THE BULLETIN OF SYMBOLIC LOGIC
Volume 4, Number 3, Sept. 1998

INCOMPLETENESS, MECHANISM, AND OPTIMISM

STEWART SHAPIRO

§1. **Overview.** Philosophers and mathematicians have drawn lots of conclusions from Gödel's incompleteness theorems, and related results from mathematical logic. Languages, minds, and machines figure prominently in the discussion. Gödel's theorems surely tell us something about these important matters. But what?

A descriptive title for this paper would be "Gödel, Lucas, Penrose, Turing, Feferman, Dummett, mechanism, optimism, reflection, and indefinite extensibility". Adding "God and the Devil" would probably be redundant. Despite the breath-taking, whirlwind tour, I have the modest aim of forging connections between different parts of this literature and clearing up some confusions, together with the less modest aim of not introducing any more confusions.

I propose to focus on three spheres within the literature on incompleteness. The first, and primary, one concerns arguments that Gödel's theorem refutes the mechanistic thesis that the human mind is, or can be accurately modeled as, a digital computer or a Turing machine. The most famous instance is the much reprinted J. R. Lucas [18]. To summarize, suppose that a mechanist provides plans for a machine, M, and claims that the output of M consists of all and only the arithmetic truths that a human (like Lucas), or the totality of human mathematicians, will ever or can ever know. We assume that the output of M is consistent. Now, since Lucas understands the proof of the incompleteness theorem, he can study M and construct its Gödel sentence G. Lucas knows that G will never be produced or "asserted" by M. He also knows that G "says" that G will never be produced by M. Thus, Lucas

Received June 2, 1998; revised July 7, 1998.

Part of this article is an extension of a paper delivered at a conference on language, minds, and machines, held in Palermo, Sicily, during August, 1977, sponsored by the Archimede Society and organized by Gianluigi Olivieri. I also gave an early version of the section on ordinals at the May, 1998 meeting of the Society for Exact Philosophy at the University of Georgia. Many thanks to both audiences and to Andreas Blass, Michael Detlefsen, Harry Deutsch, Diana Raffman, George Schumm, Neil Tennant, and an anonymous referee for helpful conversations and criticisms of an earlier draft.

1079-8986/98/0403-0002/$4.00

knows that G is true. So the mechanist was mistaken in the claim that the output of M contains all the truths that (any group containing) Lucas can know.

The eminent physicist/mathematician Roger Penrose [21, especially Chapters 4, 10] weighed in on the side of the anti-mechanist, and Gödel's unpublished writings show him to be a cautious occupant of this side of the battle line. Thus, when measured in terms of brain-power, the anti-mechanists are a formidable group. However, George Boolos [2, p. 295] correctly notes that "the arguments of these writers have as yet obtained little credence", and there is an extensive literature attacking the Lucas-Penrose position. In a recent article, Lucas [19] holds his ground, or tries to. About 200 pages of Penrose [22] are devoted to responses to various criticisms of the argument and an intriguing new version of it (see §3 below and Penrose [23]).

The second sphere considered here takes off from the observation that Gödel's incompleteness theorem is completely constructive. Given any ω-consistent formal deductive system S that contains a small amount of arithmetic, one *can effectively find* an arithmetic (Π_1) sentence G_S such that neither G_S nor its negation is a theorem of S. Moreover, if every arithmetic theorem of S is true, then G_S is true. This suggests that we just add G_S as a new "axiom" to S, producing a system $S1$. Then we can effectively find a sentence G_{S1} which is true but not provable in $S1$. Not wavering, we add G_{S1} as a new axiom, producing $S2$. And on it goes. The point is that the process is effective. A machine could carry it out as well as Lucas or Penrose, probably better. Thus, any attempt to effectively delimit the extension of arithmetic truth *effectively* leads to an arithmetic truth not so delimited. In his Gibbs lecture [11], Gödel calls this the "incompletability or inexhaustability of mathematics". Dummett [6] (see also [7]) argues that this consideration makes arithmetic truth "indefinitely extensible". This is one of his arguments against bivalence and classical logic. The issue here is what indefinite extensibility tells us about arithmetic understanding—language, minds and machines in particular. Judson Webb [33, p. vii] invokes the effectiveness of the incompleteness result to conclude that Gödel "established for the first time ... that, from the proposition 'I can find a limitation in any given machine', it by no means follows that I am not a machine". Clearly, we need to sort things out.

The third sphere comes from mathematical logic and not directly from philosophy. Consider the process of moving from a system S to its Gödel sentence G_S to the new system $S1 = S \cup \{G_S\}$. The relevant idea, traced to Turing [30] (and pursued in Feferman [9], [10]), is to extend the process into the transfinite. We collect together $S1$, $S2$, ... into a single system, which we can call $S\omega$. Then we get a Gödel sentence $G_{S\omega}$ for $S\omega$, and produce the system $S(\omega + 1) = S\omega \cup \{G_{S\omega}\}$. And onward, through the recursive ordinals. The results are some completeness theorems, of sorts.

Now, to business.

§2. **Idealization.** One problem is that the exact content of the mechanistic thesis is usually left unspecified. To belabor the obvious, the relevance of the incompleteness theorems to mechanism depends on what the mechanist claims. The raw thesis that the human mind is, or can be modeled as, a digital computer or Turing machine, is too vague to apply anything as sharp and delicate as the Gödel theorem and the Turing-Feferman extensions. My conclusion (perhaps slightly exaggerated) is that *there is no plausible mechanist thesis on offer that is sufficiently precise to be undermined by the incompleteness theorems*.

The mechanist claims that there can be a machine whose outputs are the same as those of a human or a group of humans. What sort of machine? What outputs? What aspect of what human? As for "output" , let us stick to propositions that can be rendered in the language of first-order Peano arithmetic. Penrose [23] goes so far as to restrict the output to Π_1-sentences. The totality of arithmetic sentences that a given person asserts in his lifetime is finite. The same goes for the totality of sentences asserted by any finite collection of humans, such as the professional mathematicians who lived or will live before the sun goes cold. Moreover, the totalities in question are certainly inconsistent. It only takes one mistaken calculation, later corrected. The mechanist might claim that there could be a machine whose output is one of these finite sets, or the truths among one of these sets, or the logical consequences thereof. If so, the incompleteness theorems are irrelevant.

Things get interesting only when we idealize, but things also get murky. Presumably, the mechanist and anti-mechanist are both talking about what an ideal human, or the community of ideal human mathematicians, *can prove* or know for certain. Lucas and Penrose both refer to human abilities "in principle". Of course, we must idealize on the "machines" as well. Like humans, actual digital computers have fixed limits on memory, and they are subject to hardware malfunctions and software bugs. I do not know if there could be a physical computer that matches a human being, reproducing both veridical output and error. I also do not know if there could be a physical computer whose output matches one of the finite sets in the previous paragraph. For all I know, it might not be physically possible to build a computer that big. Moreover, no actual computer can print all and only the logical consequences of one of those sets, since there are infinitely many such consequences, and we have good empirical confirmation that any machine will crash eventually. But all of this is off the point of any mechanistic claim that is supposed to be settled by Gödel's theorems.

The idealizations on the machine side are familiar, similar to idealizations made throughout mathematics. We ignore finite limits and assume that our machines never run out of memory, space, time, and attention span. We also assume that they run indefinitely without crashing. Part of the idea is to enforce the familiar distinction between hardware and software, and then completely ignore the hardware. Another part of the idea is to ignore practical or theoretical problems with limited memory and storage. In short, we deal with *Turing machines*, with their fixed programs and unlimited tapes. Some Wittgenstein-type worries about rule-following might come into play at this point, but I assume that things are pretty clear so far. There is no question of what set a given Turing machine enumerates, is there? If there is a question, set it aside.

Now, what about the human side of things? For any finite set S, there is a Turing machine that prints out the members of S and nothing else, and there is a Turing machine that prints out the logical consequences of S and nothing else (see Boyer [3]). Big deal. The principals to the present debate (try to) make idealizing assumptions about humans analogous to those of Turing machines. They do not speak of the theorems a subject does produce, but the theorems that she *can* produce. The mechanist should accommodate theorems whose shortest proofs are so long that no human can establish them without falling asleep or otherwise losing concentration, or without using up every particle in the universe. In short, the envisioned creatures have unlimited lifetimes, unlimited attention spans and energy, and unlimited materials at their disposal. Yet they are like humans in every other respect—whatever that means. Here is where rule-following considerations might become more serious. Is it clear which Turing machine Lucas (for example) would become if he were to undergo the envisioned modifications and idealizations—even if we restrict attention to Lucas's abilities concerning arithmetic sentences? Kripke [15] raises doubts about these matters, on behalf of the later Wittgenstein. So does Kreisel [14, pp. 317–318], on his own behalf. The mechanist and the Lucas-Penrose anti-mechanist must agree on a way to resolve this matter, and come up with a clear and unambiguous conception of idealized human mathematical ability. Otherwise, there is no meaningful debate. We need the idealizations *before* we can assess the relevance of the various theorems.

The principals to the debate also assume that our ideal subjects do not make mistakes—a *normative* idealization. To get around the human propensity to make mistakes, we consider the *correct* theorems that our ideal subject can produce. Implicitly, the standard move is to postulate something like an arithmetic competence/performance distinction in actual humans and then ignore problems with performance. The presupposition is that human arithmetic activity consists of following certain "routines" and "procedures".[1] The right arithmetic "software" is implemented in humans, or would be

[1]To avoid begging the question in favor of mechanism, we allow inherently informal "routines" and "procedures" (if such there be). As we shall see, Lucas and Penrose sometimes put the issue in these terms.

implemented in ideal humans free from memory and other relevant limitations. Errors that actual humans make are attributed to lack of attention or memory failure. I do not know if this presupposition is tenable. When dealing with natural organisms, can we sharply distinguish a breakdown or limitation in memory recall, for example, from an error in the "software"? Is there a clear distinction between human "hardware" and human arithmetic "software"—or what would be human software if memory limitations (etc.) were waived? The principals to the present debates presuppose that there is, and we will go along for a while in order to evaluate the debate. The normativity idealization is that human arithmetic "software" is free of arithmetic bugs, and so our idealized humans do not assert arithmetic falsehoods. At this point, the assumption is that the idealized humans produce the analogues of *theorems*, sentences in the language of arithmetic that are proved and thus known with mathematical certainty. Penrose is concerned only with what he calls "unassailable" truths.

We can put the presuppositions succinctly, and sharply delineate the mechanist dispute. Both parties assume that there is a set **K** consisting of all and only the analogues of arithmetic theorems, sentences in the language of first-order arithmetic that can be known with unassailable, mathematical certainty. Let us call **K** the set of *knowable* or provable arithmetic sentences.[2] For convenience, we acquiesce in the sloppy custom of identifying sentences with their Gödel numbers, and so we think of **K** ambiguously as a set of sentences and a set of numbers. The principals to the debate assume that **K** has sharp borders, and so we can inquire about its properties as a set. In particular, we can inquire about its arithmetic and computational properties. The mechanist asserts that all human arithmetic procedures are effective algorithms. With Church's thesis, our mechanist thus holds that there is a Turing machine that enumerates **K**. In other words, he claims that **K** is recursively enumerable. Against this, Lucas and Penrose argue that some of the routines and procedures that humans can employ—and that idealized humans do employ—cannot be simulated on a Turing machine. There are inherently $non - computational$ human arithmetic *procedures*. Lucas [19, p. 105] puts the issue in these terms: "Having once got the hang of the Gödelian argument, the mind can adapt it appropriately to meet each and every variant claim that the mind is essentially some form of Turing machine." Lucas argues that the incompleteness theorem suggests a certain procedure that a human can "get the hang of" and wield against the mechanist.

[2] I have found that many philosophers dismiss the whole Lucas-Penrose controversy, often by rolling their eyes. Penrose [23] suggests that some of them just accept, as a matter of faith, that any cognitive process must be computable. However, others may not accept the presuppositions of the issue—the idealizations in particular. This paper is a defense of (some of) the eye-rolling.

§3. Mining some gold: Gödel's Gibbs lecture. Let **T** be the set of truths of first-order arithmetic. By assumption, $\mathbf{K} \subseteq \mathbf{T}$. In his Gibbs lecture [11], Gödel refers to **T** as "objective" mathematics and **K** as "subjective" mathematics. Suppose that the language is bivalent and $\mathbf{K} = \mathbf{T}$. Let Φ be an arithmetic proposition. By bivalence, either $\Phi \in \mathbf{T}$ or $(\neg\Phi) \in \mathbf{T}$. In the former case, $\Phi \in \mathbf{K}$ and so Φ is knowable in principle. In the latter case, $(\neg\Phi) \in \mathbf{T}$ and $(\neg\Phi) \in \mathbf{K}$, and so it is knowable in principle that Φ is false. So if the language is bivalent and $\mathbf{K} = \mathbf{T}$, then for every sentence Φ in the language of arithmetic, our ideal human mathematicians are capable of determining whether Φ is true or false. Every arithmetic sentence is humanly *decidable*.

Tarski's theorem is that **T** is not definable in the language of arithmetic. *A fortiori*, **T** is not recursively enumerable. There is no effective, formal deductive system that has, as theorems, all and only the arithmetic truths. Thus, if $\mathbf{K} = \mathbf{T}$ and every arithmetic truth is ideally humanly provable, then **K** is not recursively enumerable and the mechanist is wrong. End of story.

Mechanism thus entails that $\mathbf{K} \neq \mathbf{T}$. Let $\Phi \in \mathbf{T}$ and $\Phi \notin \mathbf{K}$. Then Φ is an *unknowable* truth. In Gödel's terms, the sentence Φ is *absolutely undecidable*, as is $\neg\Phi$. Even our idealized subjects do not decide the truth value of Φ and thus in a strong sense humans *cannot* know that Φ is true. So, if the mechanist is correct, then there are absolutely undecidable sentences.

Gödel points out that if **K** is recursively enumerable, then there is an absolutely undecidable sentence of the form

$$\forall x_1 \dots \forall x_n \, \exists y_1 \dots \exists y_m \, (Px_1 \dots x_n y_1 \dots y_m = 0),$$

where P is a polynomial of degree 4 or less. As always, his conclusion is careful:

> ... the following disjunctive conclusion is inevitable: *Either mathematics is incompletable in* [the] *sense that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified* ... It is this mathematically established fact which seems to me of great philosophical interest. (p. 310)

In correspondence, Gödel suggested that humans may have the ability to decide every arithmetic truth, in which case there are no absolutely undecidable arithmetic propositions (and so mechanism is false):

> ... human reason is [not] utterly irrational by asking questions it cannot answer, while asserting emphatically that only reason can answer them ... [T]hose parts of mathematics which have been systematically and completely developed ... show an amazing degree of beauty and perfection. In those fields, by entirely

> unsuspected laws and procedures ... means are provided not only for solving all relevant problems, but also solving them in a most beautiful and perfectly feasible manner. This fact seems to justify what may be called "rationalistic optimism". (see Wang [32, pp. 324–326])

In a personal anecdote, Penrose [23, §4.2] expressed similar sentiments:

> I had vaguely heard of Gödel's theorem prior [to the first year of graduate school], and had been a little unsettled by my impressions of it ... I had been disturbed by the possibility that there might be true mathematical propositions that were in principle inaccessible to human reason. Upon learning the true form of Gödel's theorem ... I was enormously gratified to hear that it asserted no such thing; for it established, instead, that the powers of human reason could not be limited to any accepted preassigned system of formalized rules.

On the other hand, perhaps the existence of absolutely undecidable propositions is not that implausible. Boolos [2] wonders why "should there *not* be mathematical truths that cannot be given any proof that human minds can comprehend?" Once again, if **K** is recursively enumerable, then some arithmetic propositions are undecidable "by human reason" even in principle. The mechanist does a modus ponens, while Gödel and Penrose invoke modus tollens. We return to this "rationalistic optimism" in the next section.

What else follows if our mechanist is right and **K** is recursively enumerable? Let e be the Gödel number of a Turing machine that enumerates **K**, so that $\mathbf{K} = W_e$. Assume that the analogues of the Hilbert-Bernays derivability conditions hold. Roughly, (1) For any sentence Φ in the language of arithmetic, if $\Phi \in W_e$ then the arithmetic sentence stating that $\Phi \in W_e$ is in W_e; (2) For any sentences Φ, Ψ in the language of arithmetic, the arithmetic sentence stating that

$$\text{if } (\Phi \rightarrow \Psi) \in W_e \text{ then if } \Phi \in W_e \text{ then } \Psi \in W_e$$

is in W_e; (3) for any sentence Φ in the language of arithmetic, the arithmetic sentence stating that

$$\text{if } \Phi \in W_e \text{ then the statement that } \Phi \in W_e \text{ is in } W_e$$

is itself in W_e. In the terminology of Smullyan [27, Chapter 9], the assumption is that the formula $x \in W_e$ is a "provability predicate". Conditions (1) and (3) are analogues of the idea that if Φ is provable then it is provable that Φ is provable. Condition (2) is analogue of the idea that the provable sentences are closed under modus ponens.[3]

[3]The derivability conditions are needed here since the foregoing discussion turns on the second incompleteness theorem (in order to follow some of the quoted material). In some

Let Con_e be the usual arithmetic statement that W_e is consistent. By hypothesis, Con_e is true. Gödel's second incompleteness theorem entails that Con_e is not in $\mathbf{K}$. Thus, under the mechanistic assumption, Con_e is true but unknowable, and thus is absolutely undecidable in Gödel's sense. No human, no matter how idealized, could know that W_e is consistent. That is, if $\mathbf{K} = W_e$ and the derivability conditions hold, then no one can know of e that W_e is consistent. It follows that no human, no matter how idealized, can know that every sentence in W_e is true, since he would then know that W_e is consistent. In other words, no human could know that $W_e \subseteq \mathbf{T}$. The inclusion $\mathbf{K} \subseteq \mathbf{T}$ follows from the platitude that only truths are knowable. Thus, no one who knows the platitude can know that $\mathbf{K} = W_e$. In other words, even if the mechanist is right, there is no Turing machine T such that we could know (*de re*) that T enumerates all and only the knowable arithmetic sentences.

Mathematical knowability, and not just mathematical truth, is indefinitely extensible in Dummett's sense. If we know that a given Turing machine generates only knowable arithmetic truths, then we can effectively find a knowable sentence not so generated. Let me quote Gödel [11, p. 309] at length:

> It is [the second incompleteness] theorem which makes the incompletability of mathematics particularly evident. For, *it makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics*. If someone makes such a statement he contradicts himself . . . For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent. Hence, he has a mathematical insight not derivable from his axioms. However, one has to be careful in order to understand clearly the meaning of this state of affairs. Does it mean that no well-defined system of correct axioms can contain all of mathematics proper? It does, if by mathematics proper is understood the system of all true mathematical propositions [**T**]; it does not, however, if one understands by it the system of all demonstrable mathematical propositions [**K**] . . . [A]s to subjective mathematics [**K**], it is not precluded that there should exist a finite rule producing all its evident axioms. However, if such a rule exists, we with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all propositions it produces are correct.

---

cases, however, analogous points could be made by invoking the first incompleteness theorem, thus bypassing the derivability conditions.

Note the similarity to Paul Benacerraf's [1] response to Lucas [18]. If suitably idealized versions of us humans are Turing machines, then they cannot fulfil the Socratic charge: know thyself. If the ideal human is a Turing machine, he cannot know which Turing machine he is.

It follows that we cannot get ourselves in position to use W_e to generate new mathematical knowledge, since to do that, we would have to know that W_e is sound. Löb's theorem sharpens this. Let Φ be any sentence in the language of first-order arithmetic. Let $\mathbf{PR}_e$ be the usual provability predicate for the formal system corresponding to W_e and let $\ulcorner\Phi\urcorner$ be the Gödel number of Φ. Consider the arithmetic sentence:

$$\mathbf{PR}_e(\ulcorner\Phi\urcorner) \rightarrow \Phi.$$

The Löb result is that this sentence is in W_e (and thus $\mathbf{K}$) only if Φ is itself in W_e. That is, there is no unknowable sentence Φ such that we can know that *if* Φ is in W_e then Φ is true. In other words, there is no non-trivial hypothetical knowledge about the contents of W_e. By hypothesis, a sentence Φ is knowable if and only if it is in W_e. For a particular sentence Φ, we can *know that*

$$\Phi \text{ is knowable if and only if it is in } W_e$$

only if Φ is knowable.[4]

As indicated by the care in Gödel's reasoning, the mechanist has some room to maneuver. The postulated set $\mathbf{K}$ consists of the knowable sentences in the language of first-order arithmetic. Recall that both the mechanist and the opponent assume that our subjects are following procedures that deliver only sentences that can be proved and thus known with absolute mathematical certainty. Suppose that $\mathbf{K} = W_e$ and the derivability conditions hold. All that follows from the fancy Gödel (and Lucas-Penrose) arguments is that neither our idealized subjects nor we mortal humans can know (*de re*) that W_e is sound or even consistent *with that same absolute mathematical certainty*. There is nothing (so far) preventing us from concluding the soundness of W_e on less than absolutely certain evidence. The mechanist is free to argue for his thesis on empirical grounds or even on some sort of *a priori* metaphysical grounds short of mathematical proof.

In the quoted passage from the Gibbs lecture [11, p. 309], Gödel points out that someone contradicts himself if he puts forward "a certain well-defined system of axioms and rules" and claims to "perceive (with mathematical certitude)" that the axioms and rules are correct and they contain all of mathematics. The parenthetical qualification is crucial. Gödel leaves it open that a fixed formal system can reproduce "the system of all *demonstrable* mathematical propositions" (i.e., $\mathbf{K}$), but no one can claim to

[4]This conclusion is in the neighborhood of Michael Detlefsen's contribution to the conference on languages, minds, and machines mentioned in the acknowledgments.

know *with mathematical certitude* that the axioms and rules in question are correct. In a footnote (11), he states that a mechanist can consistently put forward a formal system as a candidate for **K** and claim "I believe I shall be able to perceive one after the other of the theorems to be true". Gödel goes on to argue that the soundness of the system can "at most be known with empirical certainty, on the basis of a sufficient number of instances or by other inductive inferences". He elaborates the possibility in a pair of footnotes (12, 14):

> ... it is conceivable (although far outside the limits of present-day science) that brain physiology would advance so far that it would be known with empirical certainty (1) that the brain suffices for the explanation of all mental phenomena and is a machine in the sense of Turing; [and] (2) that such and such is the precise anatomical structure and physiological functioning of the part of the brain which performs mathematical thinking.
>
> ... the physical working of the thinking mechanism could very well be completely understandable; the insight, however, that this particular mechanism must always lead to correct (or only consistent) results would surpass the powers of human reason.

By "the powers of human reason", Gödel must mean something like "absolutely certain, mathematical knowledge".

Two decades later, Kreisel [14, p. 322] comes to a similar conclusion:[5]

> ... it has been clear since Gödel's discovery of the incompleteness of formal systems that we could not have *mathematical* evidence for the adequacy of any formal system; but this does not refute the possibility that some quite specific system F ... encompasses all possibilities of (correct) mathematical reasoning ... In fact the possibility is to be considered that we have some kind of nonmathematical evidence for the adequacy of such an F.

We now move ahead two more decades, to Lucas's [19] reply to objections. As above, let S be a formal system (or Turing machine) that is put forward by a mechanist as a model of human arithmetic knowability (i.e., **K**). Let G_S be a Gödel-sentence for S and let Con_S be the usual statement that S is consistent. Lucas claims that under these circumstances, he can know that G_S is true. Putnam [24] points out that neither Lucas nor anyone else knows

[5]Kreisel expresses skepticism about evidence on these matters: "Closer inspection shows that we have ... very little experience of establishing such mathematical assertions as soundness or consistency by inductive methods and thus we have little knowledge of the *statistical principles* proper to evaluating hypothetical inductive evidence." He suggests the "less well-known ... possibility of establishing the soundness of F by *abstract*, but *nonmathematical* interpretation", but he does not elaborate this very far. Kreisel concludes that if one entertains the mechanistic thesis, "one *has* to consider unfamiliar principles of evidence, such as those involved in the inductive or philosophical approaches just mentioned" ([14, p. 323]).

that G_S is true. He only knows that *if S is consistent* then G_S is true. But the machine (or formal system) "knows" this conditional proposition as well, since

$$\mathrm{Con}_S \to G_S$$

is a theorem of S (as seen by the proof of the second incompleteness theorem). Lucas can claim to know G_S outright only if he can claim to know Con_S. But how does he establish this last premise?

Lucas's response is to shift the burden of proof. It is up to the mechanist who proposes S to show that S is consistent. If S is not consistent, then we need not take the mechanist seriously. If the mechanist can establish the consistency of his proposed model, then Lucas has the premise he needs to conclude G_S:

> Putnam's objection fails on account of the dialectical nature of the Gödelian argument. The mind does not go round uttering theorems in the hope of tripping up any machines that may be around. Rather, there is a claim being seriously maintained by the mechanist that the mind can be represented by some machine. Before wasting time on the mechanist's claim, it is reasonable to ask ... some questions about [the] machine to see whether [the] seriously maintained claim has serious backing. It is reasonable to ask ... whether [the] machine is consistent. Unless it is consistent, the claim will not get off the ground. If it is warranted to be consistent, then that gives the mind the premiss it needs. The consistency of the machine is established not by the mathematical ability of the mind, but on the word of the mechanist. (Lucas [19, p. 117])

The Gödel-Kreisel analysis reveals the error in the last sentence. The hypothetical mechanist's claim is that the system S represents all and only the arithmetic sentences that the mind can prove—the sentences that an idealized Lucas (say) can know with mathematical certainty. Lucas asks the mechanist if S is consistent. The mechanist replies "Yes, I think so. I would not have put S forward if I did not believe it to be consistent." This "word of the mechanist" does not give Lucas a premise he can use (in a modus ponens on $\mathrm{Con}_S \to G_S$) simply because this "word" does not amount to mathematical certainty. An idealized mathematician cannot invoke someone's word to justify a line in a derivation.

Lucas can surely demand that the mechanist convince us that S is consistent. Otherwise, her serious claim about S deserves little credence. If the mechanist manages to *prove* that S is consistent (Con_S), with all the rigor of mathematics, then Lucas wins the round. He just points out that Con_S is now known (by the mechanist and thus by Lucas) with mathematical certainty even though Con_S is not a theorem of S. In this scenario, the mechanist herself goes beyond S when she establishes Con_S. However, with Gödel and Kreisel, all we can conclude is that it is too much to ask that the mechanist establish her claim with mathematical certainty. It is still open for the mechanist to provide compelling non-mathematical arguments in support of the claim that S works, and thus that S is consistent.

Lucas summarizes that the "mechanist has claimed that his machine is consistent. If so, it cannot prove its Gödelian sentence, which the mind can none the less see to be true." There is a crucial ambiguity in the locution "see to be true". If it means "prove", then (for all we know so far) the mind can do no such thing. If "sees to be true" means something weaker, like "has excellent reason to believe", then there is no reason to expect G_S or Con_S to be in $\mathbf{K}$ and thus be knowable with mathematical certainty—unassailable as Penrose puts it.

Penrose [22, p. 76] comes to a measured conclusion: "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth." He concedes that the existence of an algorithm that enumerates $\mathbf{K}$ is a bare "logical possibility". However, he later [22, §§3.16, 3.23] presents an ingenious new argument aimed at showing that our idealized mathematician cannot consistently *believe* (on any grounds) that a given algorithm enumerates the provable arithmetic (or even Π_1) sentences. If correct, the new argument refutes the very possibility that $\mathbf{K}$ is recursively enumerable. To put the argument in present terms, let $\mathbf{K}'$ be the set of arithmetic sentences that idealized humans (unassailably) know to *follow from the hypothesis that* $\mathbf{K} = W_e$. That is, Φ is in $\mathbf{K}'$ if and only if the idealized human (unassailably) knows that if $\mathbf{K} = W_e$ then Φ is true.[6] Our idealized human can then reason as follows:

> Assume that $\mathbf{K} = W_e$. Then every member of $\mathbf{K}'$ is true and thus $\mathbf{K}'$ is consistent. Under the assumption $\mathbf{K} = W_e$, $\mathbf{K}'$ is presumably recursively enumerable and I (the idealized subject) can write a Gödel sentence $\mathbf{G}'$ for $\mathbf{K}'$ (by using e). So under the assumption $\mathbf{K} = W_e$, if $\mathbf{K}'$ is consistent, then $\mathbf{G}'$ is true but not in $\mathbf{K}'$. But we just saw that under the assumption, $\mathbf{K}'$ is consistent. So under the assumption, $\mathbf{G}'$ is true and not in $\mathbf{K}'$. This is a contradiction. Thus, $\mathbf{K} \neq W_e$.

In short, Penrose applies the Gödel-Kreisel-Lucas construction to the set of arithmetic sentences that the idealized human can determine *from the assumption that* $\mathbf{K} = W_e$. Nifty. There is not sufficient space here to go into this argument in detail. Notice, however, that the envisioned set $\mathbf{K}$ only contains (Gödel numbers of) *arithmetic* sentences and the original argument concerned the idealized subjects *mathematical* knowledge about the natural numbers—what he can prove within mathematics. In the new argument, the

[6] A referee pointed out that if in fact $\mathbf{K} \neq W_e$ then for every Φ, the sentence "if $\mathbf{K} = W_e$ then Φ" is true. If the idealized human knows that $\mathbf{K} \neq W_e$ then $\mathbf{K}'$ contains every sentence.

idealized subject bandies about assumptions concerning what he can and cannot unassailably know. Thus, to make the new argument rigorous, we need a term like "**K**" to be in the language of the ideal knower, with sentences containing **K** and the like subject to logical and mathematical analysis. But a straightforward diagonal argument shows that knowability is not definable (see Chalmers [4] and §6 below). Moreover, in order to state the soundness assumption (that every member of **K** and **K**′ is true), our subjects require a comprehensive truth predicate. Third, to construct the Gödel sentence for **K**′, our subject needs a Turing machine which determines how he unassailably reasons from non-mathematical assumptions (like $\mathbf{K} = W_e$). Why think the ideal subject can get this from the assumption that $\mathbf{K} = W_e$?

Penrose [23, §3] suggests that we can avoid the problems with defining truth and knowability if we start the argument with the assumption that the knowable Π_1-sentences are enumerated by a given Turing machine. Then we only need a notion of Π_1-truth, which is definable in arithmetic. The analogue of **K**′ would be the set of Π_1-sentences that the subject knows to follow from the assumption that the knowable Π_1-sentences are enumerated by the given Turing machine. But our subject now has no means to construct a Gödel sentence for the new **K**′ from the given Turing machine. How do we go from a Turing machine that enumerates Π_1-sentences to one that enumerates Π_1 consequences of a non-Π_1 assumption? We need a Turing machine for what Penrose calls "Gödelian reasoning", and the restricted hypothesis does not give us one. Also, why does the ideal subject hold that the Π_1 consequences of the non-Π_1 assumption are all true (even given the assumption)? The only reason I can think of is that the subject believes that *all* of his beliefs are true. But this is general soundness which invokes a general notion of truth. Without something like a general soundness assumption our subject has no reason to think that the non-Π_1 assumption is consistent with what he (thinks he) unassailably knows.

Penrose [23, §3.13] agrees that once we start dealing with what can and cannot be known or believed "it is important to put *some* restriction on the type of sentence to which the belief system is applied". The problem is that once such a restriction is put in place, the ideal subject must go beyond the restriction to carry out the new argument.

§4. Running up the ordinals. Does it help? No, not really. In developing techniques for iterating the Gödel construction into the transfinite, logicians have come up with some interesting results. For present purposes, however, neither the mechanist nor his opponent get much aid or comfort—but perhaps some of the issues and presuppositions are further illuminated.

Feferman [9] makes an important definition:

> By a *reflection principle* we understand a description of a procedure for adding to any set of axioms A certain new axioms whose

validity follow from the validity of the axioms A and which formally express, within the language of A, evident consequences of the assumption that all of the theorems of A are valid.

In arithmetic, Feferman's technical notion of "valid" comes to "true". A reflection principle produces new axioms for a theory A, which a theorist should accept if she already holds that every theorem of A is true. Non-trivial reflection principles give rise to the indefinite extensibility of arithmetic truth.

The reflection principle in Turing's original [30] is the Gödel sentence G_A or, equivalently, the formal consistency Con_A. As we have seen, over and over, if someone holds (for whatever reason) that every theorem of A is true, then she should also hold G_A or Con_A. Feferman [9] considers other reflection principles, the most powerful of which is a reflection ω-rule of sorts. Let A^* consist of the axioms of A together with every instance of

$$\forall x\ \mathbf{Pr}_A(\ulcorner\Phi(x)\urcorner) \to \forall x\ \Phi(x),$$

where $\mathbf{Pr}_A(\ulcorner\Phi(x)\urcorner)$ is a formula stating that the result of substituting the appropriate numeral for x in $\Phi(x)$ is provable in A.

These reflection principles can only be applied to theories A whose proof-predicate is defined in the language of A. If the theorems of A are not arithmetic, then there may be no sentence expressing the consistency of A, let alone Feferman's richer reflection principle.

The program is to repeatedly iterate a reflection principle P through recursive ordinals. Let A be a base theory, such as standard Peano arithmetic. As a first approximation, let $A(0)$ be A and, for each ordinal α, if $A(\alpha)$ is defined, then let $A(\alpha+1)$ be the theory consisting of $A(\alpha)$ together with the result of applying P to $A(\alpha)$. In Turing's case, where P is the consistency statement, $A(\alpha+1)$ is $A(\alpha)$ together with $\mathrm{Con}_{A(\alpha)}$. In Feferman's strongest case, $A(\alpha+1)$ is $A(\alpha)^*$. If λ is a limit ordinal and $A(\alpha)$ is defined for every $\alpha < \lambda$, then the theorems of $A(\lambda)$ are the union of the theorems of $A(\alpha)$ for every $\alpha < \lambda$.

For the just mentioned reflection principles, if $A(\alpha)$ is recursively enumerable, then so is $A(\alpha+1)$. If λ is a recursive limit ordinal and if the members of $\{\, A(\alpha) \mid \alpha < \lambda \,\}$ are uniformly recursively enumerable, then $A(\lambda)$ is itself recursively enumerable. Thus, theories "reflected" through recursive ordinals might make an interesting case study for mechanism. Can we consider the (non-recursive) theory consisting of the union of all $A(\alpha)$ for all recursive ordinals α?

Not yet. The theory $A(\alpha)$, as presented so far, is not well-defined since the reflection principles are not extensional. If B is a theory, then the sentence Con_B and the theory B^* depend not just on the theorems of B, but on how B is given. For example, there are theories B and C that have the same theorems, but where Con_B is not equivalent to Con_C (see Feferman [8]). The intensionality comes up again at limit ordinals. To apply the reflection principle at stage $\lambda + 1$, we need not just the theorems of $A(\lambda)$ but also a *description* of those theorems, and this depends on a description of λ. If we have two different descriptions of λ, we can end up with two different theories $A(\lambda + 1)$.

The intensionality here serves as a reminder that neither actual nor ideal humans deal with theories as such, but only with theories under a description. Consequently, the mechanist must follow suit and deal with theories under a description.

Feferman invokes a common notation for recursive ordinals, which we adopt here. Ordinals are "denoted" by natural numbers. The number 1 denotes the ordinal 0. If n denotes an ordinal α, then 2^n denotes its successor $\alpha+1$. If e is the Gödel number of a Turing machine that enumerates numbers denoting an increasing sequence of ordinals, then $3 \cdot 5^e$ denotes the limit of that sequence. Let O be the set of natural numbers that denote ordinals on this notation, and if $m \in O$ then let $|m|$ be the ordinal denoted by m. The set O is not recursively enumerable—not by a long shot (see Rogers [25, pp. 205–210]).

Let P be a reflection principle. Let $R(1)$ be a standard enumeration of the theorems of A. If $n \in O$, then let $R(2^n)$ be an enumeration of the result of applying the reflection principle to $R(n)$ (under that description). If e is the Gödel number of a Turing machine that enumerates numbers denoting an increasing sequence S of ordinals, then let $R(3 \cdot 5^e)$ be a uniform enumeration of the union of the sets $R(s)$ for $s \in S$.

The idea is that for each $n \in O$, $R(n)$ is the theory $A(|n|)$—under that description. The notation makes the intensionality explicit, since the theorems of $R(n)$ depend not just on $|n|$ but also on n. For the reflection principles under study here, we can take R to be a total recursive function on the natural numbers. If $n \in O$ and the base theory A consists only of truths, then $R(n)$ also contains only truths.

In these terms, Turing's [30] plan was to overcome incompleteness by using theories like $R(n)$, with n ranging over O. He showed that for the simple reflection principle Con_A, if Φ is a true Π_1-sentence, then there is an $n \in O$ (which can be found effectively from Φ) such that $|n| = \omega + 1$ and Φ is among the theorems of $R(n)$. This astounding result is that there is a way to iterate the Gödel construction on theories, beginning with A, so that when we collect together the finite iterations and take one more Gödel sentence, Φ is decided.[7] There is thus a certain completeness for Π_1-sentences.

Feferman extended this result. With his stronger reflection principle A^*, he showed that for *any* true sentence Φ in the language of arithmetic, there is a number $n \in O$ such that Φ is among the theorems of $R(n)$. That is,

[7] Turing [30] did not use the present notions, but his theorem is equivalent to the one stated here. See Feferman [10, §7] for readable sketches of Turing's results.

for any sentence Φ, there is a way to iterate the reflection principle (up to a small transfinite level) and decide Φ.

Turing [30, §9] was aware that his completeness result does not provide the wherewithal to decide the truth of any new arithmetic sentences: "This completeness theorem ... is of no value. Although it shows, for instance, that it is possible to prove Fermat's last theorem [with $R(n)$] (if it is true) yet the truth of the theorem would really be assumed by taking a certain" number as a member of O. Suppose that a mathematician wants to decide the truth value of the Goldbach conjecture. He calculates a number n and starts enumerating the theorems of $R(n)$, looking for the Goldbach conjecture among the output. So far so good, since all this is effective. The mathematician knows that if the Goldbach conjecture is true, then $n \in O$ and so every sentence in $R(n)$ is true. However, an examination of Turing's proof shows that if the Goldbach conjecture is false, then n is not in O, and, even worse, $R(n)$ is inconsistent. Thus, the results of the enumeration can be believed only if the Goldbach conjecture is true. This is of no help whatsoever in trying to determine the truth value of the Goldbach conjecture.

The same goes for the Feferman result. For each sentence Φ we get a formal system F that is sound if Φ is true, and if Φ *is* true, F proves it. Nothing to celebrate here. We can get that much just by adding Φ as a new axiom to the base theory A. Feferman [9, p. 262] concludes that "questions of completeness of sequences derived from progressions hinge ... on more subtle questions on how paths through O are obtained". He elaborates:

> Whenever [a mathematician] is given the information $d \in O$ he will be able to compute $[R(d)]$ and prove theorems from $[R(d)]$; moreover, if he accepts the information that $d \in O$ he should find all these theorems acceptable. Unfortunately, ... as he advances farther and farther out into the collection of systems $[R(d)]$, he may not be able to gain the knowledge necessary to decide, of any given d_0, whether or not $d_0 \in O$. In other words, in order to proceed, he may have to appeal to an 'oracle'. (p. 279)

The key to wielding these results is the ability to decide membership in O, or to find effective notations for recursive ordinals generally.

In these terms, the Lucas-Penrose contest to write and assert Gödel sentences becomes a contest to enumerate recursive ordinals. One might think that all Lucas has to do is iterate the procedure of adding Gödel sentences (or the Feferman reflection principle) far enough. The problem, however, is with the crucial notion of "far enough". At some point, we are no longer sure we are on the right road.

No machine can iterate the procedure through all and only the recursive ordinals. Can Lucas? In [19, p. 110], he envisions the shift from the Gödel sentences to the construction of ordinals (although he does not explicitly restrict the discussion to recursive ordinals). He imagines a mechanist designing machine after machine in an effort to confound the Gödelizing. Lucas keeps winning and the mechanist keeps designing new machines:

> Every now and again the mechanist loses patience, and incorporates in his machine a[n] ... operator designed to produce in one fell swoop all the Gödelian sentences the mentalist is trumping him with: this is in effect to produce a new limit ordinal. But such ordinals, although they have no predecessors, have successors just like any other ordinal, and the mind can out-Gödel them by producing the Gödelian sentence of the new version of the machine, and seeing it to be true, which the machine cannot.

Douglas Hofstadter [13, p. 475] questioned Lucas's confidence, citing the Church-Kleene theorem "that we cannot program a machine to produce names for all the ordinal numbers", as Lucas put it. Of course, there is no machine that produces names for "all the ordinals", because the collection of ordinals is a proper class, but this is not relevant here. The Church-Kleene theorem is that there is no *recursive* enumeration of every *recursive* ordinal (in a way that allows us to determine the order type of each). The non-recursiveness of O is a corollary.

There are different anti-mechanist claims that might be made at this point. The strong one is that a human can enumerate the members of O (or some other effective notation for all recursive ordinals). This is a reasonably sharp thesis, assuming we can make sense of the idealizations (see §2 above). Presently, we will examine just how plausible the view is. A weaker anti-mechanist claim is that if an idealized human is given a machine M that produces names of recursive ordinals (e.g., members of O), she can produce a name of an ordinal not produced by M. In the latter case, the human can carry the reflection principles further than M, and so this human is not identical to M in any interesting sense. However, given the parameter 'M', the thesis is pretty vague, perhaps obscure.

Lucas argues that Hofstadter begs the question in the assumption that "since there is no mechanical way of naming all the ordinals, the mind cannot do it either". Lucas says that "this is precisely the point at issue". If this *is* the point at issue, then Lucas makes the strong claim that the mind can enumerate all of the recursive ordinals (i.e., the members of O)—if not all of the ordinals.

Clearly, the presumed ability to enumerate O goes well beyond any Turing machine. So if Lucas could give some reason to think humans have this ability, he wins—hands down. So far as I can tell, however, the limitative theorems do not give a reason to think any suitably idealized human has the ability to enumerate O, and so the limitative theorems do not support this strong anti-mechanist scenario.

Moreover, the view is extremely bold. The Feferman result entails that if a human can iterate the members of O (or if she can decide membership in O) then she has the wherewithal to determine the truth value of *every* arithmetic sentence. All she has to do is systematically generate the outputs of $R(n)$, for each $n \in O$, using the Feferman reflection principle. For any arithmetic sentence Φ, either Φ or $\neg\Phi$ will eventually turn up, and the other one never will. Thus, if a being could enumerate O, then it could infallibly determine the truth value of any arithmetic sentence. When it comes to arithmetic truth, we would be not only infallible, but omniscient. A wonderful thought.

Lucas cites the authority of Gödel (see, Wang [32, pp. 324–326]) as "rejecting mechanism on account of our ability to think up fresh definitions for transfinite ordinals". Although Gödel's concerns here were with large cardinals and not with countable recursive ordinals, he did demur from mechanism. Recall his conclusion that the limitative theorems show that *either* mind is not a machine *or* there are "absolutely undecidable" arithmetic propositions (see §3 above). He demurred from the second disjunct and leaned toward the first. Gödel held that humans are not machines and we are arithmetically omniscient. We saw that Penrose also adopted this potential omniscience, which Gödel called "rationalistic optimism". Lucas is free to state that humans can enumerate O and thus join them in this rationalistic optimism. However, as Gödel pointed out, the limitative theorems do not support this conclusion.

To defeat the mechanist, perhaps Lucas does not need the strong claim that some human can enumerate O. It is enough if for any given machine M put forward by a mechanist, Lucas can "out-enumerate" that machine—if he could enumerate more members of O than M can. We must be careful. It is not at all clear just what this weak anti-mechanistic claim is. Suppose that Lucas claims that for any Turing machine M that enumerates only members of O, there is an idealized human h such that h can enumerate more members of O than M. This claim has the $\forall M\ \exists h$ form, allowing that different Turing machines M may get trumped by different idealized humans h. The claim does not undermine mechanism at all. For any Turing machine M that enumerates only members of O, there is another *Turing machine* that enumerates the same numbers that M does, plus a few more members of O. This is an analogous $\forall M\ \exists M'$ theorem.

Perhaps Lucas means that there is a *single* idealized human who can out-enumerate *any* machine that enumerates only members of O. This $\exists h\ \forall M$ claim amounts to the strong thesis that the human can enumerate all of O. Indeed, for any $n \in O$, there is a Turing machine that enumerates n (and nothing else, say). So our postulated ideal human h must also be capable of knowing that $n \in O$.

Perhaps the weak anti-mechanist claim is that there is a single idealized human L such that for any *given* machine M that a mechanist seriously puts forward as a candidate for human recursive-ordinal-generation competence, L can generate a member of O not generated by M. The claim might be that humans can use an ordinal-generating procedure that cannot be executed by any machine. Lucas [19, p. 113] suggests something like this, when he accuses Hofstadter of misconstruing the nature of the contest:

> All the difficulties are on the side of the mechanist trying to devise a machine that cannot be out-Gödelized. It is the mechanist who resorts to limit ordinals, and who may have problems devising new notations for them. The mind only needs to go on to the next one, which is always an easy, unproblematic step, and out-Gödelize whatever is the mechanist's latest offering.

Suppose that the mechanist produces a number e which is the code of a Turing machine that enumerates members of O (with the denoted ordinals in increasing order). Then Lucas calculates $3 \cdot 5^e$, which denotes the next ordinal greater than any enumerated by M. But there is nothing non-mechanical going in this calculation, just as there is nothing non-mechanical involved in writing out a Gödel sentence. Lucas is also correct that the "next" ordinal is always an easy, unproblematic step. To go from $|n|$ to $|n| + 1$, we just calculate 2^n. The calculations are grade school drills (involving *very* large natural numbers).

So what exactly is this ability that Lucas claims on behalf of ideal humans? We are back where we were before we got fancy with ordinals. For any given machine M, Lucas cannot take the "easy, unproblematic step" (to achieve knowledge of a new member of O) unless he *already knows* that M produces only members of O. How does Lucas get this prior knowledge? As before, he shifts the burden to the mechanist. If she *proves* that M produces only members of O, then she has already managed to go beyond M. However, if the mechanist only claims some sort of non-mathematical evidence on behalf of M, the Gödel-Turing-Feferman results do not apply. Lucas only knows that *if* e is the Gödel number of a Turing machine that enumerates an increasing sequence of recursive ordinals, then $3 \cdot 5^e$ denotes an ordinal greater than any enumerated by that machine, and Lucas knows that *if* n is a member of O, then 2^n denotes a greater ordinal. But these conditional statements are theorems of the base theory A, and so are "available" to all of the Turing machines.

How does Lucas get beyond the conditionals, to their consequents? He claims some sort of insight, not available to any of the mechanist's Turing machines. He writes that in the ordinal-writing duel with the mechanist, "every now and again some new, creative step is called for, when we consider all the ordinal numbers hitherto named" ([19, p. 111]). Lucas is correct that to defeat mechanism, it must be a *creative* step and not the application of an algorithm—not the mere construction of a Gödel sentence or the calculation of a new member of O. What is this creative step, and how do the incompleteness theorems indicate that we have it?

Lucas cites Turing as an ally in the weakened anti-mechanistic claim. In [30, §11], Turing defines the "activity of intuition" as the "making of spontaneous judgments which are not the result of conscious trains of reasoning". Turing suggests that "in pre-Gödel times" it was hoped that formalization could be developed "to such a point that all the intuitive judgements of mathematics could be replaced by a finite number of ... rules. The necessity for intuition would then be entirely eliminated". This is undoubtedly a reference to the Hilbert program, which was all but killed by the incompleteness results (Detlefsen [5] notwithstanding). Turing suggests that at this point,

> ... we naturally turn to 'non-constructive' systems of logic [in] which not all the steps in a proof are mechanical, some being intuitive. An example of a non-constructive logic is afforded by any ordinal logic. When we have an ordinal logic, we are in a position to prove number-theoretic theorems by the intuitive steps of recognizing formulae as ordinal formulae ...

In present terms, Turing's "ordinal logics" are like our systems $R(n)$, and the recognition of a formula as an ordinal formula is equivalent to the recognition of a natural number as a member of O.

Lucas concludes that Turing, "like Gödel, allows that the mind's ability to recognize new ordinals outruns the ability of any formal algorithm to do so". The key word here, I think, is "recognize". Lucas is not just attributing to his idealized self the ability to write out Gödel sentences or to print natural numbers, nor a mere ability to assert the conditional sentences. It is an *epistemic* ability that supposedly defeats the mechanist. The "creative step" is the presumed ability to *see that* every member of a given sequence of numbers is a member of O (or denotes a recursive ordinal), or the ability to *see that* every theorem of a given formal theory is true. Similarly, Penrose speaks often of the human ability to "understand", which any algorithmic device lacks.

Lucas's proposal seems to be a variant of the Gödel-Kreisel suggestion that humans, unlike machines, can traffic in abstract concepts. Here, the thesis is that humans are capable of epistemic states like "recognition" or "intuition". Unlike Lucas and Penrose, however, Gödel and Kreisel were aware that even if they are right about the non-mechanical nature of the human mind, there is still a burden to show that humans can out-perform any machine.

Again, the incompleteness theorems concern formal systems, algorithms, and Turing machines. To "apply" these theorems we need a sharp thesis. In this case, it is no longer clear what mechanistic thesis Lucas is refuting. What does it mean to say that an algorithm, a formal system, or a Turing machine "sees that" a number is a member of O, beyond printing out the relevant formula? In the indicated footnote to [30, §11], Turing wrote that the requirements on intuition are "very vague". We need some precise content to the supposed mechanistic theses before Lucas can wield the Gödel theorem against them.

For these reasons, I question Lucas's exegetical claim that Turing, "like Gödel, allows that the mind's ability to recognize new ordinals outruns the ability of any formal algorithm to do so". In the much reprinted "Computing machinery and intelligence" [31], Turing rejected as meaningless the question of whether machines can think. Presumably, he would also reject as meaningless the question of whether machines are capable of "recognizing" things or of having "understanding" or "intuition"—let alone the question of whether our intuition outruns theirs.[8] He proposed the Turing test, or the imitation game, as a scientifically sound substitute for these philosophical questions.

Suppose that we grant human intuition, or the ability to deal with abstract concepts. Does this give humans an edge in the Turing test? None of the present authors provide an argument that it does, and I do not see the relevance of the limitative theorems. The Turing test is to be played with real humans, not idealized specimens who never make mistakes. Moreover, the limitative theorems would be useful in the Turing game only if the human players are given the program for the machine. But in that case, the human's presumed victory would not refute anything in the neighborhood of mechanism.

§5. **Idealizations revisited.** Recall the "normative idealization" that the subjects proceed by absolutely certain methods, and they produce all and only the mathematically provable arithmetic sentences. Their productions are unassailable. These idealizations allow for the application of the limitative theorems since, in effect, we make the productions of our subjects much like a deductive system.

The normative idealization is consonant with a longstanding epistemology for mathematics. The idea is that for mathematics at least, real humans are capable of proceeding, and should proceed, by applying infallible methods. In practice (or performance) we invariably fall short of this, due to slips of the pen or faulty memory, but in some sense we are capable of error-free mathematics. We start with self-evident axioms and proceed by gap free deduction. Call this the *Euclidean* model of mathematics.

In the Gibbs lecture, Gödel [11, p. 305] comes close to endorsing the Euclidean model:

[8] Turing did predict that language use would evolve to the point where we speak of machines as "thinking" and, presumably, having intuitions. At that point, there might be a legitimate (empirical?) question of how machine intuition stacks up against human intuition.

> [The incompleteability of mathematics] is encountered in its simplest form when the axiomatic method is applied, not to some hypothetico-deductive system such as geometry (where the mathematician can assert only the conditional truth of the theorems), but to mathematics proper, that is, to the body of those mathematical propositions which hold in an absolute sense, without any further hypothesis ... [T]he task of axiomatizing mathematics proper differs from the usual conception of axiomatics insofar as the axioms are not arbitrary, but must be correct mathematical propositions, and moreover, evident without proof.

Penrose [22, Chapter 3] also seems to be in the neighborhood of the Euclidean model, with his central notion of "unassailable knowledge". He admits that even ideal subjects may be subject to error, but he insists that all errors are "correctable". Penrose sets up a plan to eliminate errors by having the subjects check each other and he provides a detailed argument that there are only finitely many sentences that need to be certified as genuinely unassailable. In a sense, Penrose maintains the Euclidean model statistically.

The treatment in the previous sections presupposes the Euclidean model, in that we assumed that there is such a thing as the set **K** of arithmetic sentences that are knowable with mathematical certainty. We saw Gödel and Kreisel resolve the dilemma by invoking "empirical" or "inductive" methods short of proof (although Kreisel reminds us that we do not have a developed epistemology for those).

Although the traditional Euclidean model has advocates today (e.g., Tennant [28], [29]), it is under serious challenge. The most popular contender (at least in North America) comes from the Quinean thesis that our beliefs are a seamless web answerable only to sensory input. The very idea of infallible methods goes the way of analytic truth and *apriori* knowledge. There is no difference in principle between a mathematical proof, an entrenched scientific thesis, and a hypothetico-deductive inference. Everything is up for revision. In principle, nothing is unassailable-in-principle.

So perhaps we need to rethink the normative idealization, so as to not tie the Lucas-Penrose argument to a contentious epistemology. Lucas himself [19, p. 120] says that to "claim to know something is not to claim infallibility, but only to have adequate backing for what is asserted". If this comment applies to both the output of the mechanist's proposed system and its Gödel sentence, then Lucas himself rejects the Euclidean model.

If we give up the normative idealization, then we do not have the above set **K** of sentences knowable with certainty. What do we focus on instead? What is the mechanistic thesis now? Lucas's opponent puts forward a formal system or Turing machine and claims something about the set of sentences it produces. Lucas then calculates a Gödel sentence for the system and makes a knowledge-claim that is supposed to refute his opponent. Whether Lucas is correct depends on what his imagined interlocutor claims on behalf of his system.

Define $\mathbf{K}_1$ to be the set of arithmetic sentences for which we can "have adequate backing", to use Lucas's phrase. Perhaps the interlocutor claims that his Turing machine enumerates $\mathbf{K}_1$ and Lucas proposes his Gödel sentence as a counterexample. Lucas argues that the mechanist must have adequate backing for the statement that his machine is consistent, and this same backing supports the Gödel sentence, even though the Gödel sentence is not produced by the Turing machine in question.

Careful. The mechanist is refuted only if (Lucas knows that) the target set enumerated by his Turing machine is consistent. Do we have adequate backing for the claim that $\mathbf{K}_1$ is consistent? Surely, we can have adequate backing for each member of a large, inconsistent set of sentences. Consider a lottery with 10,000 tickets. For each ticket t, we have adequate backing for the statement that t is not the winning ticket, even though the set of all such sentences is inconsistent with the rules of the lottery. The notion of "having adequate backing" is not monotonic.

Gödel [11, p. 309] stated that if we know that every theorem of a formal system S is true, then we know *with the same certainty* that S is consistent. The statement that S is consistent enjoys at least the level of certainty as the single statement that every theorem of S is true. However, for each axiom Φ of S, we can have good reason to think that Φ is true without having good reason to think that S is consistent. So perhaps we are dealing here with a fancy version of the lottery paradox or the preface paradox.

Surely, we do not have adequate backing for any explicit contradiction, once we realize that it is a contradiction. Thus, the set $\mathbf{K}_1$ might not be closed under deduction. We can have adequate backing for each of a large set of sentences without having adequate backing for every consequence of this set. Admittedly, the situation is paradoxical. Can there be sentences Φ and Ψ such that both Φ and $\Phi \rightarrow \Psi$ are in $\mathbf{K}_1$—we have adequate backing for both—and yet Ψ is not in $\mathbf{K}_1$? All that follows from this consideration is that the notion of "having adequate backing" is vague, and we have a version of the sorites paradox on hand. Unless the set $\mathbf{K}_1$ is precise, we cannot apply the incompleteness theorem to it.

To apply the limitative theorems in the most straightforward manner, we need an epistemological notion that is sharp and closed under deductive consequence, and which the mechanist can plausibly attribute to idealized subjects as the goal of mathematical activity. Even if "knowable with absolute certainty" is too strict a notion, "having adequate backing" is too weak and too vague. Is there something in between?

To make a stab at it, let us define a set Γ of sentences to be *stably-backed* in a given state of information if for each member Φ of Γ, Φ enjoys adequate backing in that state of information and Φ remains adequately backed no matter how many logical consequences of Γ are added to the state of information. That is, Γ is stably-backed if no member of Γ loses its backing as deductions are made on members of Γ. It follows that Γ is stably-backed only if Γ is consistent. Define Γ to be *maximally-stably-backed* if Γ is stably-backed and if there is no proper superset of Γ that is stably-backed. That is, if Γ is maximally-stably-backed then no sentence outside of Γ can get adequately backed without giving up some member of Γ. This notion is a Peircean limit, of sorts.

Suppose that someone comes up with a Turing machine and claims to have adequate backing for the statement that the set S of sentences produced by this machine is maximally-stably-backed. Assuming that S contains some arithmetic, Lucas wins the round. If someone has adequate reason to think that S is stably-backed then she has adequate backing for Con_S and if S is consistent then it does not contain Con_S. So we either give up some member of S or give up the statement that S is *maximally*-stably-backed.

The analogy with **K** is complete. For any number e, we cannot have stable backing for the statement that W_e is maximally-stably-backed (assuming the derivability conditions hold for W_e). That is, once we conclude that W_e is stably-backed, we cannot hold that W_e is maximally-stably-backed. This should not be surprising, since being maximally-stably-backed seems pretty close to the traditional "proved".

This conclusion does not threaten mechanism. How is the mechanist committed to the existence of a maximally-stably-backed set of arithmetic sentences? Even if there is such a set, why is the mechanist committed to the idea that humans are capable of producing such a set and that it is stable under Gödelian reflection? Why should the mechanist believe that if we idealize on things like limitations on life span and memory, we end up with procedures which enumerate a maximally-stable-backed set of arithmetic sentences? Again, we have to be clear about what the game is before we can decide whether Lucas can outplay any given machine.

§**6. Whither (anti-)mechanism?** Under the Euclidean assumption, it was natural to saddle the mechanist with the view that **K** is recursively enumerable. According to mechanism, if we idealize on ordinary theorem-proving activity, ignoring attention-span and simple errors, we end up with something like a Turing machine—and Lucas and Penrose are off and running. However, once we leave the Euclidean model, even the ideal agents change their minds from time to time and so the model of a Turing machine printing out truth after truth is not appropriate.

Let us try a different epistemic principle. Suppose that whenever a human asserts a contradiction, or some other arithmetic falsehood, she has the ability in principle to realize the error and withdraw it. This assumption is a minimal retreat from the Euclidean model (and nowhere near the Quinean seamless web). Call it the *semi-Euclidean* assumption. Lakatos's ([16], [17]) falsificationist picture of mathematical knowledge might be semi-Euclidean. Although Penrose is closer to the traditional, rationalist framework, he acknowledges errors on the part of ideal human mathematicians, but adds that the errors are correctable in principle. If we put aside his procedure for obtaining unassailable knowledge, we have the semi-Euclidean model.

Let M be a Turing machine with two tapes: an output tape and a scratch tape. It can write and erase on both tapes. Say that M *projects* the number n if there is a number t such that after t steps, M prints the numeral for n on its output tape and M does not erase that numeral afterwards. Let X_e be the set of numbers projected by the Turing machine with Gödel number e. A set is *Turing projectable* if there is a Turing machine that projects it. A set is Turing projectable if and only if it is recursively enumerable relative to the halting problem (i.e., Σ_2, see McCarthy and Shapiro [20]).

If the semi-Euclidean assumption is correct, then Turing projectability would be a better framework than recursive enumerability for modeling human mathematical knowability. Let M be a Turing machine. When M prints a numeral on its output tape, pretend that it has asserted the sentence with that Gödel number, and think of the erasure of a numeral from the output tape as the retraction of the corresponding sentence. So a sentence is "projected" by the machine if at some point it "asserts" the sentence and does not retract it later. So a mechanist who accepts the semi-Euclidean assumption might claim that there is a Turing machine that accurately depicts the arithmetic output of a suitably idealized human, an idealized Lucas for example. The set of sentences projected by this machine would be an analogue of the stably-backed sentences for idealized Lucas.

On the semi-Euclidean assumption, there is no assurance that the presently asserted (i.e., printed and so far unerased) sentences are consistent. Suppose that our Turing machine prints a numeral, or our idealized Lucas makes an arithmetic assertion. There is no effective procedure for determining whether the numeral will ever be erased, or whether the idealized Lucas will ever retract the sentence. However, the semi-Euclidean assumption is that the idealized Lucas will eventually discover any errors among his arithmetic assertions and will make the appropriate erasures. In a sense, our ideal subjects are *eventually infallible*.[9]

The foregoing assumptions give us a thesis sharp enough to apply the limitative theorems, and we find ourselves in pretty much the same situation

[9]Suppose that we weaken the semi-Euclidean assumption a little, and only assume that our ideal subjects will discover and retract any inconsistencies they have made. They might leave other errors in place indefinitely. The weakened assumption still makes our idealized subjects eventually infallible for Π_1-sentences, since any false Π_1-sentence is inconsistent with an elementary theorem. We can have our agent systematically assert every true Π_0-sentence and *every* Π_1-sentence. When one of the Π_1-sentences is contradicted (by one of the asserted Π_0-sentences), the agent retracts it.

as with the original Euclidean model—with a twist or two. Let e be the Gödel number of a Turing machine. Then there is an arithmetic (Σ_2) sentence $\Psi_e(x)$ that corresponds to "$x \in X_e$", or "x is projected by the Turing machine with Gödel number e".

Now suppose our mechanist asserts that e is the Gödel number of a Turing machine that represents our idealized Lucas, so that X_e is the set of Gödel numbers of the sentences that idealized Lucas will assert and never retract. Lucas proceeds as before. He calculates a fixed point for $\neg\Psi_e(x)$. That is, he (effectively!) produces a (Π_2) sentence Φ such that

$$\Phi \equiv \neg\Psi_e(\ulcorner\Phi\urcorner)$$

is provable in ordinary arithmetic. So Φ is true if and only if the given Turing machine does not project Φ.

Suppose that Φ is in X_e. Then $\Psi_e(\ulcorner\Phi\urcorner)$ is true and so Φ is false. So if every member of X_e is true, then Φ is not in X_e and thus $\neg\Psi_e(\ulcorner\Phi\urcorner)$ is true. Thus, if every member of X_e is true then Φ is true (but not in X_e). Lucas understands this argument, so if he is convinced that every sentence in X_e is true, then he asserts Φ.

Lucas can outwit any machine which he knows to project only truths. If an idealized human comes to know (and never retract) the statement that X_e contains only arithmetic truths, then this human also knows (with the same certainty) that X_e does not correspond to the sentences he will assert without retracting. He knows at least one truth that is not in X_e.

If the mechanist is correct in the claim about the Turing machine e matching idealized Lucas, then Φ is true if and only if idealized Lucas never comes to assert Φ without later taking it back. We have that Lucas does assert Φ. So if the mechanist is correct and if Lucas never retracts Φ, then Φ is false. This contradicts the semi-Euclidean assumption that idealized Lucas will eventually discover and withdraw any mathematical errors. So if the semi-Euclidean assumption holds of Lucas, then he will retract Φ (in which case Φ is true). That is, idealized Lucas will come to believe that he does not have adequate backing for the relevant fixed point Φ.

Admittedly, this looks strange, but the resolution is similar to the Gödel-Kreisel-Benacerraf analysis. Lucas knows that *if* every member of X_e is true then Φ is true. He should retract Φ if he comes to realize that he does not have adequate backing for the claim that every member of X_e is true. The "evidence" for the statement that every member of X_e is true consists of (i) the semi-Euclidean assumption about idealized Lucas and (ii) the work of the mechanist supporting the claim that X_e reproduces the relevant output of idealized Lucas. The proper conclusion is that Lucas cannot have "adequate backing" for both of these.

We cannot rule out the *existence* of a Turing machine that projects all and only the sentences that an idealized human (like Lucas) can assert without retracting. The mechanist might muster some empirical evidence for this. There may also be evidence of some sort for the claim that idealized Lucas is semi-Euclidean. But this evidence does not amount to stable-backing needed for mathematical propositions.

We have not given our ideal agent complete license to assert something just because he has *some* reason to think it is true. The semi-Euclidean assumption carries a substantial epistemic burden, even if it falls short of the Euclidean ideal of constant infallibility. Under the semi-Euclidean assumption, if an agent asserts an arithmetic falsehood, then he can eventually discover that it is false. We assume eventual infallibility. The only conclusion we have is that under all of these epistemic and empirical assumptions, if X_e does represent an ideal human's unretracted arithmetic assertions, then that ideal human does not have adequate backing for the statement that every member of X_e is true, and so he does not have adequate backing for its Gödel sentence. This is hardly an argument against mechanism.

At this point, one might argue that even the semi-Euclidean assumption is too rigid, and we should further weaken our description of idealized human arithmetic knowledge-gathering—perhaps in the direction of some contemporary holistic epistemologies, such as Quine's seamless web. Then how do the incompleteness theorems take hold? To initiate a reasonable debate over mechanism, we need a property K of arithmetic sentences (or natural numbers) such that $K(\ulcorner\Phi\urcorner)$ holds if and only if an idealized agent stably holds Φ, and we need assurance that $K(\ulcorner\Phi\urcorner)$ holds only if Φ is true. For lack of a better term, let us say that if $K(\ulcorner\Phi\urcorner)$ holds, then Φ is "knowable", or "knowable by the agent". Then we would need a reasonable *computational* property, analogous to recursive enumerability or projectability, to foist upon the mechanist. The mechanist would be committed to a claim that the set of knowable sentences has this computational property. For example, one might argue that the knowable sentences are Σ_8, or that they are no more complex than fourteen applications of the jump operator to a recursive set.

It will not do to define a fallible notion of idealized assertability (say) and claim that a sentence is knowable just in case it is both ideally assertable and true. There is no reason for our mechanist to hold that the property of being both assertable and true is computationally tractable, since the intractable notion of arithmetic truth is built in to the definition. Since arithmetic truth is not arithmetic, why think that assertable-arithmetic-truth is in any way tractable? To get a debate off the ground, the parties must come up with a reasonable notion of arithmetic assertability and independently argue that only truths are assertable.

If the principals could agree on the relevant properties and notions, then Lucas would probably win. The incompleteness phenomenon is very general. Presumably, the computational notion the mechanist attributes to the extension of K would be definable in arithmetic. As above, let $\Psi(x)$ be any formula with only x free and let Φ be a fixed point for $\Psi(x)$, so that

$$\Phi \equiv \neg\Psi(\ulcorner\Phi\urcorner)$$

is provable in ordinary arithmetic. Suppose that it is knowable by the agent that Ψ holds only of true sentences. That is, for each sentence χ, it is knowable that

$$\Psi(\ulcorner\chi\urcorner) \rightarrow \chi.$$

Then Φ & $\neg\Psi(\ulcorner\Phi\urcorner)$ follows. Moreover, the ideal agent can do the relevant deduction. Thus, Φ is knowable by the agent and it is knowable by the agent that $\ulcorner\Phi\urcorner$ is not in the extension of Ψ. *A fortiori*, it is knowable that $\Psi(x)$ does not capture the extension of knowability. In sum, there is no formula in the language of arithmetic which can be known to hold of all and only the knowable sentences (see Shapiro [26, §3]).

To repeat the now familiar conclusions, the extension of knowability might be arithmetic (for all we know so far), but there is no arithmetic formula which can be known to describe the knowable sentences. Any formula whose extension is known to contain only truths effectively leads to a knowable sentence that is not in the extension of the formula. In Dummett's sense, the notion of arithmetic knowability is indefinitely extensible. Any *arithmetic* attempt to capture the notion leads outside the notion—or else fails by calling a false sentence knowable.[10] However, this is not an argument against mechanism. The indefinite extensibility is due to the "truth" component of "knowability" and not to the "human" component.

In another context, Kreisel [14, p. 320] proposes that "human computations are more 'complicated' or, better, more abstract than the objects on which they operate—our thoughts may be more complicated than the objects thought about." He suggests that there is "an asymmetry between the 'simple' concept of natural number and the 'complicated' concept of *proof about natural numbers* . . . " (p. 325). Kreisel suggests that the human ability to reason with abstract objects might take us beyond any machine. Gödel agrees. From the present perspective, however, the further we get from the Euclidean assumption, the more the "complication" in the notion of "proof about natural numbers" lies in the included notion of "truth about natural numbers". We know how complex this last notion is, and this complexity is irrelevant to the issues of mechanism.

To sum up this long journey, we are having trouble coming up with a reasonable mechanistic thesis for Lucas and Penrose to attack. For all we know

[10]Penrose [22, Part II] predicts that an extended and futuristic version of quantum physics will solve the problem of consciousness and, presumably, overcome the Gödelian inadequacies with computational models. If this future science yields a rigorous definition of arithmetic knowability, as Penrose envisions, the science will have to go beyond arithmetic. If the science yields a rigorous definition of set-theoretic knowability, then it will have to go beyond set theory.

so far, the mechanist is free to hold that a human is a Turing machine—whatever that is supposed to mean—or she can claim that the arithmetic productions of a human (idealized or otherwise) are recursively enumerable. She can even claim that she knows an index of a Turing machine that enumerates all and only the arithmetic sentences that a human like herself or idealized Lucas can justifiably assert. But she cannot consistently claim this and claim to know that the human is infallible. Well, maybe the Euclidean is wrong and humans are fallible, even when idealized for lifetime, memory errors, etc. The mechanist is free to come up with some sophisticated procedure whereby humans weed out errors and inconsistencies. She is free to come up with an arithmetic description of the output of our idealized procedures for arithmetic. But she cannot consistently claim that these procedures are exhaustive of human competence and that they are infallible—that every false sentence (or every inconsistency) will eventually be weeded out. Given current epistemologies, it is hard to see how the mechanist is limited by these results.

REFERENCES

[1] P. Benacerraf, *God, the devil, and Gödel*, ***The Monist***, vol. 51 (1967), pp. 9–32.

[2] G. Boolos, *Introductory note* to [11], in [12], (1995), pp. 290–304.

[3] D. Boyer, *J. R. Lucas, Kurt Gödel, and Fred Astaire*, ***Philosophical Quarterly***, vol. 33 (1983), pp. 147–159.

[4] D. J. Chalmers, *Minds, machines, and mathematics*, ***Psyche***, vol. 2 (1995), no. 9, `http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html`.

[5] M. Detlefsen, ***Hilbert's program***, D. Reidel Publishing Company, Dordrecht, 1986.

[6] M. Dummett, *The philosophical significance of Gödel's theorem*, ***Ratio***, vol. 5 (1963), pp. 140–155.

[7] ———, *Reply to Wright*, ***The philosophy of Michael Dummett*** (B. McGuinness and G. Oliveri, editors), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 329–338.

[8] S. Feferman, *Arithmetization of mathematics in a general setting*, ***Fundamenta Mathematicae***, vol. 49 (1960), pp. 35–92.

[9] ———, *Transfinite recursive progressions of axiomatic theories*, ***Journal of Symbolic Logic***, vol. 27 (1962), pp. 259–316.

[10] ———, *Turing in the land of $O(z)$*, ***The universal Turing machine*** (R. Herken, editor), Oxford University Press, New York, 1988, pp. 113–147.

[11] K. Gödel, *Some basic theorems on the foundations of mathematics and their implications*, in [12], (1951), pp. 304–323.

[12] ———, ***Collected works III***, Oxford University Press, Oxford, 1995.

[13] D. Hofstadter, ***Gödel, Escher, Bach***, Basic Books, New York, 1979.

[14] G. Kreisel, *Which number theoretic problems can be solved in recursive progressions on Π^1_1 paths through O?*, ***Journal of Symbolic Logic***, vol. 37 (1972), pp. 311–334.

[15] S. Kripke, ***Wittgenstein on rules and private language***, Harvard University Press, Cambridge, Massachusetts, 1982.

[16] I. Lakatos, ***Proofs and refutations*** (J. Worrall and E. Zahar, editors), Cambridge University Press, Cambridge, 1976.

[17] ———, ***Mathematics, science and epistemology*** (J. Worrall and G. Currie, editors), Cambridge University Press, Cambridge, 1978.

[18] J. R. Lucas, *Minds, machines, and Gödel*, ***Philosophy***, vol. 36 (1961), pp. 112–137.

[19] ———, *Minds, machines, and Gödel: A retrospect*, ***Machines and thought: The legacy of Alan Turing, Volume 1*** (P. J. R. Millican and A. Clark, editors), Oxford University Press, Oxford, 1996.

[20] T. McCarthy and S. Shapiro, *Turing projectibility*, ***Notre Dame Journal of Formal Logic***, vol. 28 (1987), pp. 520–535.

[21] R. Penrose, ***The emperor's new mind: Concerning computers, minds, and the laws of physics***, Oxford University Press, Oxford, 1989.

[22] ———, ***Shadows of the mind: A search for the missing science of consciousness***, Oxford University Press, Oxford, 1994.

[23] ———, *Beyond the doubting of a shadow: A reply to commentaries on 'Shadows of the mind'*, ***Psyche***, vol. 2 (1996), no. 23, `http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html`.

[24] H. Putnam, *Minds and machines*, ***Dimensions of mind: A symposium*** (Sidney Hood, editor), New York University Press, New York, 1960, pp. 138–164.

[25] H. Rogers, ***Theory of recursive functions and effective computability***, McGraw-Hill, New York, 1967.

[26] S. Shapiro, *Epistemic and intuitionistic arithmetic*, ***Intensional mathematics*** (S. Shapiro, editor), North-Holland Publishing Company, Amsterdam, 1985, pp. 11–46.

[27] R. Smullyan, ***Gödel's incompleteness theorems***, Oxford University Press, Oxford, 1992.

[28] N. Tennant, ***Anti-realism and logic***, Oxford University Press, Oxford, 1987.

[29] ———, ***The taming of the true***, Oxford University Press, Oxford, 1997.

[30] A. Turing, *Systems of logic based on ordinals*, ***Proceedings of the London Mathematical Society***, vol. 45 (1939), pp. 161–228.

[31] ———, *Computing machinery and intelligence*, ***Mind***, vol. 59 (1950), pp. 433–460.

[32] H. Wang, ***From mathematics to philosophy***, Routledge and Kegan Paul, London, 1974.

[33] J. Webb, ***Mechanism, mentalism and metamathematics: An essay on finitism***, D. Reidel, Dordrecht, Holland, 1980.

DEPARTMENT OF PHILOSOPHY
OHIO STATE UNIVERSITY AT NEWARK
NEWARK, OHIO 43055, USA
E-mail: shapiro+@osu.edu