we humans, unlike the formal system, can know the truth of the system's unprovable sentence is central to the argument. An alternative conclusion, therefore, is that humans may be unable to know that they are consistent. Penrose, like many who think his argument persuasive, finds this simply ridiculous; but we will show that it is more plausible than it may seem at first sight (for example, we will find inconsistencies in Penrose's own published opinions) and does not have the dire intellectual consequences that he fears.[1]

Benacerraf's second point is that Lucas equivocates on the notion of proof. Once "proof for humans" is given a clear definition, the argument applies to humans as well. This point was repeated against Penrose by both Boolos [2] and Putnam [10], and we will make it again here in more detail, to show how its assumptions are very weak and plausible. While Penrose believes himself to have overcome this objection, we will show that he can do so only by adopting rather extraordinary positions in philosophy of mathematics and in psychology.

---

[1]For example: "If our mathematical reasoning were indeed fundamentally unsound, then the whole edifice of scientific understanding would come crashing to the ground!" [8].

2.1 The Bare Gödel Argument

Gödel's Second Incompleteness Theorem can be stated informally as the claim that no reasonable set of axioms for arithmetic which is strong enough to prove certain basic facts about numbers is also strong enough to prove its own consistency, unless it is actually inconsistent. On the face of it, this seems to have nothing particularly to do with the computationalist hypothesis, which mentions computers rather than axioms for arithmetic. The relevance lies in what the words "reasonable" and "prove" mean in the statement of Gödel's theorem. Essentially, these words mean something like "able to be listed by a program" and "derive according to a formal proof procedure," which is itself a kind of algorithm. The thought behind this is that without some way of objectively checking whether something is an axiom or not, a proof system is essentially unreasonable, since there is no way in principle of resolving disputes about whether something follows from it or not. Thus, the theorem can be viewed, crudely, as saying that no consistent algorithm can produce a proof of its own consistency. This gives at least some indication why this theorem is often thought to be relevant to the computationalist hypothesis.

In fact, Penrose does not actually use Gödel's theorem, but rather an easier result inspired by Gödel, namely, Turing's theorem that the halting problem is unsolvable. This turns out to be simpler to work with in practice, and every "Gödelizing" argument using the unsolvability of the halting problem has a mirror image using Gödel's actual theorem. On the other hand, notions like consistency and truth are important for both our discussion and Penrose's, and these notions are explicitly involved only in the original incompleteness Theorem. To keep the discussion manageable, we will follow Penrose by using the unsolvability of the halting problem in the most technical parts of our discussion, while talking as though we have been using Gödel's theorem the rest of the time. We will be careful to have some precise correlation of everything we are saying with the bare Turing theorem rather than the richer Gödel result.

We will first present a clean mathematical proof of the result which seems so significant to Penrose without any added philosophical assumptions, and only then introduce and discuss the philosophical claims which make the result seem so significant. Penrose's discussion mixes formal and pre-formal notions from the very beginning in a way that seems quite natural, but that leads to confusion. In fact, we intend to show just how such a procedure is bound to lead this way. We will now introduce some of the standard terminology and notation which computability theorists use. (This is for the sake of convenience and brevity of expression, so we ask the nonmathematical reader to bear with us, and the mathematical reader to be assured that we are not doing anything unusual.)

A process that can be specified precisely by rules defines a computable function. Adopting Church's thesis, we can identify the computable functions with the inputs and outputs of programs on Turing machines. Other formalisms could be used; all that matters for the arguments here is that it is possible to list the set of all programs in a natural way, assigning to each of them a so-called Gödel number, or *index*, based on its place in the list. This enables us to use numerals to refer to programs, so that predicates on integers can be interpreted as referring to programs. We use the Greek letter Φ to denote the standard enumeration of computable (partial) functions, so that $\Phi(e,x)$ denotes the result of running program e on input x. We write $\Phi(e,x)\downarrow$ ('$\Phi(e,x)$ converges') to denote that program e outputs a value, or halts, on input x. Similarly, we write $\Phi(e,x)\uparrow$ ('$\Phi(e,x)$ diverges') to indicate that the computation keeps running forever and fails to output a value. Much of the discussion will concern ways of detecting such divergence.

We can also list descriptions of the computations (*i.e.*, sequences of machine states) which occur in running programs on different inputs. Because we can Gödel number these descriptions in a computable way, we can define the T-predicate due to Kleene: $T(e,x,k)$ means that $\Phi(e,x)\downarrow$ and k is (the index of) the computation. So $\Phi(e,x)\downarrow$ if and only if $\exists k(\mathrm{T}(e,x,k))$. It is possible to show that T is itself a computable relation — in other words we can easily define a Turing machine which decides for any three numbers e, x, and k whether or not $\mathrm{T}(e,x,k)$ holds. This is an important fact, since it means that in some sense T is a universal computable relation: for any program e and input x, given $\mathrm{T}(e,x,k)$, we can recover the output of the e-th program on x by just looking at the last state listed in k. Another important concept is that of (*computable*) *enumerability*. Any relation like $\exists k(\mathrm{T}(e,x,k))$ is said to be computably enumerable, since a machine can be specified which enumerates all the pairs $\langle e,x\rangle$ such that $\exists k(\mathrm{T}(e,x,k))$.

Here then is the purely mathematical part of Penrose's argument, taken from Turing, but essentially following ideas due to Gödel. Suppose $\mathrm{A}(e,x)$ is any computably enumerable relation which determines correctly whether or not computations fail to halt. So

(1) $$\forall e \forall x(\mathrm{A}(e,x) \Rightarrow \Phi(e,x)\uparrow)$$

Now consider the *diagonalization* of A, i.e., the relation $\mathrm{A}(x,x)$. Clearly, from (1),

(2) $$\forall x(\mathrm{A}(x,x) \Rightarrow \Phi(x,x)\uparrow)$$

Since A is computably enumerable, it is straightforward to check that $\mathrm{A}(x,x)$ is as well; and being a computably enumerable relation of one variable, it must have an index in the enumeration of all such programs. In other words, there is some n_0 such that

(3) $$\forall x(\mathrm{A}(x,x) \Leftrightarrow \Phi(n_0,x)\downarrow)$$

Now, however, consider a second diagonalization by setting x to be n_0 . If $\mathrm{A}(n_0,n_0)$, then, by (2), $\Phi(n_0,n_0)\uparrow$, which yields $\neg\mathrm{A}(n_0,n_0)$ by definition of n_0. That is, $\mathrm{A}(n_0,n_0)$ implies $\neg\mathrm{A}(n_0,n_0)$; so $\mathrm{A}(n_0,n_0)$ must be false, i.e., $\neg\mathrm{A}(n_0,n_0)$; which, by (3), is equivalent to $\Phi(n_0,n_0)\uparrow$.

To this point, we have been careful not to introduce anything the least bit controversial. So we feel comfortable stating what we know so far as a well-known theorem.

Theorem (Turing, after Gödel)
Let A be any computably enumerable relation such that $\forall e \forall x(A(e,x) \Rightarrow \Phi(e,x)\uparrow)$. If n_0 is the index of A in the enumeration Φ, then $\Phi(n_0,n_0)\uparrow$ and $\neg A(n_0,n_0)$.

Now, why does this theorem seem so significant to anti-computationalists like Penrose and Lucas? Consider a human mathematician — call him Roger — who can understand this argument clearly. Suppose some computably enumerable relation R were to determine correctly whether or not computations fail to halt in exactly those cases in which Roger himself can prove that the computation does not halt. If there were such an R, then its diagonalization would certainly have an index, n_0, in which case, by the theorem just stated, $\Phi(n_0,n_0)\uparrow$ and $\neg R(n_0,n_0)$. But now, since Roger knows the theorem, he knows that $\Phi(n_0,n_0)\uparrow$ is true. This seems to show that he can prove that $\Phi(n_0,n_0)\uparrow$, but that R cannot determine this; so R cannot, contrary to assumption, have Roger's power to prove computations not to halt. Thus the assumption that Roger's knowledge is computably enumerable leads to a contradiction; and so the computationalist hypothesis is false. QED.

This is what Penrose calls the "bare Gödel argument," used in [7]. Later, in [8], it was modified to refer not to any particular human's ability, but to the accumulated wisdom of the entire human race, i.e., the set of all humanly-accessible methods; and his conclusion then is that no algorithm can fully encompass the set of *all humanly-accessible methods for ascertaining mathematical truth*. We will discuss this weaker claim later.

Turing's proof is closely related to Gödel's original theorem, which referred to provability in arithmetic rather than detecting non-termination. Both of these are related in turn to the liar paradox, which arises when we allow sentences to refer to their own truth. An informal account of Gödel's theorem is obtained by modifying the liar sentence to refer to its own provability, rather than its truth. The liar sentence then becomes an intuitive version of Gödel's true-but-unprovable sentence.[2] This familiar observation emphasizes how easily a careless use of the argument may produce a paradox rather than a conclusion, and how much care must be exercised in reasoning "intuitively" in this area. Intuition does not make the liar paradox go away. As we will show, the concept of "humanly knowable methods" leads directly to paradox.

3. TWO OBJECTIONS TO THE ARGUMENT

Many objections can be given to Penrose's argument, but we will focus here on the two mentioned above which were first made by Benacerraf [1] against the argument as originally presented by Lucas [6]. Similar points have been made by other reviewers of Penrose's books [2], although we will develop the objections in a rather different way to show how they are closely related.

The first objection concerns knowledge of consistency. For *us* to be able to prove that $\Phi(n_0,n_0)\uparrow$ as above, and hence to know this, *we* must also prove (or know) that $\forall e \forall x(\mathrm{A}(e,x) \Rightarrow \Phi(e,x)\uparrow)$. That is, we have to know that A is a sound method for deciding that our own computations fail to halt. An alternative resolution of the contradiction is therefore that a human thinker like Roger (or, perhaps, the entire human race) is unable to establish the soundness of the algorithm A that embodies his mathematical abilities. Penrose apparently considers this ridiculous, but we will show that it is more plausible than it might at first seem.

Naturally, the force of this technical objection relies on the human notion of soundness being the same sort of thing as that of a machine, namely something describable in a computable way. Penrose, it seems, believes that we have access to an intuitive notion of soundness which avoids the limitations of the machine notion, yet can still be used freely in reasoning about more precisely specified notions of soundness. This is the reason for our discussion of what we have called the second objection to Penrose's claims. We wish to more fully exhibit the fact that any such argument applies just as well to human thinkers as it does to machines. It uses the "formality" of the system only to establish that the set of algorithms is enumerable; but, as we point out below, the set of possible mathematical sentences, and even the set of possible mathematical insights, is also enumerable in the required sense. The last step in Penrose's argument can then be taken in one of two ways. If one insists that Roger really does have this insight which is forbidden to him, we simply have an epistemic version of the liar paradox. An alternative conclusion, however, is that the act of understanding the Turing proof may not in itself constitute having a mathematical proof of the unprovable sentence. We will return to this point later.

3.1 Knowing Oneself to be Sound

Penrose knows about the first objection, yet he has failed to adequately answer it. When this reply was first made to him in [2], his response (p. 693) was to claim that what was known by us and not in principle knowable by the machine A was not the fact that $\Phi(n_0,n_0)\uparrow$, but rather that *if* A is accepted as sound, *then* $\Phi(n_0,n_0)\uparrow$ must be accepted as well.

There is a clear distinction between the unprovable sentence $\Phi(n_0,n_0)\uparrow$, and the claim that this follows from the soundness of the algorithm, which is simply the assertion that the relation A is an

[2] Gödel's technical achievement was to show how this intuition could be mapped into arithmetic, so that the entire apparatus of grammatical well-formedness and derivability could be recast as arithmetic truths. In modern terminology, we might say that arithmetic supports surprisingly rich opportunities for creative hacking.

accurate predictor of non-termination, i.e., the sentence $(\forall e \forall x(\mathrm{A}(e,x) \Rightarrow \Phi(e,x)\uparrow))$:

(4) $\Phi(n_0,n_0)\uparrow$

(5) $(\forall e \forall x(\mathrm{A}(e,x) \Rightarrow \Phi(e,x)\uparrow)) \;\Rightarrow\; \Phi(n_0,n_0)\uparrow$

In several passages Penrose seems to confuse the meanings of these formulae. For example, he claims that A cannot possibly have access to (5), whereas we can, since we have proved it. However, Penrose is quite wrong here. There is absolutely no reason why A cannot incorporate (5) — in fact, (5) is a theorem of Peano Arithmetic for any computably enumerable relation A, and many computably enumerable sets include the theorems of Peano Arithmetic. His discussion of this issue continually confuses these two claims, using (some locution equivalent to) "$\Phi(n_0,n_0)\uparrow$" when describing what the machine A would not be able to determine and using "if A is sound, then $\Phi(n_0,n_0)\uparrow$" when describing what we are able to determine. For example, in [8] he argues that we must assume, as a prerequisite to any rational discussion, that the sum total of humanly accessible mathematical knowledge is consistent, and therefore that if it could be captured by an algorithm A, that algorithm would be sound. We do not find this plausible, but notice that even if one were to accept this conclusion, it would not constitute a mathematical *proof* of the soundness of A. In order to reach the contradiction required by the bare Gödel argument, it is not enough for Roger merely to believe that he is consistent: he must also believe that this can be established mathematically, i.e., that he knows it to be a *provable* fact. Matters of faith are not considered to be consequences of our mathematical methods for establishing truth.

The distinction between (4) and (5) is clear and can be rigorously established. The analogous distinction made in English is less clear, unfortunately. The ideas of truth and consistency *seem* to be intuitively clear and sharp. However, this is an area where it is notoriously difficult to trust one's intuition. It is easy to explain how we might get an informal notion of soundness or truth which avoids trouble most of the time. We come across statements made in various contexts; looking at these statements from "outside" as it were, and comparing them with our perceptions and intuitions, we count some as "true." This involves understanding the meaning of the statements, or knowing what the intended model is to which they are to be compared. The standard way of discussing these matters is to adopt this point of view, speaking about a language and its model, adopting the standpoint of a meta-language in which to make assertions about truth in the model. As long as one argues in this way, no trouble arises. No inconsistency arises from believing many statements like $\forall e \forall x(\mathrm{A}(e,x) \Rightarrow \Phi(e,x)\uparrow)$, *as long as A is not oneself*, and so one gets the feeling that one has a general idea of what "soundness" means which can even be applied to one's own utterances in a reasonable way; but there an error lurks.[3]

One can show quite rigorously that Penrose's notion of what it is to know oneself to be sound cannot itself be sound. The computable analogue of believing that the procedure indexed by e is sound is, in the context of the diagonal arguments we have been considering, to use the formula $\forall x(\Phi(e,x)\downarrow \Rightarrow \Phi(x,x)\uparrow)$ to make decisions by merely deducing first-order consequences from this formula. This procedure defines a computable function, f which takes a program index e to the index $f(e)$ for the program which uses first-order logic and the soundness assertion for procedure e to make decisions about which programs halt on their own arguments. This is a more precise description of what Penrose refers to as "automating Gödelization." We might think of f as being the formal analogue of understanding the meaning of "e is sound," since it represents the notion of asserting what follows from e's soundness. So, for a sound e, $f(e)$ gives a sound procedure, while

[3] To illustrate the error, consider the fact that anyone can stand on anyone else's shoulders, but no one can stand on their own shoulders.

for an unsound e, $f(e)$ gives an unsound procedure. If we look at f in this fashion, we can get a clear idea of what it would mean to use one's understanding of the meaning of soundness in one's reasoning.

Now, by a classical result due to S.C. Kleene, the (Second) Recursion Theorem, every recursive function like f which operates on program indices has a fixed point. Applying this result to f, we obtain a number e_0, such that the program indexed by e_0 is the same as the program indexed by $f(e_0)$. In other words, e_0 is just the sort of thing for which asserting that something follows from its soundness is the same as merely asserting that thing; we might call it a *Penrosian ideal*. However, the diagonal argument used in the basic proof shows that e_0 cannot really work in the way that it should. For, the program indexed by $f(e_0)$ converges on any input y for which the axiom "$\forall x(\Phi(e_0,x)\downarrow \Rightarrow \Phi(x,x)\uparrow)$" can be used to prove that $\Phi(y,y)\uparrow$). A particular consequence of this axiom is, of course, that $\Phi(e_0,e_0)\downarrow \Rightarrow \Phi(e_0,e_0)\uparrow$, from which it follows that $\Phi(e_0,e_0)\uparrow$. So, $\Phi(f(e_0),e_0)\downarrow$, by f's very definition. Recall, however, that $f(e_0)$ and e_0 index the same program, since e_0 is the fixed-point of f obtained by the Recursion Theorem. Hence $\Phi(e_0,e_0)=\Phi(f(e_0),e_0)\downarrow$. In other words, $f(e_0)$ indexes a program that is unsound, since it is wrong about itself, asserting of itself that it diverges on e_0, when in fact, it converges on e_0.

The algorithm need not be explicitly inconsistent, since this would involve being able to *prove* that $f(e_0)$ and e_0 yield the same output when applied to e_0. If we allow $f(e_0)$ to use Peano Arithmetic in its proofs from the soundness of e_0, then this would follow from the constructive nature of the proof of Kleene's theorem. In fact, however, we merely allowed f to construct programs which use elementary logic, as well as the minimal amount of arithmetic needed to define the notions $\downarrow$, $\uparrow$, and the enumeration Φ. Obviously we need some weak theory like this merely to express the limited idea of soundness which we are using here.

The point of involving the Recursion Theorem in this discussion is to make it clear that any precise use of even a very limited notion of soundness is bound to lead to a procedure which is unsound if applied too generally, in particular to itself. To repeat: all that f uses is first-order logic and enough arithmetic to have a language in which to express the concept of soundness. A more expressive language will result even more quickly in an unsound procedure, and probably an inconsistent one.

What we have done here is take Penrose's intuition, describe it formally — that is, mathematically — and show that this process inevitably leads to a contradiction. Now, it could be argued that this makes Penrose's case seem as strong as possible, since it shows that no consistent procedure can in any sense know that it is sound. Since, according to Penrose, we humans must assume that we do have such knowledge, this fact shows the superiority of informal intuition to formal definition. However, we believe such a position to be at odds with the very nature of science in general, and mathematics in particular. The whole purpose of introducing formal methods is to avoid the contradictions which arise from using "obvious" facts about our natural, intuitive notions. "Formal" here simply means "precise and unequivocal." The formal difficulty applies just as inevitably to the intuitions which inspired the formal description, if those intuitions are precise enough to be regarded as mathematical in nature.

Rather than merely stamping our feet, while maintaining loudly that precise definitions are better than ineffable insight, we prefer to point out how it is that our minds are in essentially the same position with respect to the truth as machine procedures are. For it turns out that even the stongest kind of limitation — that given by the recursion theorem — applies to us as well.

3.2 Knowing Oneself to be Consistent

At first glance, this may appear implausible, and it is this very implausibility which Penrose relies on to give his argument force. But how implausible is the claim that mathematicians reason effectively as machines, even given the limitations imposed on machine reasoning by the Gödel/Turing theorem? If one supposes the output of program A to be identical with that producible in principle by some particular mathematician, say Evariste Galois, and then applies the theorem, one merely has the result that there is some number n_0 which Galois could never have proved his methods would not halt on, yet for which in fact his methods would never halt. This claim might seem to involve a certain amount of hubris, since it appears to say that anyone who has read the arguments above has access to a mathematical insight that Galois could never have had, but this is really a misinterpretation of the situation. While there is no formal contradiction in another mathematician asserting Galois' methods were sound, there is one in Galois claiming to be able to prove that his own methods are sound. It is quite possible that Galois himself may have come to realize that this number n_0 existed, to see the potential contradiction lurking, and to realize that for him to claim to be able to mathematically establish his own consistency would be contradictory, and why. How this is supposed to prove that Galois' methods are non-algorithmic is a little mysterious.

There is really nothing implausible about even a genius like Galois committing just this kind of inconsistency. We can easily imagine him writing down some computable list of truths about numbers, including some version of the induction scheme, claiming that everything he believed about numbers was based on this well-behaved set of truths, and then asserting that therefore, of course, his beliefs about numbers were consistent. Taking these actions would themselves be enough to make Galois a believer in an inconsistent theory. In practice, the actual derivation of the inconsistency would have been very unlikely to ever happen, unless Galois, anticipating Godel's insight, had happened to notice the key fact that statements about numbers can be encoded as numbers.

Interestingly, when this fact was noticed, the very two logicians whose results we have been considering drew quite different conclusions from it. Gödel accepted the possibility of a machine which might be equivalent to human mathematical intuition, but which we could never prove to be sound. Turing, however, according to Penrose, actually argued that the theorem indicates that human mathematical intuition is bound to be unsound, since we do in fact believe that we know ourselves to be sound. In technical terms, this would show us to be unsound machines of precisely the self-referential type guaranteed to exist by the recursion theorem. Since machines can be explicitly designed to be consistent, whereas human brains are not the product of such conscious design, there might indeed be some sense in which our minds are necessarily different from the procedures of *consistent* machines, but equivalent to inconsistent ones. This claim may seem outrageous at first glance, and Penrose evidently takes it to be. He describes Turing's view, with what is obviously intended as ironic understatement, as being one which "many might regard as a somewhat implausible standpoint." However, the ease with which the diagonalization proof technique can extract contradictions from ideas in Penrose's own writings might persuade the reader that it is more plausible than it might seem. Ironically, Penrose considers almost this argument in [8, pp. 81-82, reply to "Q6"], but apparently fails to notice that his argument is paradoxically self-referential at this point.

The reader should note that this is not merely an objection of the *tu quoque* sort. Penrose's whole aim is to show that minds and machine procedures are different by claiming that minds can produce truths which machine procedures cannot. Our account shows how likely it is that minds are subject to the limitations that affect machines, and so removes the force of his claims.

It is worthwhile to go a little further and examine what these limitations might add up to in practice for a machine, as we have just done in the case of a human mathematician. Consider a proposed mathematical robot whose method for ascertaining truth can be summed up in a program. To keep things simple, assume that this algorithm takes (the numerical encoding of) a sentence and returns a truth-value, so it has one parameter and hence is located somewhere in our enumeration, say number r. This algorithm includes all of the robot's mathematical knowledge, among other things, so that the robot can prove p just when $A(r,p)$. We know there is a truth — $A(r_0,r_0)$ — which our robot is unable to prove.

A real robot, however, unlike a formal system, is capable of learning new truths, either by being told them or by being shown new methods for establishing truth. When it learns a genuinely new truth, its repertoire of truth-ascertaining methods enlarges. Its abilities are then embodied in a new such algorithm, one with a different index.

Suppose therefore that the robot learns the truth $A(r_0,r_0)$ and accepts this as a known fact which can properly be used in mathematical proofs. Now it has learned something, the number r_0 is no longer an accurate index of the robot's own algorithm. If we dramatize Gödel's theorem as a robot being faced with a sentence that says "you cannot know this sentence," we can see that in this case the robot is instead faced with something that says "Before you learned it, you couldn't have known this sentence." But this of course can be believed quite consistently, and one can even consistently understand why it is true. Indeed, this seems to mirror our own human situation quite accurately: as we learn new truths, we are able to see the truth of others; and we can come to understand that we were previously mistaken. Penrose mentions this objection [8, pp.78, 'Q2'], but replies that "a changing algorithm would need some specification as to the rules whereby it actually changes". This misunderstanding could be corrected by reading about machine learning. The learning process itself must be specified, but this does not amount to a specification of what is learned, and the one thing which is central to the discussion here cannot be specified in advance: whether what is learned will always lead to a sound procedure.

The robot (more exactly, the robot's new algorithm) now has a new Gödel sentence, of course, which it is unable to prove, but which (in principle) we could tell it, and so this process might continue. But at each stage, the act of learning the truth of its Gödel sentence is sufficient to change the algorithm which correctly describes the robot's mathematical abilities.

One might object that if the robot, like Roger, fully understands the import of the bare Gödel argument, it would be able to conclude without any further input that a sentence exists which is true but which it is unable to establish; and since this sentence is computable from a specification of A, the robot should be able to discover this true sentence, thereby becoming once more enmeshed in contradiction. Care is needed here, however, for while the sentence is computable, this computation may be beyond the robot's own resources. The argument as presented by Lucas and Penrose allows one to conclude only that a certain sentence *exists* which is true (but unprovable); but this does not amount to a proof of the sentence in question. Such a proof must exhibit the actual sentence and establish its particular truth. And note, it is not enough to simply discover, or be told, one's own Gödel sentence in order to be persuaded of its truth in this way. For this conviction to follow simply from one's grasp of the Turing/Gödel proof itself, one must also know that the discovered sentence is the appropriate Gödel sentence, which requires having available a complete description of one's own algorithmic techniques, knowing that it is, indeed, an accurate and complete description, and knowing that the process used to compute the Gödel sentence from this description of oneself is correct. This requires a very complete kind of self-knowledge, one that seems indeed to be quite beyond the capacity both of AI programs and of human beings: in fact, one conclusion to be drawn from the Turing-Gödel results is that such self-knowledge can never be obtained, since the very act of obtaining it changes one, as it were, into something new. And again, we do not find this to be a surprising or implausible conclusion.

This discussion should make it clear how Penrose's argument cannot in principle show that any individual mathematician's methods are incapable of being performed by a program. Of course, applying the Kleene Recursion Theorem as we did earlier might easily lead to an unsound, and even inconsistent, program that behaves like a human mathematician. This is the point of the imaginary situation of Galois — to show that there is nothing implausible about this.

3.3 The Whole Human Race

It is worthwhile to discuss here in more detail what seems to be a possible escape route for Penrose. As mentioned above, he claims in [8] that if one substitutes for the current algorithm of a mathematical robot, an algorithm consisting of every sound algorithm which the robot might ever use in the future, then the robot is bound to find itself back in a paradoxical situation. For this seems to be an algorithm which we can prove to be sound, yet which the robot can never believe sound without involving itself in contradiction. This route can easily be closed off by merely pointing out that it is evidence that this robot's mind is different from each of ours; nevertheless, it seems troubling, since it looks as though the robot should be able to achieve this insight about his future sound procedures in the same way in which we do. In [8], Penrose takes what is essentially this route, although he states it in the strongest possible way, in order to avoid the easy "different-mind" objection which we have just mentioned.

There Penrose attempts an end-run past this entire discussion by referring not to the mathematical abilities of a single human or robot, but to the abilities of the entire human race throughout history (in fact, throughout every possible history, since the concept of humanly-knowable method includes propositions which never get enunciated by any actual human mathematician, but which *could* have been proven, if the world had allowed any human to actually consider them.) His conclusion then is that this totality of humanly-knowable methods for establishing truth cannot be algorithmic in nature. As we have noted, this is hardly a firm attack on the computationalist hypothesis, but in any case, this reasoning is faulty for a new reason: the central concept of "all the procedures available to human mathematicians" (8, p.73) is incoherent.

Much of Penrose's discussion is devoted to the claim that we have some idea of this method, and that we actually know it to be sound. This last is just the seemingly-obvious claim that if a mathematician uses the correct method to decide some truth about numbers, then what he decides is really true. In the weak context we have been considering, this merely means that if some mathematician decides correctly that some arithmetic problem is unsolvable for a natural number n, then it really is unsolvable for n. This may sound reasonable, but it immediately leads to paradox.

It is quite easy to define a list of all the mathematical problems that might in principle occur to any mathematician. We could, for instance, order mathematicians lexicographically by their genetic codes, and then order their potential ideas alphabetically, or by the time at which they might occur. Since, at least according to Penrose, we know what is meant by "the method mathematicians use to decide mathematical truths," somewhere in this list there must occur the problem of using these correct methods to decide that the n-th problem in the list is unsolvable for the number n. So far, nothing we have said should be the least bit controversial: after all, Penrose himself claims to know something about this method, and even claims that we all know this method to be sound. The list of arithmetic problems quite obviously exists, apart from any claims about which problems are actually on it — there is clearly some fact of the matter about which problems can occur, and there are many ways to list them in a well-defined manner.

It should be clear where this is bound to end up, however. For if this problem of showing unsolvability is the n_0-th problem on the list, then, if mathematicians were to use correct methods to solve the n_0-th problem for n_0, they would show that the n_0-th problem is unsolvable for n_0. Since these methods are correct, this is impossible. So in fact the n_0-th problem must be unsolvable on n_0. But since, as Penrose claims, everyone knows our method to be correct, this argument has already decided correctly that the n_0-th problem is unsolvable on n_0, thereby solving the n_0-th problem on n_0, by definition of n_0. Here we have a paradoxical contradiction derived from no special assumptions at all.

Notice that the reasoning one goes through here is no different from that which we (and Penrose) used to establish the earlier theorem, and in our discussion of the Gödelization procedure *f.* We point this out to make it clear that we nowhere use any incorrect method of inference, or any sort of equivocation or other confusion. The form of this argument is identical to the completely acceptable ones above, yet it leads to an evident contradiction. This paradox arises once we are willing to accept the idea that it is actually meaningful to talk about "the method mathematicians use to correctly decide problems." Like "truth", this phrase simply cannot be given an exact meaning which mirrors in every way its informal use, and assuming that it can leads one into the kind of incoherence familiar from the liar paradox.

The problem here lies in Penrose's use of the notion of the methods available to the entire mathematical community, rather than that of the methods available to an individual mathematician. The claim of those against whom Penrose thinks he is arguing is surely not that the sum total of all the methods in principle available to every mathematician is identical with a particular algorithm, but rather that for any individual mathematician, his methods are identical with a program. Penrose slips between these meanings apparently without noticing. For example, his exposition of the basic argument in [8, pp 73-76] begins by assuming that *A* "encapsulates *all* the procedures available to human mathematicians.." — the *all* is crucial to the argument — but concludes that "Human mathematicians are not using a knowably sound algorithm to ascertain mathematical truth". To see the logical error here, consider that while no saltshaker contains all the salt in the universe, nevertheless saltshakers may contain nothing but salt.

To see why this distinction is important, suppose B is the set of all possible pairs $\langle e,n \rangle$ such that some sound computably enumerable theory proves that $\Phi(e,n)\uparrow$. In other words, the relation determines correctly whether or not computations fail to halt in exactly those cases in which in principle it is possible to construct a sound computably enumerable theory which proves that the computation does not halt. If B were itself computably enumerable then its diagonalization would certainly have an index, n_0 , in which case, by the theorem stated above, $\Phi(n_0,n_0)\uparrow$ and $\neg B(n_0,n_0)$. Clearly, if B were computably enumerable and sound, the theory generated by all "$\Phi(e,n)\uparrow$" such that $B(e,n)$, together with the statement that $\Phi(n_0,n_0)\uparrow$,would also be sound, and computably enumerable by an obvious program. But then, by definition of B, $B(n_0,n_0)$, since it is the sum total of all such sound theories. This contradicts B's soundness (in fact, even its consistency). Hence, B cannot be computably enumerable.

This shows that any computably enumerable relation A, which determines correctly whether or not computations fail to halt, cannot be identical with what in principle is correctly determinable by computations not to halt, since some computation can determine that $\Phi(n_0,n_0)\uparrow$, where n_0 is A's index, and A cannot determine this. It follows that what can in principle be done by computations cannot be identical with any particular computation. At this point, if we argued as Penrose does, we would draw the conclusion that computations cannot be computations — an evident absurdity. This shows plainly how illegitimate it is to move from the conclusion that no machine operation is identical with all of human mathematical understanding to the conclusion that machine operations cannot be identical with the mathematical understanding of particular humans. Clearly, since the method used by any individual mathematician is weaker than the sum total of the methods of the entire mathematical community, the claim that no algorithm can implement this weaker method is a stronger claim than the claim that no algorithm can implement the postulated stronger method.

It should be clear that for Penrose's arguments against AI to have any real force, he needs to establish the stronger claim, since the computationalist hypothesis only claims that the theorems resulting from the correct methods of individual mathematicians can be reproduced by individual machines. This does not imply that the theorems that can be produced using the infinite sequence of all possible correct methods available to individual mathematicians can be produced by a particular formal system or machine. Although Penrose does not make his argument transparently clear at this point, he seems to have been misled by his carelessness with the notions of universal machines and enumerations. A universal machine is one which is capable of producing all computations; the machine that computes our enumeration F is such a machine. What Penrose evidently believes is that because we can define such machines, we could, at least in principle, run the computations which use the individual correct methods through some device like time-sharing in order to have some putative computation of the totality of correct computational methods.

To set the readers mind at ease, we can give a brief technical description of the true state of affairs. Such a universal machine is easily constructed using Kleene's T-predicate, which is computable. The T-predicate is true of three integers i, j and k just when the algorithm with index i, when run with input j, produces the computation with index k. All one need do to produce any computation whatsoever from T is to search for the least k such that $T(e,n,k)$. If such a k is found, one then looks at the final state of the computation encoded in k and reads the result of $\Phi(e,n)$ from that. The relation B described earlier is not partially computable using T, however, as our argument shows. The solution to the apparent contradiction lies in the fact that while each individual procedure that goes to make up B does have a program-index e which can be used by T to run the procedure, it is an unsolvable problem to determine exactly which numbers e are indices for programs that are sound in the relevant sense. Penrose has missed the crucial point — in order to run this "super-computation," we would have to be able to give a method for distinguishing the sound methods from the unsound ones, and this is in fact an unsolvable problem[4].

3.4 On Knowing Ourselves

At this point, we have shown as plainly as possible that Penrose's arguments do not have the kind of force which he attributes to them. He has not produced unassailable mathematical proofs of his claims, but rather arguments of an ordinary kind which some may find more or less plausible. Notice that our second point against Penrose, that humans seem to be subject to the same limitations that machines are, has been developed to a point where it can be viewed as another aspect of the first point --- that is, that the extra assumption of soundness is needed to force the algorithm under consideration into inconsistency. With our stories about Galois and the robot, we have tried to show that it is plausible that humans might use a program that is either (1) unsound or (2) not known to be sound. In fact, the conclusion of the previous discussion could be summed up in the fact that no matter how implausible these alternatives might both seem to be, we have to accept one of them, since to deny them both leads to a paradox.

Some readers may feel that some error must have been committed in the discussion above — after all, it is surely just common sense that reasonable people can perceive that they are being reasonable,

[4]This essentially follows from the paradoxical argument about B above. This should be clear, since if the problem were solvable, the paradoxical B above would be explicitly definable. The fact can, of course, be proved directly using the same sort of dialgonalization we have been using throughout.

and so the reduction of this to a paradox must be based on some sort of intellectual sleight-of-hand. The actual explanation for this common sense belief that we know precisely what we are doing when we practice mathematics is, however, nothing particularly magical, and, although the process may be a little involved, analyzing this issue in some detail should alleviate the feeling that there is some trickery here that needs to be explained.

Penrose discusses these matters in the third chapter of [8]. He distinguishes three possible viewpoints that one who believes that some algorithm underlies mathematical understanding might take:

(i) The algorithmic procedure A that underlies mathematical understanding is consciously knowable, and that it is the algorithm underlying mathematical understanding is knowable.

(ii) The algorithmic procedure A that underlies mathematical understanding is consciously knowable, but that it is the algorithm underlying mathematical understanding is not knowable.

(iii) The algorithmic procedure A that underlies mathematical understanding is not consciously knowable, and the fact that it is the algorithm underlying mathematical understanding is unknowable.

Penrose ignores the remaining logical possibility, that some algorithmic procedure A that underlies mathematical understanding might be unknowable, but the fact that it is indeed the algorithm underlying mathematical understanding is knowable. He has been taken to task for this by Putnam in [10], correctly, in our view.

Penrose argues against case (i) as follows (p. 131): If such an A exists, it is clear that one must believe that A is a sound algorithm, since it is known to underlie mathematical understanding, which is believed to be sound. By the theorem above, the soundness of A entails that some computation $\Phi_F(n_0,n_0)$ fails to halt, and A can never determine that $\Phi_F(n_0,n_0)$ fails to halt. Since the theorem is believed, and the soundness of A is believed, it must also be believed that $\Phi_F(n_0,n_0)\uparrow$ and that A can never determine this truth, all of which contradicts A being believed, much less known, to be the algorithm underlying mathematical practice.[5] However, it is important to realize that to believe a mathematical statement to be true is not the same as to believe that one can establish it to be true by mathematical methods. To give a real-world example, most mathematicians believe the axiom of choice to be true, yet few believe it to be possible to establish its truth by mathematical methods. The situation is exactly the same here. Although one might know that A underlies mathematical understanding, and believe that A is sound, this is a long way from believing that one can *show* that that A is sound, which is clearly what it would take to know that $\Phi_F(n_0,n_0)\uparrow$.

Penrose's tendency to make this mistake is indicated quite clearly by his confusion between the sentences (4) and (5). His argument is based on the claim that we must assume that we are consistent, for if we are not, then the foundation of knowledge will fail and the walls of science will tumble. Even if one accepts this plea, however, it hardly amounts to a mathematical argument for our own consistency. It has more the flavor of a cry of faith than an explanation of truth.

3.5 What Does it Mean to be Formal?

[5]Penrose expanded this argument in more detail in his Psyche reply to Chalmers, being careful there, as we discussed above, to emphasize that the "mathematical understanding" referred to was essentially "all mathematical truth that is knowable in principle." We have already pointed out the technical error in this move.

Here is the place to emphasize the difference between believing that what one says is true and treating “what I say is true” as an axiom. The Gödel arguments show us that we cannot use this belief, that what we think is true, explicitly in an argument formalizable in arithmetic. It is a belief of a very strange and special kind, involving notions like “I” and “true” which no one has given any convincing explanation of how to use in reasoning. In fact, all we know is that reliance on ordinary logic using the unrestricted concept of truth leads directly to paradox.

While this is surely now well established, we often ignore it and refer to “truth” as though the concept were clear. And Penrose believes that it is. He believes that mathematics describes a real, external, objective Platonic realm of mathematical abstractions. On this view, the notions of truth and consistency have a meaning which quite transcends any local attempt to define or formalize them. Our beliefs alter and change, and they may only be an approximation to this single, absolute truth which we all perceive more or less clearly and, as mathematicians, strive to see more clearly; but the subject matter of mathematics simply *exists*, abstract and eternal. But now, if one's assertions and beliefs are to be judged by how close to this objective Platonic world one can make them, what can be made of someone who asserts, or even allows the possibility, that their own beliefs are inconsistent? From this perspective, such a position would seem to be an extreme kind of abdication of mathematical responsibility, since if one is inconsistent then conformity to any kind of reality is evidently impossible. One can see, therefore, how Penrose would be led to claim that to assert something, and to assert that it follows from one’s own consistency, are essentially to say the same thing, confusing (4) with (5). From his perspective, all meaningful mathematical assertions *must* be made within a framework which includes the assumption of the consistency of the speaker's beliefs.

This uncritical version of epistemological Platonism claims that any mathematical statement which one proves can be known by direct mathematical intuition. To make this claim plausible, Penrose [8, p. 68] gives an example concerning "hexagonal" numbers, which amounts to showing a few pictures and making some remarks about them in order to establish the fact that for every

$$n \geq 0, \sum_{k=0}^{n} (1 + 6\sum_{j=0}^{k} j) \text{ is } (n+1)^3.$$

The point seems to be that one can grasp this algebraic fact by the “informal” means of visualization, which indicates that formalization is not necessary for mathematical understanding. In Penrose's words, “part of my purpose here is to show that there are sound methods of mathematical reasoning that are not 'formalized' according to some preassigned system of rules.” What Penrose misses here is that whenever one gives a precise description of what mathematicians are saying in a particular instance, one does effectively produce some formal rules. Calling an argument “formalizable” amounts to no more than calling it “precise and unambiguous.” Even such directly intuitive graphic arguments as his hexagonal number illustration are quite within the scope of computational formalization. (Ironically, an active subarea of AI has recently emerged concerned precisely with the mechanization of such directly geometric intuitive processes. See, for example, the comprehensive review [3].)

Although Penrose is clearly offended by the idea of a mathematical insight (specifically, his own) being circumscribed and limited by a “preassigned” system of rules, the Gödel argument says nothing about *pre*assignment. All that is needed for the argument to go through is the *existence* of a formal system, not its *pre*-existence. (In fact, in Penrose’s own Platonic view of mathematical reality, such things as formal systems, being mathematical in nature, have no temporal aspect to their existence at all.) All it requires to be preassigned is the standard ordering Φ; but this places no constraint upon the formal systems being considered, just as the existence of a standard alphabetical ordering places no constraints at all upon the possible sentences of English. (In order to accommodate the notorious lexical creativity of mathematicians we could even allow a countably infinite alphabet, and still have a predefined alphabetic ordering on all possible sentences.) The only substantive requirement for the proof to go through is that the sentences themselves are finite. But since any mathematician, no matter how creative, can produce only a finite number of symbols in his or her lifetime, this seems to place no real constraint on the undecidability theorem's range of applicability. In particular, if we take any argument offered by Penrose and ask him to make the meanings of the words in it quite clear (even using pictures and geometric intuitions if he wishes) and to make all his assumptions quite explicit, then by doing so he will have “formalized” it sufficiently well to satisfy the conditions of his own argument.

One might reply that there is no limit to the range of mental devices that a mathematician might use to discover truth, and this set of mental insights may not be computable. But this is irrelevant to our point here, and does not allow one to escape the force of the paradox. The formalization only has to reproduce the sentences uttered, or asserted, by the mathematician in expressing mathematical insights; and there is a standing condition on responsible mathematicians to be able to actually express their meaning unambiguously and be willing to reveal their assumptions, if challenged. It will never be enough to claim that something is true simply because its truth has been revealed to one in some inexplicable way, or that its truth will be obvious if the reader is willing to perform a certain ritual. Such claims are regularly made in human discourse, but we do not call this mathematics. And this requirement still applies even if the truth seems obvious to someone, unless that person is able to convey the grounds for his or her confidence to the rest of the human race. Suppose, for example, that someone were to claim that she could simply *see* that the axiom of choice is false. Such insight is not available to the rest of us; so, in order to qualify as mathematics, we would require the claimant to explain and justify this insight, in spite of its being directly apparent to her; and if she were unable to do so, it would not be regarded as having been mathematically established. Unambiguous clarity and explicit declaration of assumptions are all that we require to have a formalization in the sense being used here.

3.6 Intuitive Soundness and Set Theory

While Penrose's arguments are faulty, his conclusions may be consistent with one coherent philosophical position concerning the nature of mathematical truth. While we will not attempt to refute this position, it is worth pointing out just how very idiosyncratic it is, and how ill-fitted to the wealth of technical results in the foundations of mathematics which have been developed during this century.

It seems that what Penrose imagines is something like the following. Through some process akin to the geometric visualization described in his discussion of hexagonal numbers, one can grasp that Peano Arithmetic and various stronger theories are sound, as follows: since we know by intuition that there is a set of natural numbers, and we can perceive directly that each of the Peano axioms is true of these numbers, we know that these axioms are consistent and, in fact, sound. We thus know that adjoining their consistency statement to them also results in a sound theory and, even more, that the whole sequence of all such adjoinings will result in a sequence of sound theories; and this is a statement that none of the first-order arithmetic theories so generated can possibly imply.

The process of making explicit Penrose's reasoning here does not have to be a merely hypothetical one. The work has already been done, since the entire argument Penrose is thinking of in this context can be formalized in Zermelo-Frankel set theory. Since most mathematicians, even if they are not themselves "formalists," are at least willing to use the method of constructing a formal proof of a statement from the ZF axioms to establish a mathematical claim, they apparently believe that they can in principle *know* that all these arithmetical theories are consistent. The problem arises in Penrose's idea that the consistency of set theory itself is known in the same kind of transparent way.

In fact, the wide acceptance of set theory by mathematicians is different from the similarly widespread acceptance of arithmetic in a way that is relevant to this very issue. For what was discovered by Cantor, Russell, and others was just this: that the very intuitions which we are naturally inclined to accept lead directly to contradictions. Nobody has ever given a convincing explanation of why it is illegitimate to refer to the set of all sets. We know that this is illegitimate because of Russell's paradox, but we still lack an intuitive basis for the judgment that it is illegitimate. In fact, it is quite clear that in almost every context where mathematicians reason about sets, they are psychologically using the unrestricted comprehension axiom of Frege, which leads to paradox. Can we really believe that when a working mathematician uses, say, the notion of the set consisting of all the subsets of the set of differentiable real-valued functions on the reals, that this analyst actually goes through the mental process of checking that this set is "small enough" to be an object in the relevant way? And if not, then how plausible can the claim be that mathematicians are actually using sound methods which they *know* to be sound?

The set-theory axioms which encapsulate what we are willing to accept as "unassailable," in Penrose's words, came about precisely because it was recognized that the actual intuitions which mathematicians were using led to inconsistency. After analyzing what was really needed to prove the results which were so dear, the ZF axioms were produced. In what sense can we claim to know that these axioms are consistent? If we follow Penrose, this knowledge would be achieved by something akin to the geometric visualization, which he discusses in the example mentioned. If this is the case, it seems clear that what is really known to be consistent is the finite second-order version of the axioms, in which a single second-order statement takes the place of the infinitely many axioms of the replacement scheme. (This is nothing controversial. The replacement scheme only arises as a technical first-order approximation to the obvious second-order statement, and it seems more likely that what we believe is itself a finite sentence rather than an infinite collection of sentences.)

For Penrose, the arithmetic axioms are known to be consistent because they are seen to be true about the intended structure, the natural numbers. To many, this in itself might seem to be illegitimate, since it is a little unclear how we as finite beings can *see* things to be true about the infinite object consisting of all the natural numbers. However, the fact that we have accepted the axiom of infinity in our set theory gives us at least a means of talking about this infinite object which we believe makes sense, and so this claim might not be too implausible. In the case of the second-order set-theory axioms, the situation is very different. The smallest object which could be a model of these axioms has the size of an inaccessible cardinal.[6] For one to intuit the soundness of the set-theory axioms in a natural way involves having an intuition of the whole universe below an inaccessible cardinal.

Such an intuition would be very different from the intuition that there is an infinite set of natural numbers. All that is involved in the case of the natural numbers is that one should have an idea of

[6]For those who know some set theory: In the case of the first-order axioms, there are, of course, countable models if the theory is consistent, but these cannot possibly be the intended model in Penrose's sense, since the interpretation of "uncountable" in such a model is the wrong one—in set-theorist jargon, the notion fails to be absolute. The same holds for all first-order models below the first inaccessible, essentially because they have the wrong notion of their own cardinality. This is why we have to adopt the second-order axioms to make sense of Penrose's claims. The problem does not arise for the inaccessible, since the existence of the inaccessible cannot be proved in the theory.

the smallest set closed under the relatively simple operation of taking the successor of a number. Because the successor operation is so simple, it seems reasonable to think that mathematicians have a clear idea of everything produced by applying the successor operation successively to zero. In the case of the inaccessible cardinal, it is not the successor operation that is involved, but (among others) the power-set operation. This is a much harder process to grasp intuitively than that of adding one to an integer. To form a complete idea of the set of all subsets of an infinite set, even the set of natural numbers, would involve, for instance, knowing its cardinality. But nobody knows the cardinality of this set. To know it would be to know whether the continuum hypothesis were true or not; and far from knowing this, we know that we cannot know it. One of the significant achievements of logic in this century was Cohen's development of forcing, by which one can show that the power sets of infinite sets can be consistently taken to be almost anything. The method of forcing shows very clearly that even the simplest, most natural, question about the power-set operation cannot be inferred from any current intuitions about the nature of sets which can be expressed in the language of set theory.[7]

Imagine that we were unable to tell how much bigger the successor of a number was than the number itself — even whether or not it was finitely bigger — would we be likely to claim that we had a clear intuition of "the smallest set closed under the successor operation"? The answer is obviously no. Yet this is just the sort of thing we have to accept if we want to claim we have an intuition that set theory has a model.

Notice, however, that Penrose cannot stop even there. Whatever his claims to intuitive mathematical knowledge are, it is clear that something must connect them to the formal notions of consistency and satisfiability; these are, after all, themselves the subject-matter of a part of mathematics. But to claim to have knowledge of the satisfiability of an axiom system like that of set theory, and to claim further that this knowledge is itself about something that can be used in a mathematical argument of the kind we have been considering, is to claim that one has, at the very least, knowledge of formal satisfiability. As pointed out, in the case of set theory, this claim involves the intuition of a very large infinite object. Worse yet, Penrose believes that it is legitimate for him to treat his own soundness as known in the same way other mathematical facts are known. What this self-referential notion of soundness means is that there is no way to stop. For now, the intuition of the inaccessible cardinal, which provides a model for the ordinary set-theory axioms, is itself an intuition of an actual object, and so must be seen as existing in some model of still higher cardinality which cannot be seen to exist from the mere existence of the inaccessible. Then this model too must exist somewhere, and so on and on. Penrose's claim, to the extent that it has any precise meaning, has to be understood as a claim that one can intuit the whole mathematical universe, no matter what that universe contains. And this seems unlikely to be the case. It is one thing to be an epistemological Platonist about the natural numbers, but the claim that we can have direct, unassailable intuitions about the structure of all inaccessible cardinals reflects a degree of self-confidence that few mathematicians would be likely to admit.

Since an inaccessible cardinal can be proved to be what is needed to provide a model for our set-theoretic principles, yet such a thing seems to be "too big" for us to have a clear intuition of it, there might seem to be a mystery as to how we can accept our set theory as true. However, what we actually do here is clear. We apply our set-theoretic principles in ordinary contexts and then make a kind of "leap of faith" that there is actually a model of our natural axioms.

The difference between this and Penrose's optimistic approach to mathematical intuition is precisely in his conviction that this is not a leap of faith, but a mathematical truth which, like all the others, can be established mathematically. But surely the lesson of our century, if it has one, is that human

[7]There are non-intuitive axioms like V=L, or the Proper Forcing Axiom which do settle such questions, but no one believes these statements to be self evident in anything like the way the ordinary axioms of ZFC are thought to be.

thinkers are both more complex and less reliable than our naive intuition tells us. There is no bedrock of mathematical certainty on which the edifice of science must be based, no *direct route* to mathematical Truth. If mathematics is really the empirical science of the Platonic realm, then Popper's difficulties apply to it just as they do to every science. We can never be *absolutely* sure that we have things right, even in mathematics, and still less can we be certain that all truths will eventually be vouchsafed to us. Penrose thinks he knows he is consistent, but he will never be able to establish this, for he doesn't really know it. He only thinks he knows.

4. CONCLUSION

The Gödel/Turing diagonalization argument is a powerful tool: so powerful that it can easily destroy its user if used carelessly. As we have shown, arguments with the identical form and power of those used by Lucas and Penrose can be used to derive quite ridiculous, even paradoxical, conclusions. Somewhere in these arguments one can always find an illegitimate step where two senses of a concept are confused with each other, and meta-mathematics is transformed into philosophy. But as we have emphasized, this area perhaps above all others is one where such transformations must be examined very carefully, as this meta-mathematical reasoning skirts dangerously close to philosophical paradox.

We have located these crucial ambiguities in Penrose's publications, but our case is broader: *any* attempt to utilize the undecidability and non-termination results to attack the computationalist thesis is bound to be illegitimate in this way, since these results are quite consistent with the computationalist thesis. Theorems of the Gödel and Turing kind are not at odds with the computationalist vision, but with a kind of grandiose self-confidence in the notion that human thought has some kind of magical quality which resists rational description. The picture of the human mind sketched by the computationalist thesis accepts the limitations placed on us by Gödel, and predicts that human abilities are limited by computational restrictions of the kind that Penrose and others find so unacceptable. The tension which Penrose and others perceive arises only if one adds further assumptions, often about the nature of truth, human insight, or computation itself, which are already incompatible with the computationalist hypothesis, and indeed often have been explicitly rejected by those working in these areas. Whether or not the hypothesis is considered acceptable, therefore, the heavy-duty meta-mathematics adds nothing to the case being made.

This point deserves emphasis. The Lucas/Penrose argument refers to computations, and so its *reductio* conclusion is stated as a result about computations. It is salutary, however, to ask what “computational” properties it uses. In fact, the only substantive property that computations must have in order for the argument to be made are that they somehow express propositions about integers and that the set of them all is enumerable. These are extremely weak properties, and apply just as well to sentences in a well-defined language, axiomatizations, diagrams, or almost any way in which rational agents could express themselves clearly. In our view, they also apply to human thoughts. Penrose undertakes to escape from this by hypothesizing that conscious human thoughts arise from field collapse in quantum theory and are therefore not enumerable. We do not find this plausible, but mention it here only to emphasize the lengths to which one must go in order to avoid the power of the argument. And as we have mentioned earlier, the set of every unambiguous sentence that a human mathematician could utter is enumerable, so in order to escape the self-referentiality of the diagonal construction, the set of mathematical insights that quantum indeterminacy might provide must apparently include some thoughts that can never be said.

At various stages in his odyssey, Penrose has assumed that human mathematical thought is consistent and knows itself to be; that if we were an algorithm, we would necessarily know what algorithm that was; that if we were each an algorithm, then the sum total of our thinking would also be algorithmically describable, and that being "sound" in some informal sense means the same as being true and consistent. All of these extra assumptions are computationally implausible, and some can be actually refuted within computational mathematics. His books are full of many other notions which are computationally implausible (such as the idea of a well-defined process which takes uncomputable decisions); we mention these only to emphasize that his intuitions are clearly very different from ours. We do not claim to absolutely refute Penrose's conclusions; for, ultimately, the differences between us involve controversial assumptions about deep issues in the foundations of mathematics and the nature of thought. But he has not, and indeed he could not have, *established* them.

5. REFERENCES

[1] P. Benacerraf, God, the Devil, and Gödel, *The Monist* **51** (1967) 9-32.

[2] G. Boolos, and others, an Open Peer Commentary on *The Emperor's New Mind* in *Behavioral and Brain Sciences* **13**(4) (1990) 655.

[3] J. Glasgow, N. Hari Narayan and B. Chandrasekaran, eds, *Diagrammatic Reasoning* (AAAI Press/MIT Press, Menlo Park, CA, 1995).

[4] D. Lewis, Lucas against mechanism, *Philosophy* **44** (1969) 231-233.

[5] D. Lewis, Lucas against mechanism II, *The Canadian J. of Philosophy* **9** (1989) 373-376.

[6] J.R. Lucas, Minds, machines, and Gödel, *Philosophy* **36** (1961) 120-124.

[7] R. Penrose, *The Emperor's New Mind* (Oxford University Press, New York, 1989).

[8] R. Penrose, *Shadows of the Mind* (Oxford University Press, New York, 1994).

[9] R. Penrose, discussion of *Shadows of the Mind* in *Psyche* **2**(23) (1996) online filename: psyche-96-2-23-shadows-10-penrose.

[10 H. Putnam, Review of *Shadows of the Mind* in *Bulletin of the American Mathematical Society* **32** (1995) 370-373.