A Refutation of Penrose's Gödelian Case Against Artificial Intelligence*

Selmer Bringsjord & Hong Xiao
Dept. of Philosophy, Psychology & Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180 USA
selmer@rpi.edu • http://www.rpi.edu/∼brings

February 23, 2000

Abstract

Having, as it is generally agreed, failed to destroy the computational conception of mind with the Gödelian attack he articulated in his *The Emperor's New Mind*, Penrose has returned, armed with a more elaborate and more fastidious Gödelian case, expressed in Chapters 2 and 3 of his *Shadows of the Mind*. The core argument in these chapters is enthymematic, and when formalized, a remarkable number of technical glitches come to light. Over and above these defects, the argument, at best, is an instance of either the fallacy of denying the antecedent, the fallacy of *petitio principii*, or the fallacy of equivocation. More recently, writing in response to his critics in the electronic journal *Psyche*, Penrose has offered a Gödelian case designed to improve on the version presented in *SOTM*. But this version is yet again another failure. In falling prey to the errors we uncover, Penrose's new Gödelian case is unmasked as the same confused refrain J.R. Lucas initiated 35 years ago.

*We are indebted to Martin Davis, Kelsey Rinella, Marvin Minsky, David Chalmers, Jim Fahey, Michael Zenzen, Ken Ford, Pat Hayes, Bob McNaughton, and Kostas Arkoudas. Selmer would like to express special thanks to Roger Penrose for debate and conversation concerning many of the issues treated herein. Bringsjord and Penrose both believe that the mind is beyond computation; both *also* believe that Gödelian results can be deployed to demonstrate this. However, as this paper testifies, they differ over how to carry out the demonstration.

Table 1: Some of Bringsjord's Arguments Against "Strong" AI.

Argument	**Reference**
Argument from Creativity	Chapter 5 in (Bringsjord and Ferrucci 2000)
Argument from Irreversibility	(Bringsjord and Zenzen 1997)
Argument from Infinitary Reasoning	(Bringsjord 1997*b*)
Argument from Mental Imagery	(Bringsjord and Bringsjord 1996)
Argument from Free Will	Chapter VIII in (Bringsjord 1992)
Argument from Introspection	Chapter IX in (Bringsjord 1992)
Argument from Possibility of Zombies	(Bringsjord 1999)

1 Introduction

Those who go in for attacking artificial intelligence (AI) can be indefatigable. John Searle tirelessly targets at least one brand of AI ("Strong" AI) with variations on his Chinese Room (he fired the first shot in (Searle 1980)). One of us (Bringsjord) has at last count published 13 formal arguments against "Strong" AI. (For a sampling, see Table 1.) And Roger Penrose appears to have the same endless energy when it comes to producing Gödelian attacks on AI: Having, as it is generally agreed, failed to improve on Lucas' at-best-controversial primogenitor (Lucas 1964) with the argument as formulated in his *The Emperor's New Mind* (*ENM*) (Penrose 1989),[1] Penrose has returned, armed with a new Gödelian case, expressed in his *Shadows of the Mind* (*SOTM*) (Penrose 1994). This case, unlike its predecessor, does more than recapitulate Lucas' argument, but it nonetheless fails, as we shall see. The great irony is that this case is based on Penrose's near-deification of logico-mathematical reasoning, but such reasoning, as we show herein, can be used to refute Penrose's case.

The heart of *SOTM*'s Chapter 2 is a diagonal argument designed to show that there is no "knowably sound" algorithm for classifying computations as non-halters. About this diagonal argument Penrose says:

> Admittedly there is an air of the conjuring trick about the argument, but it is perfectly legitimate, and it only gains in strength the more minutely it is examined. ((Penrose 1994), p. 75)

Unfortunately, we *have* examined the argument minutely, and Penrose is stone cold wrong: at best, it's enthymematic, and when formalized, a remarkable number of technical glitches come to light. Over and above these defects, the argument, at best, is an instance of either the fallacy of denying the antecedent, the fallacy of *petitio principii*, or the fallacy of equivocation.

In Chapter 3, Penrose (working under the assumption that the argument of Chapter 2 is sound) tries to rule out the remaining possibility: viz., that there *is* an algorithm, but not a knowably sound one, for classifying computations as non-halters. Here again he fails — and once more the problems are formal in nature.

More recently, writing in response to his critics in the electronic journal *Psyche*, Penrose has offered a Gödelian case designed to improve on the version presented in *SOTM*. But this version is yet again another failure.

In falling prey to the errors we uncover, Penrose's new Gödelian case is unmasked as the same confused refrain J.R. Lucas initiated 35 years ago.

Our plan herein is as follows. In section 2 we set out the main foundational positions on AI, chief among which are Penrosean versions of "Strong" and "Weak" AI. In section 3 we explain why "Weak" AI is pretty much invulnerable, and therefore Penrose's target should be "Strong"

AI only. In section 4 we review the mathematical background presupposed by Penrose's Gödelian case. Section 5 covers the core diagonal argument for this case. In section 6 we review the formal machinery we use to expose the invalidity of this diagonal argument, in section 7 we formalize this argument, and in section 8 we explain why the argument is fallacious. Section 9 is devoted to considering and rebutting replies on behalf of Penrose. In section 10 we show that even if the diagonal argument is sound, Penrose's overall case fails. In section 11 we give Penrose a last chance: we consider a version of his Gödelian case that he gave in the electronic journal *Psyche* in an attempt to improve upon the version featured in *Shadows of the Mind*. We sum things up in section 12, and briefly consider there the future of Gödelian attacks on computationalism, including, specifically, the recent claim (LaForte, Hayes and Ford 1998) that no such attack can possibly work.

2 The Main Positions on AI

Penrose begins by setting out in propositional form what he sees as the four fundamental positions on AI (p. 12, (Penrose 1994)):

$\mathcal{A}$ All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.

$\mathcal{B}$ Awareness is a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness.

$\mathcal{C}$ Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.

$\mathcal{D}$ Awareness cannot be explained by physical, computational, or any other scientific terms.

$\mathcal{A}$ is intended to encapsulate so-called "Strong" AI; put in terms of future robots and the *Total* Turing Test,[2] the thesis boils down to the claim that robots able to pass TTT will arrive, and will moreover have full-blown conscious mental states. $\mathcal{B}$ is supposed to encapsulate "Weak" AI; again, put in terms of future robots, the idea is that TTT-passing robots are headed our way, but despite impressive *behavior*, they will lack consciousness: they will be zombies.[3] $\mathcal{C}$ is Penrose's position; and $\mathcal{D}$ is what Penrose calls the "mystical" stance, one apparently affirmed by, among others, Kurt Gödel.

Penrose's four-fold breakdown, upon reflection, is disturbingly imprecise. The main problem is that $\mathcal{B}$ needlessly unites *three* seemingly separate claims, viz.,

$\mathcal{B}_1$ Awareness is a feature of the brain's physical action.

$\mathcal{B}_2$ Any physical action can be simulated computationally.

$\mathcal{B}_3$ Computational simulation cannot by itself evoke awareness.

Some thinkers deny $\mathcal{B}_1$ but affirm $\mathcal{B}_2$. Indeed, Bringsjord is such a specimen. He is at present agnostic on whether or not substance or property dualism is true (and hence agnostic on $\mathcal{B}_1$), but he wholeheartedly affirms $\mathcal{B}_2$. It seems possible that someone could coherently hold to

$$\mathcal{B}_2 \wedge \neg\mathcal{B}_3 \wedge \mathcal{A}$$

as well, but in the interests of economy we leave this possibility aside. As to $\mathcal{B}_2$ itself, this thesis entails

$\mathcal{B}_2^M$ Physical action relevant to mentation can be simulated computationally.

But there well might be thinkers who affirm $\mathcal{B}_2^M$ but reject the stronger $\mathcal{B}_2$.

Here and hereafter let's identify "Weak" AI with $\mathcal{B}_2^M$. Part one of *SOTM* is a sustained argument for

$$\neg(\mathcal{A} \vee \mathcal{B}_2^M).$$

Given this, it follows by propositional logic that if $\mathcal{A} - \mathcal{D}$ exhaust foundational takes on AI, and if the niceties we've noted in connection to $\mathcal{B}$ are ignored,

$$\mathcal{C} \vee \mathcal{D}.$$

If the mystical $\mathcal{D}$ is unacceptable to nearly all scientists and engineers, as in fact it doubtless is (as Penrose points out), we are left with $\mathcal{C}$ by disjunctive syllogism; and Penrose spends the second part of *SOTM* exploring and explaining the "uncomputable" physics needed (given $\mathcal{C}$) in order to explain consciousness. Obviously, if the argument of part one fails, part two is little more than a curiosity.

3 Why "Weak" AI is Invulnerable

Though Penrose focuses on human mathematical reasoning of a most abstract and esoteric sort (witness the example pertaining to hexagonal numbers discussed in the next section), and though the $\mathcal{A}$–$\mathcal{D}$ quartet is itself rather abstract, there is a firm connection between this reasoning and "Strong" AI: If people are computing machines and cognition is computation, then mathematical reasoning, however far removed it may or may not be from "everyday" cognition, must in the end *be* computation. As Penrose points out:

> It might well be argued that the building of a robot mathematician is very far from the immediate aims of AI; accordingly, the finding of such an F [= a theorem-proving machine on par with human mathematicians] would be regarded as premature or unnecessary. However, this would be to miss the point of the present discussion. Those viewpoints which take human intelligence to be explicable in terms of algorithmic processes implicitly demand the potential of such an F ((Penrose 1994), p. 137).

Of course, by *modus tollens* it follows that if no such F exists, AI, at least of the "Strong" variety, cannot be right, that is $\neg\mathcal{A}$. Unfortunately for Penrose, the connection between human mathematical reasoning and $\mathcal{B}_2^M$ is nothing like what he thinks it is; here's why.

The problem for Penrose is that a machine might *appear* to be doing all sorts of mathematical proofs of the type that Penrose venerates, and yet might be doing so on the strength of "mindless" simulation. Selmer has such a simulation available to him on the machine he is currently typing this sentence into: this is a simulation, by the theorem prover known as OTTER, of Gödel's first incompleteness theorem (Gödel I).[4] Selmer can run OTTER and after a bit of time, bingo, (an encoded version of) this theorem is proved and printed. The important point to realize is that this simulation has none the mental states Gödel instantiated when he carried out his famous proof. For that matter, the simulation has none of the mental states logic instructors like Selmer instantiate when they prove Gödel I for their students. (For details on machine proofs of Gödel I see (Bringsjord 2001).)

We can bring the point here directly against Penrose, as follows. (This objection is one Searle has ingeniously articulated as well, in slightly different form: (Searle 1997).) The key phenomenon for Penrose, the one he believes to be beyond computation, is that of a mathematician "ascertaining mathematical truth." As an example consider this proposition:

SUM The sum of two even numbers is always an even number.

When a mathematician attempts to ascertain whether SUM is true, he or she is attempting to decide whether or not a certain computation will ever halt. What computation? This one:

$$0 + 2 \text{ odd?}, 2 + 2 \text{ odd?}, 0 + 4 \text{ odd?}, 4 + 2 \text{ odd?}, 2 + 4 \text{ odd?}, 0 + 6 \text{ odd?}, \ldots$$

Obviously, this computation will never halt; knowing just a little bit about arithmetic is enough to grasp this fact. Of course, professional mathematicians would be engaged with propositions rather more challenging than SUM. Let P be such a proposition, and let P_C denote the corresponding computation. Now suppose that Penrose carries out mathematical reasoning over a stretch of time from t_1 to t_{10} which eventuates in his correct declaration that P_C doesn't halt.[5] Assume that we take a snapshot B_{t_i} of Penrose's brain at each t_i, and that each snapshot has a corelate B'_{t_i} in an artificial neural network N that we build to process an encoding of P from t_1 to t_{10}. Suppose as well that N yields the answer "Doesn't halt" at t_{10}. The problem for Penrose is that N, for every proposition like P, can *in fact* be built. This is evident once one realizes that N needn't have any of the actual mental states Penrose has from t_1 to t_{10}. N, after all, is just a *simulation*. From the perspective of $\mathcal{B}_2^M$, *even if* human mathematical reasoning produces a verdict on whether computation C halts via information processing beyond that which is computable (super-Turing processing, e.g.; cf. (Bringsjord 1998), (Siegelmann 1995), (Siegelmann and Sontag 1994), (Kugel 1986)), it is a trivial matter to build a system that yields this verdict through standard computation. (Analogously: Kasparov may for all we know routinely surpass the Turing limit when playing chess as he does, but Deep Blue can still beat him by running standard algorithms.)

The predictable rebuttal on behalf of Penrose is that N here isn't *really* a simulation, because there isn't a sufficiently close correspondence between Penrose's rationcination from t_1 to t_{10} and the states of N through this interval. There is a fatal problem afflicting this rebuttal: the B'_{t_i} can approximate the B_{t_i} to a degree of fidelity that far exceeds what we normally demand in cases of "real world" simulation. To see this, consider the grandest and greatest computational architecture used to simulate human cognition: ACT-R (Anderson 1998). ACT-R is intended by John Anderson to mark the fulfillment of Alan Newell's dream of "a unified theory" of all human cognition.[6] ACT-R is composed of two elementary formalisms and one overarching algorithm, a trio used routinely in AI (fully covered, e.g., in (Russell and Norvig 1994)). The first formalism is a frame-based representation system, which is merely another way of expressing facts in first-order logic. The second formalism is a production system, which is merely, again, a system that allows for conditional reasoning in first-order logic. The most recent version of ACT-R, version 4.0, is set out in a book that explains, in painstaking detail, how this architecture simulates humans carrying out elementary arithmetic (see Chapter 9 of (Anderson 1998).) Chapter 11 of this book is devoted to providing experimental evidence for the view that ACT-R 4.0 can be used to simulate the cognition involved in human scientific discovery. Both simulations involve an exceedingly weak correspondence between real, human cognition and inferencing in first-order logic. Indeed, the correspondence is a good deal weaker than that between Penrose's decision with respect to P_C and the behavior of N.

Interestingly enough, in *SOTM* Penrose does consider a (weaker — because 'simulation' is used in a sense not in play in AI) version of the objection from simulation that we have given immediately above. Here's how Penrose expressed the objection in *SOTM*:

> **Q7. The total output of all the mathematicians who have ever lived, together with the output of all the human mathematicians of the next (say) thousand years is finite and could be contained in the memory banks of an appropriate computer. Surely this particular computer *could*, therefore, simulate this output and thus**

> **behave (externally) in the same way as a human mathematician — whatever the Gödel argument might appear to the tell us to the contrary.** ((Penrose 1994), p. 82–83)

Penrose responds to this objection as follows.

> While this is presumably true, it ignores the essential issue, which is how we (or computers) know which mathematical statements are true and which are false. ... The way that the computer is being employed in **Q7** totally ignores the critical issue of *truth judgement*. ((Penrose 1994), p. 83)

As a rebuttal against a proponent of "Weak" AI, what Penrose says here is, alas, worthless. "Weak" AIniks, proponents of $\mathcal{B}$ (or, more precisely, $\mathcal{B}_2^M$), explicitly ignore such genuinely mental phenomena as truth judgment. They only care about *simulating* such phenomena. On the other hand, no proponent of $\mathcal{A}$ would articulate **Q7** in the first place.

The upshot of all this is that from this point on we view Penrose's Gödelian case as an argument exclusively against $\mathcal{A}$, for it's just a brute fact that he will do no damage to $\mathcal{B}_2^M$.

Ironically, Penrose's work is likely to catalyze "Weak" AI projects; specifically, formidable attempts to implement systems capable of establishing results currently within the reach of only human mathematicians. (Hubert Dreyfus' widely publicized claim that chess grandmasters would forever be restricted to *homo sapiens* seemed to motivate many to prove him wrong.) This is not to say that there is in *SOTM* material that can provide a partial blueprint for building an artificial mathematician; we see no such material. However, we do see specific examples of human mathematical achievement that are likely to be taken by some as targets for a robot mathematician. One such example involves the proof of an interesting theorem about *hexagonal numbers* and *cubes*; this theorem serves as a pivot around which the formal background for Penrose's new Gödelian case revolves. We turn now to this background, and the theorem in question.

4 Background for Penrose's New Gödelian Case

It's uncontroversial that there are certain Turing machines (TMs)[7] which provably never halt. Following one of Penrose's examples, begin by considering the hexagonal numbers,

$$1, 7, 19, 37, 61, 91, 127, \ldots$$

i.e., the numbers that can be arranged as ever-increasing hexagonal arrays (see Figure 1). Now consider the cubes:

$$1 = 1^3, 8 = 2^3, 27 = 3^3, 64 = 4^3, 125 = 5^3, \ldots$$

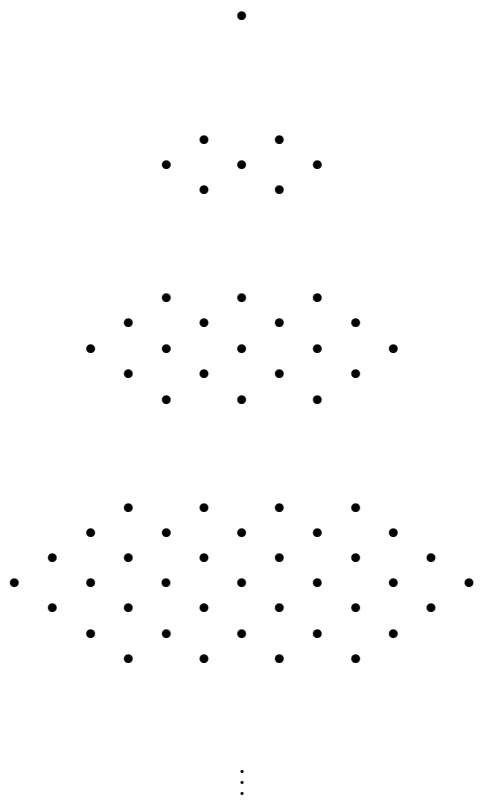
Let TM $M_{\bar{c}}$ be defined as follows. $M_{\bar{c}}$ adds together the hexagonal numbers successively, starting with 1, checking to see if each sum is a cube. If so, the machine keeps working away; if not, it halts. Does $M_{\bar{c}}$ halt? No, and mathematicians can prove it, for that the pattern

$$1 = 1, 1 + 7 = 8, 1 + 7 + 19 = 27, 1 + 7 + 19 + 37 = 64, 1 + 7 + 19 + 37 + 61 = 125, \ldots$$

continues forever is a theorem.

The basic background idea to be derived from the foregoing is that there is some procedure[8] (let's call it '$\Re$') by virtue of which mathematicians correctly classify some Turing machines (or their "user-friendly" equivalents, e.g., algorithms) as non-halters. Penrose's negative objective in *SOTM* is to establish that $\Re$ is uncomputable.[9] This objective is to be reached, according to his plan, via first demonstrating that

Figure 1: Hexagonal Numbers as Arrays



$\mathcal{G}$ For every "knowably sound" algorithm A for classifying Turing machines as non-halters, $\Re \neq A$.

After this attempted demonstration (Chapter 2), Penrose's plan calls for ruling out the other "computationalist" position (Chapter 3; this is the position others seem to regard as quite formidable, e.g. see (Chalmers 1995)), namely, that $\Re$ is indeed an algorithm, just not a "knowably sound" one. We first consider Chapter 2's argument against $\mathcal{G}$.

5 The Core Diagonal Argument

The heart of the Gödelian Case against $\mathcal{G}$ is a less-than-one-page presentation of a certain diagonal argument designed to show that assuming $\Re$ to be some sound set of computational rules (= some TM) A results in the revelation that $\Re$ cannot be such a set (= such a TM). Here — reproduced verbatim to preclude any inaccuracy — is the diagonal argument for $\mathcal{G}$. It begins with the assumption that

> A is just *any sound* set of computational rules for ascertaining that some computations $C_q(n)$ do not ever halt. Being dependent upon the two numbers q and n, the computation that A performs can be written $A(q, n)$, and we have
>
> **(H)** If $A(q, n)$ stops, then $C_q(n)$ does not stop.
>
> Now let us consider the particular statements **(H)** for which q is put equal to n... we now have:
>
> **(I)** If $A(n, n)$ stops, then $C_n(n)$ does not stop.
>
> We now notice that $A(n, n)$ depends upon just *one* number n, not two, so it must be one of the computations $C_0, C_1, C_2, C_3, \ldots$, since this was supposed to be a listing of *all* the computations that can be performed on a single natural number n. Let us suppose that this is in fact C_k, so we have:
>
> **(J)** $A(n, n) = C_k(n)$.
>
> Now examine the particular value $n = k$. We have, from **(J)**,
>
> **(K)** $A(k, k) = C_k(k)$.
>
> and, from **(I)**, with $n = k$:
>
> **(L)** If $A(k, k)$ stops, then $C_k(k)$ does not stop.
>
> Substituting **(K)** in **(L)** we find:
>
> (M) If $C_k(k)$ stops, then $C_k(k)$ does not stop.
>
> From this, we must deduce that the computation $C_k(k)$ does not stop. (For if it did then it does not, according to **(M)**!) But $A(k, k)$ cannot stop either, since by **(K)**, it is the *same* as $C_k(k)$. Thus, our procedure A is incapable of ascertaining that this particular computation $C_k(k)$ does not stop even though it does not. Moreover, if we *know* that A is sound, then we *know* that $C_k(k)$ does not stop. Thus, we know something that A is unable to ascertain. It follows that A *cannot* encapsulate our understanding. ((Penrose 1994), pp. 74–75)

Immediately after presenting this argument Penrose says

> At this point, the cautious reader might wish to read over the whole argument again ... just to make sure that we have not indulged in any 'sleight of hand'! Admittedly there is an air of the conjuring trick about the argument, but it is perfectly legitimate, and it only gains in strength the more minutely it is examined. ((Penrose 1994), pp. 75)

Unfortunately, after taking up Penrose's challenge, having examined the argument minutely, we find that he is stone cold wrong: the argument, in the end, *is* nothing more than prestidigitation. Parts of it are at best enthymematic, and when the whole thing is rendered precisely, a remarkable number of technical glitches come to light. But over and above these defects, there is a fatal dilemma afflicting the argument: at best, it is an instance of either the fallacy of denying the antecedent, the fallacy of *petitio principii*, or the fallacy of equivocation. In falling prey to these fallacies, Penrose's new Gödelian Case is unmasked as the same confused refrain J.R. Lucas (Lucas 1964) initiated 35 years ago.

6 Formal Machinery

In order to expose Penrose's legerdemain, we need only formalize the argument. Our formalization, and the evaluation thereof, will be charitably naive — because they will be carried out without exploiting the fact that Penrose has misdescribed the connection between his diagonal argument and the associated meta-theorems in mathematical logic (not the least of which are Gödel's incompleteness results themselves).[10] We will start by attempting to formalize Penrose's diagonal argument in $\mathcal{L}_I$ (full first-order logic). As will be seen, the formalization will eventually call for $\mathcal{L}_{II}$ (second order logic).[11]

For our formalization we follow the notation of (Ebbinghaus, Flum and Thomas 1984), and hence deploy atomic formulas

$$M_t : u \rightarrow v$$

to denote the fact that TM M_t, starting with u as input on its tape, halts and leaves v as output. Similarly,

$$M_t : u \rightarrow \text{halt}$$

and

$$M_t : u \rightarrow \infty$$

denote, respectively, that the TM in question halts and doesn't halt (on input u). Next, assume that the alphabet with which our TMs work is of the standard sort, specifically $\{|, \bullet\}$, where a natural number n is coded as a string of n |s, and $\bullet$ is used solely for punctuation. Finally, fix some enumeration of all Turing machines and a corresponding Gödel numbering scheme allowing us to reference these machines via their corresponding natural numbers.

With this machinery, humble as it is, it's easy to formalize certified diagonal arguments, like the classic one traditionally used to establish the halting problem, that is, that there is no TM $M_{h \vee \bar{h}}$ which can ascertain whether or not a given TM halts.[12] In the following Fitch-style formalization, with '1' used to signal 'Yes' and '0' 'No,' one starts by assuming that some $M_{h \vee \bar{h}}$ *does* exist, from which a contradiction for *reductio* is derived.[13]

1	$\exists p \forall r \forall s[(M_p : r \bullet s \rightarrow 1 \Leftrightarrow M_r : s \rightarrow \text{halt}) \wedge$ $(M_p : r \bullet s \rightarrow 0 \Leftrightarrow M_r : s \rightarrow \infty)] = \phi$	supp.
2	$\phi \Rightarrow \exists m \forall n[M_m : n \rightarrow \text{halt} \Leftrightarrow M_n : n \rightarrow \infty]$	Lemma 1
3	$\exists m \forall n[M_m : n \rightarrow \text{halt} \Leftrightarrow M_n : n \rightarrow \infty]$	1, 2 MP
4	$\quad \forall n[M_a : n \rightarrow \text{halt} \Leftrightarrow M_n : n \rightarrow \infty]$	supp.
5	$\quad M_a : a \rightarrow \text{halt} \Leftrightarrow M_a : a \rightarrow \infty$	$\forall$E
6	$Z \wedge \neg Z$	3, 4−5 $\exists$E & RAA
7	$\neg \phi$	1−6 RAA

One comment on this proof,[14] an important one given our coming formalization of Penrose's diagonal argument. Note that once an implicit contradiction $p \Leftrightarrow \neg p$ is obtained in line 5, a contradictory formula, devoid of the constant a, is obtained (via the fact that everything follows from a contradiction). That the instantiating constant a not occur in line 6 is required by one of the standard restrictions on the rule $\exists$E existential elimination. This requirement, as is well-known, ensures that the instantiating constant plays only an intermediary role. To violate it is to allow for absurdities such as that from the fact that there is a negative number it follows that two is a negative number.[15]

7 Formalizing Penrose's Diagonal Argument

Now, what does Penrose's diagonal argument look like once it's formalized using the machinery at our disposal? The initial part of the formalization is straightforward. Penrose begins by assuming that there is some set A of computational rules (we use 'M_a' to refer to A as TM; this is an identification Penrose himself, following standard mathematical practice, explicitly sanctions in Appendix A of *SOTM*) such that: if A yields a verdict that some TM M fails to halt on input n, then M *does* fail to halt on n. He then moves, via quantifier manipulation, through **(H)** to **(I)**. Here's how this initial reasoning runs:

$1'$	$\exists m \forall q \forall n[M_m : q \bullet n \rightarrow \text{halt} \Rightarrow M_q : n \rightarrow \infty]$	supp.
$2'$	$\quad \forall q \forall n[M_a : q \bullet n \rightarrow \text{halt} \Rightarrow M_q : n \rightarrow \infty] = \mathbf{(H)}$	supp.
$3'$	$\quad \forall n[M_a : b \bullet n \rightarrow \text{halt} \Rightarrow M_b : n \rightarrow \infty]$	$2'\,\forall$E
$4'$	$\quad M_a : b \bullet b \rightarrow \text{halt} \Rightarrow M_b : b \rightarrow \infty$	$3'\,\forall$E
$5'$	$\quad \forall n[M_a : n \bullet n \rightarrow \text{halt} \Rightarrow M_n : n \rightarrow \infty] = \mathbf{(I)}$	$4'\,\forall$I

At this point we reach the reasoning from **(I)** to **(J)**, and things begin to turn a bit murky. The reasoning, recall, is (from p. 75 of (Penrose 1994)):

> We now notice that $A(n, n)$ depends upon just *one* number n, not two, so it must be one of the computations $C_0, C_1, C_2, C_3, \ldots$, since this was supposed to be a listing of *all* the computations that can be performed on a single natural number n. Let us suppose that this is in fact C_k, so we have:
>
> **(J)** $A(n, n) = C_k(n)$.

What, formally speaking, sanctions this reasoning? What entitles Penrose to infer that a TM operating on input $n \bullet n$ is identical to one operating on just n? One possibility that comes to mind is a pair of elementary theorems like

($\mathbf{T}_1$) $\forall m \forall n[M_m : n \bullet n \rightarrow \text{halt} \Rightarrow \exists q(M_q : n \rightarrow \text{halt})]$

($\mathbf{T}_2$) $\forall m \forall n \forall o[M_m : n \bullet n \rightarrow o \Rightarrow \exists q(M_q : n \rightarrow o)]$

($\mathbf{T}_1$) is easily established by construction. Given a TM M_1 that operates on input $n \bullet n$ and eventually halts, it's easy to build a TM M_2 which starts with just n on its tape, calls a TM M_2 which copies n so that $n \bullet n$ is written on the tape, and then proceeds to simulate M_1 step for step. The same sort of simple trick verifies **($\mathbf{T}_2$)**.

Neither of these two theorems, however, can be what Penrose presupposes, for he needs the *identity* **(J)**. Moreover, that which might do the trick for him is false. Specifically, these two propositions aren't theorems (and are, in fact, easily counter-exampled):

($\mathbf{T}_3$) $\forall m \forall n[M_m : n \bullet n \rightarrow \text{halt} \Rightarrow M_m : n \rightarrow \text{halt}]$

($\mathbf{T}_4$) $\forall m \forall n[M_m : n \bullet n \rightarrow \text{halt} \Rightarrow \exists q(M_q = M_m \wedge M_q : n \rightarrow \text{halt})]$

Charity suggests devising something to rescue Penrose's reasoning. What did he have in mind? We think Penrose, during his moves from **(I)** to **(J)**, had in mind a rationale like that behind the likes of **($\mathbf{T}_1$)** and **($\mathbf{T}_2$)**. His use of identity may stem from an erroneous but tempting conflation of *Turing machine* with Turing machine *computation*, where the latter (following the textbook view) is a sequence of configurations of a TM. In order to grasp our exegesis, think back to the machines M_1, M_2, and M_3 in the justification given above for **($\mathbf{T}_1$)**. Here, it's no doubt safe to say that though M_1, strictly speaking, is diverse from the composite machine M_3 composed of M_2 and M_1, M_1 and M_3 are virtually identical — because the computations involved differ only in regard to the trivial duplication of n.[16] So let us say that in cases like this the machines in question are *approximately identical*, written (in this case) $M_1 \approx M_3$. And let us affirm, on Penrose's behalf, the appropriate sentence (see $6'$ in the next portion of the derivation), as well as inference rules for $\approx$ paralleling those for $=$. Now we can continue our careful rendition of Penrose's diagonal argument. Indeed, we can make it beyond **(M)**:

$6'$	$\forall n[(M_n : n \bullet n \rightarrow \text{halt} \Rightarrow M_n : n \rightarrow \infty) \Rightarrow$	
	$\exists q(M_q \approx M_n \wedge (M_q : q \rightarrow \text{halt} \Rightarrow M_q : q \rightarrow \infty))]$	Lemma
$7'$	$M_a : a \bullet a \rightarrow \text{halt} \Rightarrow M_a : a \rightarrow \infty$	$5'\,\forall\text{E}$
8	$(M_a : a \bullet a \rightarrow \text{halt} \Rightarrow M_a : a \rightarrow \infty) \Rightarrow$	
	$\exists q(M_q \approx M_a \wedge (M_q : q \rightarrow \text{halt} \Rightarrow M_q : q \rightarrow \infty))$	$6'\,\forall\text{E}$
9	$\exists q(M_q \approx M_a \wedge (M_q : q \rightarrow \text{halt} \Rightarrow M_q : q \rightarrow \infty))$	$7', 8\,\text{MP}$
10	$M_k \approx M_a \wedge (M_k : k \rightarrow \text{halt} \Rightarrow M_k : k \rightarrow \infty)$	supp.
11	$M_k : k \rightarrow \text{halt}$	supp.
12	$M_k : k \rightarrow \text{halt} \Rightarrow M_k : k \rightarrow \infty = (\mathbf{M})$	$10\,\wedge\text{E}$
13	$M_k : k \rightarrow \infty$	$11, 12\,\text{MP}$
14	$M_k : k \rightarrow \text{halt}$	$11\,\text{R}$
15	$M_k : k \rightarrow \infty$	$11-14\,\text{RAA}$

At this point it may be thought that things are beginning to look decidedly up for Penrose. For not only have we reached his **(M)**, but we have also achieved, in line 15, the formal equivalent to his assertion that "$C_k(k)$ does *not* in fact stop." Two problems, however, stare us in the face.

Problem 1 is that everything to this point is based on two *undischarged* suppositions, $2'$ and 10, in which existentially quantified variables are instantiated to what are supposed to be arbitrary and intermediary constants, a in line 2 and k in line 10. There is no indication whatsoever from the text in question that Penrose intends to discharge these assumptions. In fact, the text clearly indicates that Penrose intends to rest his diagonal argument with the constants a (for his A) and k undischarged. Though we don't understand how this could be (given the traditional mathematical use of, and restrictions on, instantiating arbitrary constants), we will assume for the sake of argument that this defect can be remedied.

Problem 2 is that it's impossible to derive that which would coincide in our formalization with "But $A(k, k)$ cannot stop either, since by **(K)**, it is the *same* as $C_k(k)$." What would suffice to validate this prose is a derivation of $M_a : k \bullet k \rightarrow \infty$; but this formula can't be derived. (We *can* derive $M_a : k \rightarrow \infty$, by first isolating the "identity" $M_k \approx M_a$ from line 10 and then using this "identity" with line 15 and the indiscernibility of identity.) We see no way to rectify this problem (after burning more than a few grey cells in the attempt to extend and/or modify the proof), but, once again, for the sake of argument we're prepared to assume that somehow Penrose can survive, that is, that he can continue the proof. So we then have:

$1'$		
$\vdots$		
15	$M_k : k \rightarrow \infty$	$11-14\,\mathrm{RAA}$
$\vdots$	$\vdots$	$\vdots$
n	$M_a : k \bullet k \rightarrow \infty$	
$n+1$	$(M_k : k \rightarrow \infty) \wedge (M_a : k \bullet k \rightarrow \infty)$	$15, n \wedge \mathrm{I}$

8 Penrose's Dilemma: Either Way a Fallacy

With the diagonal argument done, Penrose now purports to wrap things up:

> Thus, our procedure A is incapable of ascertaining that this particular computation $C_k(k)$ does not stop even though it does not. Moreover, if we *know* that A is sound, then we *know* that $C_k(k)$ does not stop. Thus, we know something that A is unable to ascertain. It follows that A *cannot* encapsulate our understanding. ((Penrose 1994), p. 75)

The first sentence in perhaps a bit misleading. What we *can* conclude from line $n+1$ is that M_a doesn't yield a verdict on whether M_k halts on input k, and M_k doesn't halt on k. This, combined with a relevant reinstantiation of $2'$, viz.,

n+2 $M_a : k \bullet k \rightarrow \text{halt} \Rightarrow M_k : k \rightarrow \infty$

gives nothing helpful. Though we admit it's a rather unflattering view, Penrose may fall prey here to the fallacy of denying the antecedent, for with it invoked on lines $n+2$ and n he would have the negation of the consequent in $n+2$, $M_k : k \rightarrow \text{halt}$, which would contradict line 15 — which would in turn give him, by *reductio ad absurdum*, the denial of line $1'$. He would then have proved that there is no Turing machine (algorithm, set of computational rules, etc.) that can do what $\Re$ does (viz., give correct verdicts on non-haltingness), which is certainly what he wants to ultimately establish.

The other possibility would seem to be that Penrose's imprecision has led him to confuse the conditional $2'$ with a conditional running in the opposite direction, i.e.,

$2''$ $\quad \forall q \forall n[M_a : q \bullet n \rightarrow \text{halt} \Leftarrow M_q : n \rightarrow \infty]$

For note that $2''$, once instantiated (with both q and n to k), yields, by *modus tollens* with line n, a contradiction with line 15; this in turn allows for the *reductio* that would, by the reasoning just explained, make Penrose's day.

Unfortunately, $2''$ taken as premise[17] begs the entire question. The reason is as follows. First, $2''$ and $2'$ combine to produce

$2'''$ $\quad \forall q \forall n[M_a : q \bullet n \rightarrow \text{halt} \Leftrightarrow M_q : n \rightarrow \infty]$,

Second, (as is well-known) there exists a TM M_h[18] such that

$$\forall r \forall s(M_h : r \bullet s \rightarrow 1 \Leftrightarrow M_r : s \rightarrow \text{halt}).$$

Third, any TM which halts on input m can be effortlessly adapted to print a string u on its tape and *then* halt. It follows that $2'''$ implies that there is some machine $M_{a'}$, adapted from M_a, that can solve the halting problem! Since no Turing machine can solve this problem, and since (under the setup Penrose has erected) what implies that A *can* solve it is that A can match $\Re$, it follows

that A cannot match $\Re$; and this, again, is precisely what Penrose wants to show. The problem, of course, is that this reasoning has as a premise that $\Re$ can solve the halting problem, and whether or not some human cognition is capable of this is precisely what's at issue in the debate of which *SOTM* and *ENM* are a part! So on this reading, Penrose begs the question. And that is the first part of his dilemma: either way, so far, the Gödelian Case is fallacious.

But what of the final moves in Penrose's argument? That is, what of

> Moreover, if we *know* that A is sound, then we *know* that $C_k(k)$ does not stop. Thus, we know something that A is unable to ascertain. It follows that A *cannot* encapsulate our understanding. ((Penrose 1994), p. 75)

It is easy to render this precise, given our analysis. And what this analysis reveals, alas, is that Penrose is once again perched atop the same dilemma, the horns of which are the fallacies cited above: That "if we *know* that A is sound, then we *know* that $C_k(k)$ does not stop" amounts (courtesy of the proof above from $1'$ to $n+1$, and concessions that both $6'$ is separately provable and failing to discharge 10 is somehow surmountable) to knowledge of

$$\{1'\} \vdash 15.$$

All that remains now is to cash out the final two sentences in the previous quote. These two sentences seem to suggest that "our understanding" is capable of ascertaining whether or not $\Gamma \vdash \psi$, where Γ is some set of first-order formulas, and ψ is one such formula. But of course ascertaining whether or not $\Gamma \vdash \psi$ is provably equivalent to ascertaining whether or not a given Turing machine halts (e.g., see the elegant proof in (Boolos and Jeffrey 1989)). Hence, A obviously cannot solve the problem of ascertaining whether or not such implications hold. Moreover, to assume that $\Re$, our understanding, can, is to beg the question in the manner discussed above (since $\Re$ would here be capable of solving the halting problem). The other horn, again, is that if Penrose retreats to the circumspect view he started with, that A simply cannot yield a verdict on whether or not $\{1'\} \vdash 15$ (because A is only a *sound* procedure, in the sense that if A says "Yes" then $\{1'\} \vdash 15$, and nothing more), he needs to invoke the fallacious rule $\{\psi \Rightarrow \gamma, \neg\psi\} \vdash \neg\gamma$ in order to get the contradiction $\{1'\} \vdash 15$ and $\{1'\} \nvdash 15$.

9 Possible Replies

What might Penrose say for himself? How might he try to dodge the dilemma? We suspect he might say that the diagonal argument is included in another, wider argument which we have mangled; he might claim, more specifically, that he never intended to generate a contradiction from assuming $1'$. In fact, as alert readers will doubtless have already noticed, this *must* be his claim, for $1'$ is in fact an easily proved theorem. For the (trivial) proof, note that all Turing machines can be recast as flow diagrams. Note, in particular, that any TM represented by a flow diagram having as part the fragment shown in Figure 2 would be a non-halting TM (because if started in state 1 with its read/write head scanning a block of |s it will loop forever in this fragment). So we obviously have near at hand a Turing machine M_4 which, upon being given as input a TM M and a string of |s as the input to M, does the following:

- converts M into a flow diagram;
- checks to see if the fragment of Figure 2 is in the diagram;
- if so, it outputs a '0' (to indicate that M is a non-halter);

- if not, it goes into an infinite loop of its own.

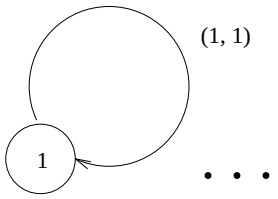


Figure 2: Flow-Diagram Fragment That Entails Non-Halting.

Proposition $1'$ follows from the existence of M_4 by existential introduction. Penrose might also complain that we went wrong in working from the previous quote to the general case of whether or not $\Gamma \vdash \psi$, for all first-order formulas ψ and sets Γ thereof.

What, now, is the wider argument that Penrose can be read as giving? In order to construct it on his behalf, we return to the observation that should Penrose establish that

$$\text{For every TM } M, \text{if } M \text{ is "knowably" sound in the sense of } 1', \text{ then } \Re \neq M$$

he will have succeeded. The next thing to note is that this proposition can of course be established by conditional proof and universal introduction. That is, one could start by assuming that some arbitrary TM M' is sound (in the sense of $1'$ or $2'$; abbreviate this property by 'SOUND'), derive from this somehow that $\Re \neq M'$, move to the conditional, and supplant the constant denoting M' with a variable within the scope of a universal quantifier. If we once again pretend that somehow Penrose can work a miracle on discharging assumptions, we can capitalize on our earlier analysis to quickly build a candidate version of his reasoning:

α_1	$2'$	
$\vdots$		
α_n	$(M_k : k \rightarrow \infty) \wedge (M_a : k \bullet k \rightarrow \infty)$	$15, n \wedge \text{I}$
α_{n+1}	$\Re$ yields a verdict of *doesn't halt* w.r.t. $k \bullet k$ by virtue of proving $\{\alpha_1\} \vdash \alpha_n$.	
α_{n+2}	$\forall x \forall y(x = y \Leftrightarrow \forall X(Xx \Leftrightarrow Xy))$	Leibniz's Law
α_{n+3}	$\Re \neq M_a$	$\alpha_n, \alpha_{n+1}, \alpha_{n+2}$
α_{n+4}	$2' \Rightarrow \Re \neq M_a$	$\alpha_{n+1} - \alpha_{n+3}$ Cond. Proof
α_{n+5}	$\forall q(\text{SOUND}(M_q) \Rightarrow \Re \neq M_q)$	α_{n+4} $\forall$I

Have we finally arrived, then, at a victorious version of Penrose's new Gödelian Case? Hardly. In fact, yet another fallacy rears up here — the fallacy of equivocation. An instance of this fallacy is found in the currently enthymematic presentation of

$$\{\alpha_n, \alpha_{n+1}, \alpha_{n+2}\} \vdash \alpha_{n+3}.$$

The concept of *yielding a verdict* is used equivocally in these inferences: in connection with the TM M_a this concept is used in the straightforward, well-understood sense of a TM doing some work and then printing out '0'; in connection with $\Re$, however, the concept means something quite different — it means carrying out a meta-proof. In order to verify our diagnosis, you have only

to look more carefully at how LL is used. In order to use this proposition, the variable X must obviously be instantiated. Let's say that it's instantiated to V. What are V's truth conditions? On the one hand we have the normal, well-understood sense applicable to M_a: V is true of some triple (m, n, o) iff M_m halts upon taking $n \bullet o$ as input. On the other hand, we have the sense according to which $\Re$ is supposed to have this property: V is true of some triple (m, n, o) iff m can deliver a meta-proof showing that a TM M_n goes to infinity on o.[19]

The problem can be made painfully clear if we spotlight the implicit inferences, which run as follows. First, let the meaning of V be the well-understood sense. Now instantiate LL to get

$$\Re = M_a \Leftrightarrow \forall X(X\Re \Leftrightarrow XM_a).$$

Next, suppose for *reductio* that $\Re = M_a$. Then by *modus ponens* we have

$$\forall X(X\Re \Leftrightarrow XM_a).$$

We now observe that $\neg VM_a$, from which it follows (by universal instantion and biconditional elimination) that $\neg V\Re$. But from α_{n+1} we have $V\Re$ — contradiction.

The problem, of course, is the invalid inference from α_{n+1} to $V\Re$.[20] And the problem arising from an equivocal use of V is unavoidable: If one starts the reasoning we've just gone through with the "meta-proof" sense of V, then we can no longer count on knowing $\neg VM_a$.

Penrose's last chance, at this point, is to somehow define V disjunctively, taking account of both definitions. For example, perhaps he could stipulate that $\Re$ be "operationalized," so that its verdict is delivered by way of something suitably mechanical, perhaps the checking of a box tagged with `DH` for *doesn't halt*. Unfortunately, this doesn't help solve the problem in the least. This is so because what we *know* of M_a is that it fails to halt; for all we know, over and above this, this TM is capable of checking boxes, writing stories, ... carrying out astounding meta-proofs. In searching for a way to straddle the equivocation, Penrose catapults himself back to square one, for as Bringsjord (Bringsjord 1992) — and others as well (Slezak 1982), (Webb 1980) — has pointed out elsewhere, there is no reason whatever to think that Turing machines (of which, of course, M_a is one) can't deliver the meta-proofs (concerning number theory, recall) with which Penrose framed the entire investigation.

10 Given $\mathcal{G}$, The Other Possibilities

Now let's assume for the sake of argument that Chapter 2's argument for $\mathcal{G}$, contrary to what we have seen in the foregoing, succeeds. This leaves the following four computationalist possibilities (which Penrose isolates on pages 130-131) with respect to $\Re$:

P1 $\Re$ is unknowable and sound

P2 $\Re$ is sound and knowable, but not *knowably* sound

P3 $\Re$ is unsound (i.e., mathematicians unwittingly use an unsound algorithm)

P4 there are different algorithms for different mathematicians (so we cannot speak univocally of $\Re$)

It seems to us that nearly all of the arguments Penrose gives against P1-P4 are rather sloppy. (This is not to say that these arguments fail to fascinate. His treatment of P1 — including as it does the attempt to show that "$\Re$ is unknowable" implies the mystical $\mathcal{A}/\mathcal{D}$ notion that $\Re$ is the result of divine intervention — is quite ingenious.) We don't have the space to treat each possibility and each argument; we focus on P3, and on Penrose's attempt (pp. 137–141) to rule this possibility out. Similar analysis, with similar results, could be given for the remaining trio.

As readers will recall, and as Penrose well knows, if $\Re$ is unsound, then any argument against computationalism from Gödel's first incompleteness theorem will necessarily fail — because the hypothesis of this theorem is that the axioms in question are consistent.[21] What some readers may not know, and what some others may have forgotten, is something Penrose (tendentiously?) fails to even mention, let alone discuss, in *SOTM*: viz., Gödel's *second* incompleteness theorem, and the connection between this result and P3. Here's the theorem in question:

Gödel II: Where Φ is a set of Turing-decidable first-order formulas built from the symbol set $\{+, \times, 0, 1\}$, where $+$ and $\times$ are binary function symbols (intepreted as addition and multiplication, resp.) and 0 and 1 are constants denoting the numbers zero and one, and where $\Phi \subset \Phi_{PA}$, i.e., Φ is a subset of the Peano axioms for arithmetic, then it's not the case that $\Phi \vdash \text{Consis}_{\Phi}$, where Consis_{Φ} abbreviates a formula expressing the proposition that from Φ one cannot derive '0 = 1,' i.e., that Φ is consistent.[22]

Since classical mathematics includes ZFC set theory or some formal equivalent, it follows immediately from Gödel II that Penrose's belief that $\Re$ is not unsound cannot be the product of the sort of (to use Penrose's respectful phrase) "unassailable reasoning" at the heart of classical mathematics. Given this, why does Penrose reject P3? At first, it appears that his rationale is extraordinarily weak, for he asks at the outset of his discussion of P3:

> But is it really plausible that our unassailable mathematical beliefs might rest on an unsound system — so unsound, indeed, that '1=2' is in principle part of those beliefs? ((Penrose 1994), p. 138)

This question is anemic. To say that a system is inconsistent (e.g., that the set Φ of Gödel I is inconsistent) is to say that from it a contradiction can be derived, from which it does indeed follow by propositional logic that *any* proposition, including, then, '1=2,' can be derived. *But the question is whether the contradiction can be found; only if it can can an absurdity be produced.* Finding the contradiction is the issue!

Penrose himself seems to realize that his question (reproduced in the block quote immediately above) is little more than a rhetorical trick: he explicitly considers the possibility that the contradiction could be a "hidden" one (p. 138); and he offers Russell's paradox as an example. As Penrose says: "Without the contradiction having been perceived, the methods of reasoning might well have been trusted and perhaps followed by mathematicians for a good long while" (p. 139). So the obvious question is: Why isn't it possible that P3 is true, and therefore that $\Re$ is unsound, because there is a contradiction hidden in classical mathematics that no one has yet found? For that matter, why isn't it possible that there is a contradiction that will *never* be found?

Penrose's response to these questions is first to say that Russell's paradox could not have gone undetected for any great length of time (p. 139). Because Russell's paradox is so simple we concede for the sake of argument that this response is cogent. This possibility still remains, however: there could be *extraordinarily complicated* contradictions buried within classical mathematics. Penrose himself seems to clearly recognize that this is at least a conceptual possibility, for he writes as follows.

> One might imagine some much more subtle paradox, even lying implicit in what we believe to be unassailable mathematical procedures that we allow ourselves today — a paradox which might not come to light for centuries to come ((Penrose 1994), p. 139).

But immediately after reading this we face an exceedingly peculiar part of *SOTM*: Penrose promptly proceeds to identify the objection based upon the possibility of a "more subtle paradox" with the objection that there is no fixed $\Re$ underlying present-day mathematical understanding, but rather a series of algorithms in constant flux (p. 139). But these are two different objections.

Figure 3: Yablo's Paradox

Imagine an infinite sequence of sentences $s_0, s_1, s_2, \ldots$ each to the effect that every subsequent sentence is untrue:

(s_0) for all $k > 0, s_k$ is untrue,

(s_1) for all $k > 1, s_k$ is untrue,

(s_2) for all $k > 2, s_k$ is untrue, ...

Formalizing the sentences with a truth predicate, T, we have that for all natural numbers, n, s_n is the sentence $\forall k > n, \neg T s_k$. Note that each sentence refers to (quantifies over) only sentences later in the sequence. No sentence, therefore, refers to itself, even in an indirect, loop-like, fashion. There seems to be no circularity.

Given this set-up, the argument to contradiction goes as follows. For any n:

$$Ts_n \Rightarrow \forall k > n, \neg T s_k \qquad (*)$$
$$\Rightarrow \neg T s_{n+1}$$

But:

$$Ts_n \Rightarrow \forall k > n, \neg T s_k \qquad (*)$$
$$\Rightarrow \forall k > n+1, \neg T s_k$$
$$\Rightarrow T s_{n+1}$$

Hence, Ts_n entails a contradiction, so $\neg Ts_n$. But n was arbitrary. Hence $\forall k \neg Ts_k$, by Universal [Introduction]. In particular, then, $\forall k > 0, \neg Ts_k$, i.e., s_0, and so Ts_0. Contradiction (since we have already established $\neg Ts_0$).

The first objection is that $\Re$ and classical mathematics may well be unsound (in which case, as Penrose is forced to admit, his Gödelian case is derailed). The second objection is that talk of the singular and determinate $\Re$ is unjustified. We are not concerned with the second; the first, however, is ours. And Penrose does nothing to disarm it.

Penrose and his supporters might at this point ask: "Well, what paradox do you have in mind? If Russell's paradox doesn't do the trick for you, what does?" It's hard to see how this question can help Penrose. If $\Re$ is unsound because of a hidden contradiction, then the contradiction is just that: hidden. So we can hardly be obligated to display it. At most, the challenge to us is to say what *sort* of paradox might point the way toward the hidden contradiction. This challenge is easy to meet. First, the kind of paradoxes we have in mind are purely logico-mathematical; "physics-related" paradoxes, such as those attributed to Zeno (see e.g., (Salmon 1975)), are irrelevant. Second, the paradoxes we see as supporting the notion that there may for all we know be hidden contradictions in $\Re$ are ones that aren't solved. Our current favorite is the Yablo Paradox (Yablo 1993), presented in Figure 3.

Our point is not that Yablo's Paradox in insoluble. The point is that we seem to have met the challenge to characterize the sort of paradox that should give Penrose pause. $\Re$, for all we know, involves some hidden contradiction of the sort exposed by Yablo's Paradox, a contradiction that is much more subtle than that produced by Russell's simple paradox.[23]

As a matter of fact, one of us (Selmer) wrote this paradox out in person for Penrose to subsequently attempt to solve. No solution has arrived. While Graham Priest (Priest 1984) has

diagnosed the version of the paradox given above as self-referential, this is not to solve the paradox, for self-referentiality, in and of itself, is quite innocent (as Gödel I itself shows). And as Priest concedes, *fully* infinitary versions of the paradox aren't self-referential. It is infinitary versions of Yablo's Paradox that should prove particularly disquieting for Penrose, since he is avowedly relaxed about the meaningfulness of infinitary mathematical reasoning.[24] To adapt Yablo's paradox so as to produce a fully infinitary version, the first trick is to replace the dangerous[25] imperative "Imagine an infinite ..." with the benign "Recall the natural numbers $\mathbf{N} = \{0, 1, 2, \ldots\}$. For each $n \in \mathbf{N}$ there exists a corresponding sentence

$$\forall k > n, s_k \text{ is untrue } \ldots"$$

At this point we construct the list $(s_0), (s_1), (s_2), \ldots$ as before — so that, e.g.,

$$(\star) \qquad Ts_0 \text{ iff } \forall k(k > o \Rightarrow \neg Ts_k).$$

The second trick is to turn the n in the finitary *reductio* from a free variable to a particular natural number. Notice that for each particular natural number n, a proof by contradiction exists. By the ω-rule

$$\frac{\alpha(1), \alpha(2), \ldots}{\alpha(n)}$$

we can infer

$$\forall n \neg Ts_n$$

and we reach a contradiction again by instantiating the biconditional $(\star)$. (Put in these terms, the paradox is couched in the logical system $\mathcal{L}_{\omega_1\omega}$.) We challenge Penrose to solve this paradox while at the same time clinging to his views on the meaningfulness of infinitary mathematical reasoning. More importantly, we simply observe that Penrose's view that no such contradiction is hidden in the foundations of mathematics is nowhere substantiated. It follows, then, that even if $\mathcal{G}$ is somehow true, Penrose's attack on $\mathcal{A}$ is at best inconclusive.

11 Penrose's Last Chance

As was mentioned in passing earlier, *SOTM* was evaluated by a number of thinkers who then published their critiques in the electronic journal *Psyche*. Penrose then wrote a sustained response to these critiques.[26] In this response Penrose gives what he takes to be the perfected version of the core Gödelian case given in *SOTM*. Here is this version, verbatim:

> We try to suppose that the totality of methods of (unassailable) mathematical reasoning that are in principle humanly accessible can be encapsulated in some (not necessarily computational) sound formal system F. A human mathematician, if presented with F, could argue as follows (bearing in mind that the phrase "I am F" is merely a shorthand for "F encapsulates all the humanly accessible methods of mathematical proof"):
>
> > (A) "Though I don't know that I necessarily am F, I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F' is F supplemented by the further assertion "I am F." I perceive that it follows from the assumption that I am F that the Gödel statement $G(F')$ would have to be true and, furthermore, that it would not be a consequence of F'. But I have just perceived that "If I happened to be F, then $G(F')$ would have to be true," and perceptions of this nature would be precisely what F' is supposed to

> achieve. Since I am therefore capable of perceiving something beyond the powers of F', I deduce that I cannot be F after all. Moreover, this applies to any other (Gödelizable) system, in place of F." ((Penrose 1996), ¶ 3.2)

Unfortunately, (A) is a bad argument, as is easily seen. In order to see this, let's follow Penrose directly and set

$$\psi = \text{ "}F \text{ encapsulates all the humanly accessible methods of mathematical proof"}$$

and

$$F' = F \cup \psi$$

What the hypothetical human mathematician can now conclude, as Penrose tells us, is that on the assumption that ψ,

(16) $G(F')$ is true.

(17) $F' \nvdash G(F')$ and $F' \nvdash \neg G(F')$

The idea is really quite simple. It is that there is a contradiction arising from the fact that the hypothetical mathematician, i.e. F, can conclude that (16) $G(F')$ is true on the one hand, and yet (17), which "says" that F cannot conclude $G(F')$, is true on the other. But wait a minute; look closer here. Where is the contradiction, exactly? There is no contradiction. The reason is that (16) is a *meta*-mathematical assertion; it's a claim about *satisfaction*. More precisely, where $\mathcal{I}$ is an interpretation, (16) is just

(16′) $\mathcal{I} \models G(F')$ is true.

And for all we know, F can prove (16′) while being bound by (17)! So we see here again what we saw above in section 9: Penrose conflates proofs within a fixed system with meta-proofs.

12 Conclusion; The Future

So, Penrose has tried three times to refute "Strong" AI by central appeal to Gödelian theorems, and each time he has flatly failed. Are Penrose's arguments in the end any improvement over Lucas' (Lucas 1964) tiny-by-comparison primogenitor? By our lights, the answer is both "Yes" and "No." The answer is "Yes" because certainly Penrose has fleshed out the line of thought only adumbrated by Lucas. After all, *SOTM* is a big, beautifully written tour, the guide for which is an engaging polymath. Those who read it ought not to be thereby convinced that minds aren't machines (they should be convinced by the likes of the arguments listed in Table 1), but they will learn lots of things about computability theory, mathematics, and physics. On the other hand, the answer is "No" because, alas, Penrose's core Gödelian arguments are no better than those of Lucas. This we have painstakingly shown. Now, this raises the obvious question: Should we conclude that a denial of the computational conception of mind simply cannot be deduced from Gödelian results? To infer an affirmative answer on the basis of what we have shown would be to commit a non sequitur.

On the other hand, LaForte, Hayes, and Ford have recently argued that Gödel's theorem (and related results) cannot refute computationalism, period (LaForte et al. 1998). Here is what they say:

> *Any* attempt to utilize the undecidability and non-termination results to attack the computationalist thesis is bound to be illegitimate . . ., 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 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. ((LaForte et al. 1998), p. 285)

This is marvelously ironic. The L-H-F trio reason here exactly as Penrose does: informally. Identify the computationalist hypothesis L-H-F have in mind with our $\mathcal{A}$. Now, where is the *proof* that Gödel I and $\mathcal{A}$ and propositions expressing the powers of the human mind are consistent? L-H-F don't provide this proof; they don't even sketch it; they just baldly assert this consistency. At least Penrose has *tried* to offer arguments. What would happen if L-H-F did try to prove what they merely assert? They would quickly learn that the proof is rather hard to come by. To see this, let the set $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ enumerate the familiar properties often offered as candidates for capturing, together, the essence of personhood. (These properties are listed and discussed in (Bringsjord 1997*a*).) If $P_i \in \mathcal{P}$, then $P_i^\star$ denotes a technical, precise correlate for P_i suitable for use in careful argumentation. What L-H-F need to provide is a proof that for *all* $P_i^\star$, these propositions are consistent:

1. Persons have $P_i^\star$
2. Gödel I (or non-termination theorem)
3. $\mathcal{A}$

This proof would consist in specifying a model on which these propositions are at once demonstrably true, and would probably be in the running for the most complicated proof ever conceived!

The upshot is that the future, with respect to Penrose's goal, is still open. One of us (Bringsjord) is in complete and utter agreement with Penrose that it *is* possible to derive the denial of "Strong" AI from Gödelian facts, and is working on producing the demonstration.

Notes

[1]For an explanation of why standard Gödelian attacks fail, see "Chapter VII: Gödel" in (Bringsjord 1992).

[2]The straight Turing Test, as is well known, tests only for linguistic performance. The TTT, devised by Stevan Harnad (Harnad 1991), requires that the human and robot (or android) players compete across the full range of behavior. For example, the judge in TTT can ask questions designed to provoke an emotional response, and can then observe the facial expressions of the two players.

[3]We refer to *philosophers'* zombies (creatures who are behaviorally indistinguishable from from us, but who have the inner life of a rock), not those creature who shuffle about half-dead in the movies. Actually, the zombies of cinematic fame apparently have real-life correlates created with a mixture of drugs and pre-death burial: see (Davis 1985), (Davis 1988). One of us (Bringsjord) has recently discussed zombies in connection with the view that thinking is computing: (Bringsjord 1999).

[4]OTTER can be obtained at

`http://www-unix.mcs.anl.gov/AR/otter/`

[5]Here and hereafter we leave aside categorization of such propositions as Π_1^0, etc.

[6]Newell expressed his dream for a unified (production system-based) theory for all of human cognition in (Newell 1973).

[7]We assume readers to be familiar with Turing machines and related aspects of computability theory and logic.

[8]Not necessarily an *effective* procedure: to assume that the procedure is effective is of course to beg the question against Penrose, since he purports to show that the procedure in question *isn't* effective.

[9]His *positive* objective is to lay the foundation for a science of the mind which accommodates his negative results.

[10]Feferman (Feferman 1995) and Davis (Davis 1980) catalogue the inadequacy of Penrose's scholarship when it comes to mathematical logic. However, we don't think the quality of this scholarship, however questionable it may be, creates any fundamental problems for Penrose's core Gödelian arguments against "Strong" AI.

[11]It's important at this point to note that while many of the mathematical issues relating to Penrose's Gödelian arguments are not expressible in $\mathcal{L}_I$, the core arguments themselves must conform to the basic, inviolable principles of deductive reasoning that form the foundation of technical philosophy. *SOTM* is an essay in technical philosophy; it's not a mathematical proof. This paper is itself technical philosophy. We make reference to logical systems beyond $\mathcal{L}_I$, but our core reasoning is intended to meet standards for deductive reasoning circumscribed in $\mathcal{L}_I$ and $\mathcal{L}_{II}$. See (Ebbinghaus et al. 1984) for a nice introduction to these logical systems, as well as more advanced ones, such as the infinitary systems related to Yablo's Paradox, which we discuss later.

[12]The form of the halting problem we use is specifically that given a TM M and input to this machine, no TM can ascertain whether or not M halts on this input.

[13]Note that we use '∀E' to denote the inference rule of universal quantifier elimination, etc.

[14]We make no comments about what we've called 'Lemma 1.' For an elegant proof of this lemma (which, if formalized, would require more lines centered around the rules ∃E, ∀E, ∃I, ∀I), see (Boolos and Jeffrey 1989).

[15]The proof would be: $\exists x Nx$ as supposition, Nt as supposition, Nt by reiteration, and Nt with second supposition discharged by application of the (here erroneous) rule of ∃E.

[16]This is as good a place as any to point out that though Penrose's use of the term 'computation' is, to say the least, relaxed, a little analysis reveals that by it he can invariably be read as ultimately making reference to a Turing machine. For example, Penrose sometimes refers to "computations" by way of imperatives, as when he writes:

> Suppose we had tried, instead, the computation
>
> **(B)** Find a number that is not the sum of four square numbers.
>
> Now when we reach 7 we find that it *is* the sum of *four* squares: $7 = 1^2 + 1^2 + 1^2 + 2^2$,
> ...

It's clear from the commentary here after the presentation of **(B)** that Penrose's use of imperatives is elliptical for a description of an algorithm — the algorithm to be used in the attempt to meet the imperative.

[17]To put it more precisely, $2''$ would follow from taking as premise a sentence like $1''$, which would be $1'$ with $\Rightarrow$ changed to $\Leftrightarrow$.

[18]Which simply simulates, step for step, the TM M_r.

[19]Notice that there are other technical issues that arise here — issues a good deal more subtle than the sort Penrose customarily deals with. For example, the domain behind deployment of LL can't be the set of all TMs, as it has been to this point — because $\Re$, for all we know at this stage in the proof, isn't a TM. (The solution here would perhaps be to adopt as domain not only those machines in the computable part of the Arithmetic Hierarchy, but this set union those "machines" in the initial fragment of the uncomputable part of AH.)

[20]We leave aside issues arising from the self-referentiality of the situation: $\Re$ apparently carries out a proof in which it itself figures.

[21]Here, for reference, is a fairly careful statement of **Gödel I**:

> Let Φ be a consistent, Turing-decidable set of first-order formulas built from the symbol set $\{+, \times, 0, 1\}$, where $+$ and $\times$ are binary function symbols (intepreted as addition and multiplication, resp.) and 0 and 1 are constants denoting the numbers zero and one, and where Φ is *representable*. (For details on representability, see (Ebbinghaus et al. 1984).) Then there is a sentence ϕ built from the same symbol set such that: $\Phi \nvdash \phi$ and $\Phi \nvdash \neg\phi$.

Cognoscenti will note that here and herein we drop Gödel's original concept of ω-consistency in favor of "modern" versions. Penrose has conceded (e.g., see section 2, "Some Technical Slips in *Shadows*," in (Penrose 1996)) that in *SOTM* he erred in his use ω-consistency. (The errors in question were pointed out by Feferman (Feferman 1995) and others.) We believe these errors are complete red herrings.

[22]For details on how to construct Consis_{Φ}, see Chapter X in (Ebbinghaus et al. 1984).

[23]In personal conversation Penrose seemed to be inclined to complain that Yablo's paradox appears to go beyond what can be derived from axiomatic set theory (e.g., ZFC). This reply is odd, for *SOTM* is quite literally filled with reasoning that appears to go beyond first-order logic. (E.g., consider the diagrammatic "proof" concerning hexagonal numbers we visited above. Such diagrams seem to move beyond first-order logic (Bringsjord and Bringsjord 1996).)

[24]This is why one of us (Selmer) has long said that the Bringsjordian "Argument from Infinitary Reasoning" against "Strong" AI (Bringsjord 1997*b*), listed in Table 1, may well capture Penrose's core intuitions better than any argument appealing to Gödel I.

[25]Dangerous because we may well be attempting to imagine something that is incoherent.

[26]The dialectic appeared in 1996 in volume **2.23** of *Psyche*, which can be accessed via

- `http://psyche.cs.monash.edu`

And of course *Psyche* can be located using any standard search engine.

References

Anderson, J. R. (1998), *The Atomic Components of Thought*, Lawrence Erlbaum, Mahwah, NJ.

Boolos, G. S. and Jeffrey, R. C. (1989), *Computability and Logic*, Cambridge University Press, Cambridge, UK.

Bringsjord, S. (1992), *What Robots Can and Can't Be*, Kluwer, Dordrecht, The Netherlands.

Bringsjord, S. (1997*a*), *Abortion: A Dialogue*, Hackett, Indianapolis, IN.

Bringsjord, S. (1997*b*), An argument for the uncomputability of infinitary mathematical expertise, *in* P. Feltovich, K. Ford and P. Hayes, eds, 'Expertise in Context', AAAI Press, Menlo Park, CA, pp. 475–497.

Bringsjord, S. (1998), Philosophy and 'super' computation, *in* J. Moor and T. Bynam, eds, 'The Digital Phoenix: How Computers are Changing Philosophy', Blackwell, Oxford, UK, pp. 231–252.

Bringsjord, S. (1999), 'The zombie attack on the computational conception of mind', *Philosophy and Phenomenological Research* **59.1**, 41–69.

Bringsjord, S. (2001), 'Is Gödelian model-based reasoning computational?', *Philosophica* .

Bringsjord, S. and Bringsjord, E. (1996), 'The case against ai from imagistic expertise', *Journal of Experimental and Theoretical Artificial Intelligence* **8**, 383–397.

Bringsjord, S. and Ferrucci, D. (2000), *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Lawrence Erlbaum, Mahwah, NJ.

Bringsjord, S. and Zenzen, M. (1997), 'Cognition is not computation: The argument from irreversibility?', *Synthese* **113**, 285–320.

Chalmers, D. (1995), 'Minds, machines, and mathematics', *Psyche* **2.1**. This is an electronic publication. It is available at http://psyche.cs.monash.edu.au/psyche/volume2-1/psyche-95-2-7-shadows-7-chalmers.html.

Davis, M. (1980), 'How subtle is godel's theorem?', *Behavioral and Brain Sciences* **16**, 611–612.

Davis, W. (1985), *The Serpent and the Rainbow*, Simon & Shuster, New York, NY.

Davis, W. (1988), *Passage of Darkness: The Ethnobiology of the Haitian Zombie*, University of North Caronlina Press, Chapel Hill, NC.

Ebbinghaus, H. D., Flum, J. and Thomas, W. (1984), *Mathematical Logic*, Springer-Verlag, New York, NY.

Feferman, S. (1995), 'Penrose's godelian argument', *Psyche* **2.1**. This is an electronic publication. It is available at http://psyche.cs.monash.edu.au/psyche/volume2-1/psyche-95-2-7-shadows-5-feferman.html.

Harnad, S. (1991), 'Other bodies, other minds: A machine incarnation of an old philosophical problem', *Minds and Machines* **1.1**, 43–54. This paper is available online at ftp://cogsci.ecs.soton.ac.uk/pub/harnad/Harnad/harnad91.otherminds.

Kugel, P. (1986), 'Thinking may be more than computing', *Cognition* **18**, 128–149.

LaForte, G., Hayes, P. and Ford, K. (1998), 'Why Gödel's theorem cannot refute computationslism', *Artificial Intelligence* **104**, 265–286.

Lucas, J. R. (1964), Minds, machines, and Gödel, *in* A. R. Anderson, ed., 'Minds and Machines', Prentice-Hall, Englewood Cliffs, NJ, pp. 43–59. Lucas' paper is available online at http://users.ox.ac.uk/∼jrlucas/mmg.html.

Newell, A. (1973), Production systems: models of control structures, *in* W. Chase, ed., 'Visual Information Processing', Academic Press, New York, NY, pp. 463–526.

Penrose, R. (1989), *The Emperor's New Mind*, Oxford, Oxford, UK.

Penrose, R. (1994), *Shadows of the Mind*, Oxford, Oxford, UK.

Penrose, R. (1996), 'Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind*', *Psyche* **2.3**. This is an electronic publication. It is available at http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html.

Priest, G. (1984), 'Yablo's paradox', *Analysis* **57**(4), 236–242.

Russell, S. and Norvig, P. (1994), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Saddle River, NJ.

Salmon, W. C. (1975), *Space, Time and Motion: A Philosophical Introduction*, Dickenson, Encino, CA.

Searle, J. (1980), 'Minds, brains and programs', *Behavioral and Brain Sciences* **3**, 417–424. This paper is available online at http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html.

Searle, J. (1997), Roger penrose, kurt Gödel, and the cytoskeletons, *in* J. Searle, ed., 'The Mystery of Consciousness', New York Review of Books, New York, NY, pp. 53–93.

Siegelmann, H. (1995), 'Computation beyond the turing limit', *Science* **268**, 545–548.

Siegelmann, H. and Sontag, E. (1994), 'Analog computation via neural nets', *Theoretical Computer Science* **131**, 331–360.

Slezak, P. (1982), 'Gödel's theorem and the mind', *British Journal for the Philosophy of Science* **33**, 41–52.

Webb, J. (1980), *Mechanism, Mentalism and Metamathematics*, D. Reidel, Dordrecht, The Netherlands.

Yablo, S. (1993), 'Paradox without self-reference', *Analysis* **53**, 251–252.