*Philosophical Perspectives, 15, Metaphysics, 2001*

# THE LIMITS OF HUMAN MATHEMATICS*

Nathan Salmon
University of California, Santa Barbara

## I

What, if anything, do Gödel's incompleteness theorems tell us about the human intellect? Do they inform us, for example, about human insight and creativity? Or perhaps about the human mind's capacity for *a priori* certainty? Ernest Nagel and James R. Newman write:

> Gödel's conclusions bear on the question whether a calculating machine can be constructed that would match the human brain in mathematical intelligence. ...as Gödel showed in his [first] incompleteness theorem, there are innumerable problems in elementary number theory that fall outside the scope of a fixed axiomatic method... The human brain...appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines. ... Gödel's proof [of the first incompleteness theorem]...does mean that the resources of the human intellect have not been, and cannot be fully formalized, and that new principles of demonstration forever await invention and discovery. ... The theorem does indicate that the structure and power of the human mind are far more complex and subtle than any nonliving machine yet envisaged.[1]

More recently, Roger Penrose has declared that "from consideration of Gödel's theorem...we can see that the role of consciousness is non-algorithmic when forming *mathematical* judgments, where calculation and rigorous proof constitute such an important factor."[2] J. R. Lucas provided an argument in support of a similar (if slightly stronger) conclusion:

> Gödel's [first incompleteness] theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true—i.e., the formula is unprovable-in-the-system—but which we

> can see to be true. It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines.
>
> ...The conclusions it is possible for the machine to produce as being true will...correspond to the theorems that can be proved in the corresponding formal system. We now construct a Gödelian formula in this formal system. The formula cannot be *proved-in-the-system.* Therefore the machine cannot produce the corresponding formula as being true. But *we* can see that the Gödelian formula is true: any rational being could follow Gödel's argument, and convince himself that the Gödelian formula, although unprovable-in-the-given-system, was nonetheless—in fact, for that very reason—true. Now any mechanical model of the mind must include a mechanism which can enunciate truths of arithmetic, because this is something which minds can do... But...for every machine there is a truth which it cannot produce as being true, but which a mind can. This shows that the machine cannot be a complete and adequate model of the mind. It cannot to *everything* that a mind can do, since however much it can do, there is always something which it cannot do, and a mind can. ... The Gödelian formula is the Achilles' heel of the cybernetical machine. And therefore we cannot hope ever to produce a machine that will be able to do all that a mind can do: we can never, not even in principle, have a mechanical model of the mind.[3]

Anticipating this argument, Hilary Putnam exposed an apparently fatal fallacy.[4] We are to suppose, for a *reductio ad absurdum*, that we have been given in full detail a complex logistic ("formal") system that adequately and completely formalizes the mathematical abilities of a human mind. It is by no means a foregone conclusion that the mind can prove the proposition expressed by the Gödelian sentence for this system—a sentence that indirectly says of itself (in a well-defined sense) that it is not provable-in-the-given-logistic-system. What is proved is conditional: that the proposition is true *provided the logistic system is consistent.* Indeed, this much is provable within the very logistic system in question. Proving that the system is consistent (free of contradiction) would yield the target proposition as an immediate corollary. Gödel's second incompleteness theorem states that the logistic system, if it is consistent, cannot in this sense prove its own consistency. (The second theorem itself is proved precisely by noting the corollary that would otherwise result.) For some relatively simple logistic systems of arithmetic, we may know with mathematical certainty, even though this is not provable within the system, that its primitive deductive basis (the axioms and primitive rules of inference) does not generate any contradiction. In these cases, there may be a sense in which it is true that the human mind relevantly "sees" the truth expressed by the Gödelian sentence, since this provably follows from the system's consistency. But there are other logistic systems for mathematics with respect to which the system's consistency is anything but obvious. In particular, the second incompleteness theorem calls into serious question whether the human mind is capable of a proof of consistency for a logistic system sufficiently complex to capture all of humanly demonstrable mathematics, i.e. a logistic system adequate to formalize the human capacity for proving mathematical theorems.[5]

Perhaps a more guarded conclusion can be legitimately drawn. In his 1951 Josiah Willard Gibbs Lecture to the American Mathematical Society, Gödel himself derives from his second incompleteness theorem a disjunctive conclusion which, though weaker than the conclusions of Newman and Nagel, et. al., Gödel says is a "mathematically established fact which seems to me of great philosophical interest":

> Either mathematics is incompletable in this 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 (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives).[6]

This disjunction is evidently not subject to the same response that Putnam made to Nagel and Newman and company. For Gödel judges only that the human mind surpasses any theorem-proving machine *provided that the mind is in principle capable of solving any purely mathematical problem, including the question of its own mathematical consistency*. This more cautious conclusion is nevertheless philosophically substantive. Gödel proceeds to draw disjunctive philosophical conclusions from it, by inferring consequences of the first disjunct about the human mind's capacity for outperforming any finite computing machine, including whatever theorem-proving machinery there is in the human brain, and consequences of the second disjunct about the independence and objectivity of pure mathematics. If the theorem-proving machinery of the human brain is a computer, then either the human mind surpasses the human brain or humankind does not deserve credit for creating pure mathematics (or as some might see it, humankind does not deserve the blame). Thus, the human mind either surpasses the very organ in which it evidently resides or else it is not responsible for the existence of pure mathematics—or both, as Gödel himself believed (and I agree).[7] Here follows the relevant passage in which Gödel derives the disjunction:

> 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. [*Gödel's note*: If he only says "I believe I shall be able to perceive one after the other to be true" (where their number is supposed to be infinite), he does not contradict himself. (See below.)] 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; it does not, however, if one understands by it the system of all demonstrable mathematical propositions. I shall distinguish these two meanings of mathematics as mathematics in the objective and in the subjective sense: Evidently no well-defined system of correct axioms can comprise all [of] objective mathematics, since the proposition which states the consistency of the system is true, but not demonstrable in the system. However, as to subjective mathematics, 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; [*Gödel's note*: For this (or the consequence concerning the consistency of the axioms) would constitute a mathematical insight not derivable from the axioms and rules under consideration, contrary to the assumption] or in other terms, we could perceive to be true only one proposition after the other, for any finite number of them. The assertion, however, that they are all true could at most be known with empirical certainty, on the basis of a sufficient number of instances or by other inductive inferences. ... If it were so, this would mean that the human mind (in the realm of pure mathematics) *is* equivalent to a finite machine that, however, is unable to understand completely its own functioning. [*Gödel's note*: Of course, 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.][8]

There appears to be the following sort of argument: Suppose that the human mind's capacity for conceiving proofs is an effectively describable phenomenon, like the deterministic workings of a Turing machine, so that the very process by means of which the mind attains, or can attain, purely mathematical knowledge with mathematical certainty is thus fully captured by some finite effective rule (even if it is very long). It is a consequence of the second incompleteness theorem that the mind cannot know with mathematical certainty that this rule generates only correct results, or even that its results are internally consistent. For if the mind can know with mathematical certainty of all the propositions of pure mathematics it is able to prove that all of them are true, then it can also know with mathematical certainty that they are formally consistent—something that is precluded by the theorem. Since the consistency of the system of theorems can be recast as a purely mathematical proposition, it follows that if the mind, in its theorem-proving capacity, is a finite machine, then there are purely mathematical truths it cannot know with mathematical certainty; in particular, it cannot prove its own consistency, and hence cannot completely understand its own functioning.

George Boolos has claimed that Gödel's disjunction—that either the human mind is not equivalent to a finite machine or there exist absolutely undecidable mathematical propositions—though it is weaker than the conclusions of Nagel and Newman, et. al., is still not validly derivable from the incompleteness theorems.[9] Boolos deems the above argument inconclusive owing to obscurity in the idea that "the human mind is equivalent to a finite machine."

Even assuming, for the sake of argument, that the theorem-proving aspect of the human mind is mechanistic, it does not straightforwardly follow that in that case the mind's theorem-proving mechanism meets the conditions for being a Turing machine and is therefore incapable of proving its own consistency. For it is in the first place excessively unclear what is meant by saying (or by denying) that *the* human mind, or even that a single mind, simply *is* a Turing machine. And if what is meant is that the theorem-proving aspect of the mind, or of a single mind, is (or is not) *represented* by a Turing machine, Boolos objects, Gödel does not specify exactly how the representation is supposed to go.

The argument does indeed raise troubling questions of this sort, and more. A Turing machine is the formal counterpart of a deterministic computational process. It does much more than merely represent a recursive function in the abstract, mathematical sense. The function is fully represented by the machine's input and output, and may be aptly represented equivalently by a set of ordered sets of numbers. By contrast, a Turing machine is the program that produces a specific output for a given input; it represents the process of *calculating* the value of the function for any argument. In the opening paragraph of the Gibbs Lecture, just before arguing for his disjunction, Gödel cites Turing machines as providing the most satisfactory way of defining the concept of an effective calculation or algorithm (a "finite procedure")—thereby indicating his acceptance of Church's thesis (at least as restricted to numerical functions, and sets characterizable by numerical functions). Is the "finite machine" of which Gödel speaks in the quoted passage supposed to mirror, in the manner of a Turing machine, the method and procedures by which the human mind is able to construct or discover (as the case may be) mathematical proofs? If so, we need to know exactly how, and exactly to what extent, the finite machine does this in order to assess Gödel's conclusion. Lacking this additional information, the most that can be justified is the supposition that the machine delivers the same theorems that the mind is able to prove, though perhaps by a completely different construction.

Filling in the gaps, Boolos proposes a reconstruction of Gödel's argument culminating in a circumscribed conclusion concerning not the actual process of proving theorems, but just the results thereby obtained. Though still somewhat vague, Boolos grants that the following is a consequence of the second incompleteness theorem: If there is a theorem-proving Turing machine whose output is the set of sentences expressing just those mathematical propositions that can be proved by a mind capable of understanding all polynomials with integer coefficients (and therefore capable of understanding a mathematical sentence tantamount to the meta-theoretic observation that the mind's theorem-proving mechanism is consistent), then there is a true mathematical proposition that can be understood but cannot be proved by that same mind—namely, the mathematically recast assertion of its own consistency. (See note 10 below.) Thus, any mind whose theorem-proving capacity is representable by some Turing machine *in terms of the theorems it proves (as opposed to the proofs it produces and/or the process by which it conceives those proofs)* is in principle incapable

of solving certain mathematical problems indirectly about its own theorem-proving capacity. On Boolos's reconstruction, the machine passively represents the mind's potential output of theorems. Boolos's conclusion concerns those theorems only in the sense that it is about that *class* of theorems, not their production. The machine does not necessarily represent the mind's potential proofs of that potential output, let alone the active process by which the mind can generate those proofs.

Boolos's conclusion is comparatively strikingly narrow. It is a trivial, disappointingly anti-climactic restatement of the second incompleteness theorem's corollary that no theorem-enumerating Turing machine prints a sentence tantamount to an assertion of its own consistency. Any possible generating activity whose output coincides, for whatever reason (or for no reason at all), with that of a theorem-enumerating Turing machine fails to produce a mathematical proposition tantamount to the consistency of that output—regardless of whether the activity is teleologically assisted by an understanding of the output, hence even if it is a room full of monkeys at typewriters.[10] One might also point out, in much the same spirit, that anyone whose feats in manipulating geometric figures, as it happens, do not exceed those geometric tasks that can be performed using only a compass and straightedge, does not trisect an angle. In confining his attention to the mathematical theorems themselves, setting aside the epistemological character of their potential proofs by the human mind, Boolos disengages his conclusion from the philosophical issues that drive Gödel's. Gödel's argument does not concern hypothetical minds of a precisely delimited capacity. It concerns the capability of the human mind, such as it is, to attain certainty in mathematics. It is about human mathematics at its edges—both the initial starting points and the ultimate upward limits. Does the obscurity of the very idea that the human mind is equivalent to a machine block us from any such sweeping conclusion, and force a disappointingly restrictive retreat? I believe it does not and that, *contra* Boolos, Gödel's argument about the limits of human mathematics is reasonably secure, or can be made so.

## II

Gödel's principal argument does not make any essential detour through Turing machines, or machines of any sort. One can dispense with machines altogether and make an end run for a disjunctive conclusion of just the sort from which Gödel draws philosophical conclusions about the human mind and the objectivity of mathematics.

Following Gödel, let us distinguish between mathematics proper (i.e., all the truths of pure mathematics) and what I have called *human mathematics* (Gödel's "subjective mathematics")—that portion of mathematics that the human mind, or any intelligence (whether biological or artificial) that is epistemologically similarly situated to human intelligence, is capable of knowing with mathematical certainty ("mathematical certitude"). It is useful for this purpose

to introduce some artificial terminology. Let '**HuMath**' designate the class ("system") of all true propositions of human mathematics. This is a subclass of the class **Math** of all purely mathematical truths. **HuMath** almost certainly extends well beyond all the mathematics that will ever have been known with mathematical certainty by humans—by some human or other at some time or other. Take note: it is not assumed that **Math** and **HuMath** are distinct, nor is it assumed that they are identical. It is not even assumed that **HuMath** includes every purely mathematical proposition that mathematicians take to be true. **HuMath** is restricted to those purely mathematical propositions that are knowable, hence true. If (contrary to our expectation) there should be any false purely mathematical propositions of which mathematicians have been persuaded (e.g., by a subtly fallacious argument), they are excluded from **HuMath**. Since all of **HuMath** are true, **HuMath** is *a fortiori* consistent, i.e. no contradiction is correctly deducible from it. Notice also that **HuMath** excludes any purely mathematical truths that are only knowable by the human mind to some degree short of mathematical certainty.[11]

**HuMath**'s definition invokes the generic notion of knowability by the human mind, and this notion is somewhat obscure. What is knowable by one human mind may be unknowable by another. It may be that no single, existent human mind (past, present, or future) is capable of knowing everything that the human mind is capable of knowing. It may even be that no *possible* human mind can know all of the facts each of which, taken individually, the human mind is capable of knowing.[12] As Boolos notes, it does not follow that no proposition involving the notion of human knowability is validly deducible from a mathematical theorem. Boolos cites the particular inference: *91 is composite; therefore, it is not humanly knowable that 91 is prime*. This instance depends on the fact that knowledge entails truth. Gödel's derivation of his disjunction, by contrast, depends on the fact that knowledge entails epistemic *justification*. But this does not, in itself, provide a reason to doubt that Gödel's argument is sound. The basic epistemological assumption is that, whatever differences there are among humans, certain epistemic mechanisms—ways of coming to know—are in principle accessible to the human mind.[13] At a minimum, there is an epistemic mechanism that is characteristically human, in this sense, and yields mathematical knowledge with mathematical certainty. The principle does not require that one be able to determine with any certainty whether a particular alleged phenomenon (e.g., telepathy) is a human epistemic mechanism, in this sense, or whether a particular alleged fact is knowable by a human mechanism. It may well be that this fundamental epistemological principle is not itself known with mathematical certainty, and to the extent that Gödel's argument presupposes the principle, the derived disjunction is also not so known. But the principle is known (even if not with mathematical certainty), and is not typically subject to doubt. If a proposition is validly inferred from a mathematical theorem using this epistemological principle, it is not unreasonable to say that the inferred proposition is a mathematically established fact.

The epistemic mechanism by which the elements of **HuMath** are knowable with mathematical certainty by humans is evidently that of mathematical *proof*. Gödel notes that if any purely mathematical knowledge is obtained by proof on the basis of truths antecedently known with mathematical certainty, then some purely mathematical knowledge is not.[14] For proofs must have starting points, and knowledge obtained by proof is derived ultimately from knowledge of those starting points. The latter knowledge Gödel calls the "evident axioms." (It includes axioms of both logic and mathematics proper.) This epistemic mechanism for attaining certainty in pure mathematics is aptly represented by the logistic method.[15] There is a proper subclass **Ax** of **HuMath** consisting of epistemologically foundational axioms—purely mathematical "first truths" each knowable with mathematical certainty by the human mind (i.e., by some possible human mind) without proof from other purely mathematical truths but through direct mathematical intuition or insight ("perception"), or perhaps derived from something more fundamental than pure mathematics (including logic)—while the rest of **HuMath** are knowable with mathematical certainty only by proof, i.e. only by deductive derivation ultimately from the mathematical axioms, using logical (primitive) rules of inference together perhaps with purely mathematical rules of inference over and above the axioms. **HuMath** is the deductive closure of **Ax** under the rules of human mathematical reasoning. In this sense, the union of **Ax** with the rules of human mathematical inference form the deductive basis of human mathematics. Let us call it '**Basis**'.[16]

**Ax** may extend beyond all those fundamental truths of pure mathematics that will ever have been known by humans with mathematical certainty without independent mathematical proof, i.e. without proof from antecedently known purely mathematical truths. It is not assumed that any particular human mathematician, or even any possible human mathematician, can know all the elements of **Ax**. However, each of the axioms, taken individually, must be humanly knowable with mathematical certainty without independent mathematical proof. If we cannot know an axiom, then we also cannot know anything derived from it—except by some independent epistemological means. Genuinely inferential knowledge requires knowledge of that from which it is inferred. Moreover, each of the rules of inference of human mathematical reasoning must be not only valid (i.e., such as to preserve truth in any model), but also of a sort that transfers, through the cognitive act of immediate inference, the sort of epistemic justification that yields mathematical certainty. It is not independently required that we know each of the inference rules to be valid (let alone that we know this with mathematical certainty), but knowing this may be inextricably bound up with the rules' being such as to transfer mathematical certainty to the immediately inferred conclusion from that from which the conclusion is immediately inferred. In any event, it is reasonable to suppose that we can know of each inference rule of the required sort, with mathematical certainty and without independent mathematical proof, that it is indeed valid.

It is frequently assumed in discussion of Gödel's incompleteness results (especially of their philosophical implications) that they entail that any well-defined

deductive basis for arithmetic, if consistent, is incomplete and fails to decide in particular a recast assertion of its own consistency. From this it would follow directly that, contrary to David Hilbert, there are purely mathematical truths the human mind is incapable of proving, including an assertion of its own mathematical consistency. (Recall that **Ax** is a subclass of **HuMath**, which is restricted to truths, and that the rules are valid; hence **Ax** is consistent.) But the assumption often involves a mistake, and Gödel did not believe its conclusion. There exist deductive systems for arithmetic (in a broad sense of 'deductive system') that are both consistent and complete—Gödel's theorems notwithstanding. This simple fact, although sometimes overlooked, is essential to a proper understanding of Gödel's disjunction and the argument for it. One way to obtain a consistent deductive system for arithmetic whose theorems are exactly those sentences of the language that express truths of arithmetic is to take all and only those sentences as axioms.[17] No object-theoretic Gödelian sentence indirectly asserting its own unprovability-in-this-system exists. On the other hand, the axiom set is unwieldy—as unwieldy as possible without allowing for the deduction of falsehoods. It is all over the map. Each expressible truth of arithmetic, regardless of how complex or abstract, is provable in this system in a single line. We are currently in no position to determine whether certain sentences are axioms of this system—for example, the sentence expressing Goldbach's Conjecture. By contrast, the elements of **Ax** are narrowly confined to those purely mathematical truths that are humanly knowable with mathematical certainty without independent mathematical proof. The envisaged complete, consistent system does not come close to adequately representing the way the human mind achieves knowledge with mathematical certainty in arithmetic. Part of the significance of Gödel's incompleteness results derives from the fact that they obtain for deductive systems that do at least approach the way the human mind attains mathematical knowledge.

A requirement that the axioms be written out in full would be excessive, since it excludes the possibility of a logistic system with infinitely many axioms. Instead, it is customary to consider deductive systems whose primitive bases are recursively enumerable (if not indeed primitive recursive)—so that even if there are infinitely many axioms there is an effective procedure by which theoretically one could enumerate them (allowing repetitions) and calculate what the *n*th axiom is for any natural number *n*. This condition (or something that entails it, perhaps given Church's thesis) is typically built into the definition of a *logistic* or *formal* system or theory.[18] It is only in that case that the deductive system can be effectively specified (in an intuitive sense) in a finite description. Moreover, if the deductive basis is effectively decidable, then so is the notion of a proof. Suppose that the elements of **Ax** constitute a *recursively enumerable set of propositions*, in the following sense: that there is a recursive numerical function from whole numbers onto a set *A* of Gödel numbers of sentences of a possible formal language expressing each of the elements of **Ax** in that possible language—so that there is an effective procedure by which theoretically one could calculate what the *n*th element of **Ax** is for any natural num-

ber $n$.[19] Suppose also that the rules of inference are analogously recursively enumerable. (See note 16.) Gödel showed how, in that case, the notions of a proof-from-**Ax** and of contradiction, and therewith the statement of **HuMath**'s consistency (which is meta-theoretic), can be put into object-theoretic form. Specifically, if the elements of **Ax** form a recursively enumerable set, and so do the inference rules, then there is a purely mathematical binary relation *Proof* which is designated by an open formula $\phi_{Proof}(x, y)$ of a possible formal language suitable for arithmetic and which provably holds between a pair of numbers $n$ and $m$ if and only if $n$ is the Gödel number of a sequence of formulae that collectively express, in that same formal language, a proof from **Ax**, by way of the inference rules, of the proposition expressed, in that language, by the formula whose Gödel number is $m$. Likewise, there is a purely mathematical relation *Contradict*, designated by an open formula $\phi_{Contradict}(x, y)$ of the same language, which provably holds between a pair of numbers if and only if they are the Gödel numbers of formulae one of which is the negation of the other. There is then a corresponding sentence $\varphi_{Cons}$ of the form $\ulcorner\sim(\exists x)(\exists y)[\phi_{Contradict}(x, y) \wedge (\exists z)\phi_{Proof}(z, x) \wedge (\exists z)\phi_{Proof}(z, y)]\urcorner$, which is mathematical code *via* Gödel numbering for the consistency of the logistic system generated by the set $A$ of axioms and the inference rules. The sentence $\phi_{Cons}$ expresses a mathematical proposition *Cons* which we know with mathematical certainty to be equivalent to the logistic system's formal consistency.[20] On the assumption that the elements of **Ax** and the rules constitute recursively enumerable sets, Gödel's second incompleteness theorem implies that $\varphi_{Cons}$ is not provable from **Ax**. For the theorem (as extended by Barkley Rosser) states that if an axiomatic basis suitable for arithmetic is both recursively enumerable and consistent, then the corresponding object-theoretic statement (constructed thus *via* Gödel numbering) of the theory's consistency, though true, is not provable from those axioms.[21] Since each of the propositions expressed by the elements of $A$ is knowable, *a fortiori* each is true. And since all of the them are true and the rules are valid, $A$ is *a fortiori* consistent. Thus, if the elements of **Ax** constitute a recursively enumerable set, and so do the rules, then *Cons* is a purely mathematical truth that does not belong to **HuMath**.

In this sense, either the axiomatic basis of human mathematics (i.e., the purely mathematical truths knowable by the human mind with mathematical certainty without independent mathematical proof, together with the rules of human mathematical inference) is not reducible to a recursively enumerable set (and thus they do not yield a logistic or formal system, in the technical sense), or else some purely mathematical truths—including a mathematical encoding of the consistency of human mathematics—are in principle unknowable by the human mind with mathematical certainty. This result already goes significantly beyond Boolos's conclusion that any mind capable of understanding all polynomials with integer coefficients and whose provable theorems exactly coincide with the output of a theorem-proving Turing machine is incapable of proving a mathematical truth that it apprehends. But Gödel takes matters further still.

Enter the argument about a "finite rule" and the prospect of the human mind being "equivalent to a finite machine." Against the interpretation placed on this by Boolos and others, the imagined rule, as it is understood and intended by Gödel, does not generate proofs of the elements of **HuMath**—let alone does it capture the method or procedure by which the mind constructs or discovers proofs.[22] Whereas Gödel's argument is concerned with the epistemological character of potential proofs by the human mind, the actual cognitive process whereby the human mind might conceive or discover its proofs is irrelevant. Let it be by a mechanistic process or let it be utterly non-mechanistic, by an indescribable mathematical inspiration, by a vital, non-deterministic spark of creativity. Let it be by supernatural revelation, or by divine intervention. It makes no difference to the argument.

Nor is the envisaged "finite rule" merely supposed to produce the mathematical theorems provable by the human mind—the elements of **HuMath**—even if by a potentially different construction. What the speculated rule *is* supposed to generate are the "evident axioms," i.e., not the elements of **HuMath** themselves but their axiomatization in **Basis**. If **Basis** is recursively enumerable, there is an effective procedure that enables one to enumerate its elements (possibly with repetitions). According to Church's thesis (construed so as to include the effective enumeration of a set of propositions), the converse obtains as well. Suppose there is a finite rule that produces all the elements of **Basis**—for example, finite instructions enabling one automatically to write out the sentences of a possible mathematical language, one after another, which express just the elements of **Ax** as well as the inference rules. Mathematical certainty that the rule, properly characterized, generates no inconsistencies would then be unattainable. It follows from the second incompleteness theorem (and Church's thesis) that if there are such instructions, then even though each of the propositions expressed by the sentences they produce is true and humanly knowable with mathematical certainty without proof, and even though each of the generated rules are valid and such as to transfer mathematical certainty *via* the immediate inference, we cannot know of the instructions, with the same certainty, that their product is even consistent. Therefore, either there is no such rule—equivalently, no recursive function that enumerates the elements of **Basis**—or again there are purely mathematical truths of a certain type that are humanly unprovable. This is, nearly enough, Gödel's disjunction. It is, in effect, a trivial transformation in propositional logic of the following: *If the elements of **Basis** constitute a recursively enumerable set, then **HuMath** is a proper subclass of **Math**.*[23]

Gödel expands on his first disjunct—that there is no effective procedure producing exactly the axiomatic basis of human mathematics—by drawing an inference concerning the human mind *vis a vis* a finite machine. If indeed there is no such rule, then the human mind's capacity for attaining certainty in mathematics surpasses that of a theorem-proving computer—at least insofar as the computer's theorem-proving capacity is restricted to procedures that correspond to a recursive notion of proof. There is no assertion here that the theorem-

proving mechanism of the human brain is not a computing machine (if 'machine' is the right term to use) whose theorem-proving capacity is not restricted in this way. Boolos's worries about the vagueness of the general notion that "the human mind is equivalent to a finite machine," while they may be an appropriate reaction to an attempt to derive some such more sweeping conclusion than this, are not pertinent here. The difficulty of likening the theorem-proving capacity of the human brain to a computer is not so much that the brain's cognitive processes are not mechanistic. Nor is it that a machine cannot know the fundamental axioms of human mathematics. (Although it cannot. Strictly speaking, it is a person, and not the person's brain, that knows things.) The difficulty comes in the very *design* (let alone the construction) of a theorem-proving machine when there is no effective procedure for delimiting its proofs' admissible starting points. Either there is no such procedure with regard to the human mind's capacity for attaining knowledge with mathematical certainty in pure mathematics, or else there are purely mathematical problems of a certain sort that are in principle unsolvable by the human intellect. This is Gödel's disjunction.

## III

Gödel remarks in passing (in effect) that the correctness of a set of propositions (i.e., truth of all the elements) entails their formal consistency, and hence knowledge with mathematical certainty of the former yields knowledge with mathematical certainty of the latter. Call this 'Gödel's thesis'.[24] It follows that knowledge with mathematical certainty of a proposition $p$ (which may be a conjunction of propositions) yields the knowledge, with the same certainty, that $p$ is consistent. Insofar as **Ax** consists of propositions that the human mind is capable of knowing with mathematical certainty, one might expect the mind to be able to know the conjunction of those axioms (perhaps by repeated applications of a familiar logical rule of inference). From the latter, according to Gödel's thesis, we could deduce the conjunction's consistency, and from this the Gödelian undecidable proposition. Does Gödel's thesis provide support for Lucas's assertion that the mind can after all see the truth of Gödel's undecidable proposition, which indirectly says of itself that it is not provable from the axioms?

Not without further argument. **Ax** is presumably infinite. The conjunction of its elements would then be an infinite conjunction. But there is a question of whether there even exist such propositions. If such propositions do exist, there is still a question of whether the human mind can comprehend them. Furthermore, though each element of **Ax** is knowable with mathematical certainty without independent proof, it does not follow that the conjunction of all the axioms is itself knowable with mathematical certainty—even assuming that this conjunction is humanly comprehendible. In order for a proof to confer knowledge with mathematical certainty, one must know each of the axioms employed in the proof with the same certainty. Even if one is thus capable of knowing with certainty the conjunction of axioms used in any proof that one may construct or discover, since proofs are finite this yields knowledge with certainty of con-

junctions of finite subsets of elements of **Ax**, not yet knowledge with certainty of the conjunction of *all* elements of **Ax**.[25]

Suppose the human mind were able to know the conjunction of all of **Ax** at once. Suppose the inference rules are finite. According to Gödel's thesis, we could then know with mathematical certainty that if the conjunction of such-and-such axioms is correct, then the conjunction of such-and-such axioms (these same ones) is formally consistent. The fact concerning **Ax**—that if all those propositions are correct then they are consistent—is something we would be able to know with mathematical certainty if we were capable of apprehending **Ax** all at once, and if we are capable of any mathematical knowledge at all. Hence, if we could but know the conjunction of all elements of **Ax** with mathematical certainty, we could infer their consistency by *modus ponens* (an inference rule of just the sort required). But if the elements of **Ax** constitute a recursively enumerable set, then we cannot know *Cons* (which is provably equivalent to the consistency of **Ax**) with mathematical certainty. Therefore, by *reductio ad absurdum*, either the elements of **Ax** are not recursively enumerable, or else their conjunction is not humanly provable. Or to put the point somewhat differently from Gödel: Though each of the elements of **Ax**, taken individually, is humanly knowable with mathematical certainty, if those elements are recursively enumerable, then even though they are, their conjunction is not humanly knowable with mathematical certainty. This result does not advance the position of Nagel and Newman, et. al.

By Gödel's thesis, if the elements of **Ax** are recursively enumerable, then the human mind is barred from knowing their conjunction with mathematical certainty. This does not mean that if the elements of **Ax** are recursively enumerable, then the human mind is barred from knowing with mathematical certainty the general meta-theoretic proposition that *all the purely mathematical propositions knowable with mathematical certainty by the human mind without independent mathematical proof are true*. On the contrary, the latter proposition appears to be something of which we are certain (setting aside worries about the so-called paradox of the knower), on the basis of the analytic truth that whatever is known is true. (See note 25.) What it does mean is that if the elements of **Ax** are recursively enumerable, we are barred from knowing *of* those propositions (*de re*) with mathematical certainty that all are true, by inference from anything of the form ⌜Every $x$ such that $\phi(x)$ is true⌝ where ⌜$\phi(x)$⌝ designates **Ax** in a manner provably equivalent to its designation in $\phi_{Proof}$ and $\phi_{Cons}$. In particular, even if the elements of **Ax** are recursively enumerable, we cannot know with mathematical certainty of any recursive function that enumerates it, that it generates only Gödel numbers of true sentences—with the enumerating function characterized so as to yield a formula ⌜$\phi(x)$⌝ of the indicated sort.

Again suppose there is a finite rule that produces exactly the axioms of human mathematics. Under certain circumstances (e.g., where one fully understands the possible language in question), knowing of the envisioned effective instructions that they produce only sentences expressing truths is tantamount to

knowing those propositions expressed to be consistent. It follows that if there are such instructions, then even though each of the propositions expressed by the sentences they produce is true and humanly knowable with mathematical certainty without proof, we cannot know of the instructions, with the same certainty, that their product is correct. If there are effective instructions that produce sentences expressing exactly the axioms of human mathematics, we are incapable of knowing of those instructions with mathematical certainty that they do so. If we were to stumble upon such a rule we could not prove it to be such, or even that it produces only truths.[26]

Lucas, like Nagel and Newman and others who have discussed the philosophical import of Gödel's incompleteness results, evidently tacitly assumes that insofar as the mathematical capabilities of a human mind is represented by a deductive system at all, the axioms constitute a recursively enumerable set, if not indeed a recursive set.[27] It follows from this assumption, taken together with Gödel's thesis, that though each of the axioms is humanly knowable with mathematical certainty, the mind is incapable of deducing their conjunction. This in itself does not refute Lucas's argument. The position of Nagel and Newman, et. al., appears to be that, whatever one's axioms for mathematics may be at a given time, the human mind, unlike the logistic system it instantiates at that time, is capable of augmenting its primitive deductive basis through a non-mechanistic mathematical insight that goes beyond what is strictly provable from those axioms. The mind can both prove that the axioms cannot prove their own consistency, and at the same time *see* (without proving this from the current axioms) that those same axioms *en toto* are correct, hence consistent. The mind thereby expands its deductive basis, empowering itself to prove the incompleteness of the previous axioms from the new set. The mind can then repeat the maneuver with respect to its new deductive basis, and then again with the yet newer basis, and so on in an ongoing dialectic. More important, the vital mathematical faculty or insight that fuels the dialectical progression also yields knowledge with mathematical certainty of its own correctness, and hence consistency, and thereby of the correctness, and consequent consistency, of the entire system generated by the initial axioms and inference rules taken together with the special non-mechanistic faculty itself. The hypothesized vital mathematical insight would strikingly set the human mind apart from any machine or mechanistic process that lacks it.

Unfortunately, this view of human mathematics as a dynamic process of continuing discovery fueled by a unique kind of non-mechanistic and self-validating mathematical insight does not solve the problem. **Ax**, by definition, includes every purely mathematical truth that is humanly knowable without proof from other purely mathematical truths. If there is any special, self-validating faculty or intuition of the sort hypothesized, whatever is humanly knowable by its means is thus already included in **Ax**. The only way for the mind to come to know a purely mathematical truth with mathematical certainty that does not belong to **Ax** is to prove it ultimately from **Ax** (i.e., to prove it from **Ax**, or to prove it from theorems proved from **Ax**, or from theorems proved from theo-

rems proved from **Ax**, etc.). **HuMath** is completely axiomatized by **Basis**, i.e., **Ax** together with the inference rules. In light of Gödel's second theorem, if **Basis** is recursively enumerable, the recast assertion of its consistency is not humanly knowable with mathematical certainty. Rather than making the case for the position of Nagel and Newman, et. al., this result spells trouble for it.

Assume for the moment, with Hilbert, that the human mind is capable, in principle, of solving any purely mathematical problem. It then follows from Gödel's disjunction that the mind's capacity for proving theorems surpasses that of any theorem-proving computer whose primitive deductive basis is recursively enumerable. The mind's superiority over any such machine (in this sense) is explained not so much, or not directly, by the mind's being able to "see" that which cannot yet be proved, but instead by the fact that its primitive deductive basis is essentially richer than the computer's. The richness of human mathematics would in that case result from some human faculty or intuitive insight—which would indeed separate man from those machines without it—but this special mathematical faculty or intuition might be the very same faculty that provides us with even the simplest axioms, not something further and different. Moreover, its consistency may not be reducible to any purely mathematical proposition, and therefore it need not be self-validating to be mathematically complete.[28] In any event, there remains the unproven assumption that the human mind can prove every purely mathematical truth.

## IV

Gödel's derivation pointedly places a special focus on a question that is ignored in Lucas's argument: Are the elements of **Basis** recursively enumerable? Or put another way (under the assumption of Church's thesis, applied to the effective enumeration of a set of propositions): Is there an effective procedure for enumerating the rules of human mathematical inference together with those purely mathematical truths that the human mind is capable of knowing with mathematical certainty without independent mathematical proof? If there is not, then the mathematical capacity of the human mind surpasses that of any mathematics machine whose deductive basis is subject to such a procedure; and otherwise, the human mathematical mind is, in a certain sense, in principle incapable of resolving the question of its own consistency.

In particular, could it be that **Ax** is not effectively enumerable? Many logicians would regard this prospect as quite impossible. Church argued, in effect, that nothing should count as a genuine *proof* unless the totality of axioms form a recursively enumerable set, indeed a recursive set. He posed his argument in the context of a logistic system, construed syntactically. Church imposed as an inviolable restriction on any logistic system that "the specification of the axioms shall be effective in the sense that there is a method by which, whenever a well-formed formula is given, it can always be determined effectively whether or not it is one of the axioms."[29] Unless there is an effective procedure for deciding whether a given formula is or is not one of the axioms,

the notion of proof itself will not be effective. Church's justification for the restriction is given with characteristic eloquence and force:

> There is then no certain means by which, when a sequence of formulae has been put forward as a proof, the auditor may determine whether it is in fact a proof. Therefore he may fairly demand a proof, in any given case, that the sequence of formulae put forward is a proof; and until this supplementary proof is provided, he may refuse to be convinced that the alleged theorem is proved. This supplementary proof ought to be regarded, it seems, as part of the whole proof of the theorem, and the primitive basis of the logistic system ought to be so modified as to provide this, or its equivalent. Indeed it is essential to the idea of a proof that, to any one who admits the presuppositions on which it is based, a proof carries final conviction.[30]

Lecturing on Gödel's incompleteness theorems in 1974, Church gave the following related argument, as reconstructed from my notes (edited and approved by Church at the time for distribution to the class):

> The initial reaction to an incompleteness proof for a logistic system is to search for additional axioms, postulates, or rules of inference which, when added to the incomplete system, yield a complete system. But there does not seem to be any way of doing this for the logistic system $A^2$ [a formalization of second order Peano arithmetic]. The Gödel proof does not make great use of the particular axioms, postulates, and rules of inference of $A^2$. The proof is of such generality that it is easily extended to a logistic system obtained from $A^2$ by the addition of particular axioms, postulates, and rules of inference.
>
> The reason for the incompleteness of $A^2$ does not lie in the axioms, postulates, or rules of inference, but rather in the notion of mathematical proof. A proof must carry conviction; one who accepts the axioms and rules of inference, if he has once seen a proof of a particular theorem, must then not be able justifiably to doubt the theorem. But if axiom schemata or rules of inference are non-effective, the situation can arise that one who has seen a proof may still doubt, because he is unable to verify that what is before him is in fact a proof. Thus the notion of proof must be effective, that is, there must be an effective procedure for determining whether an alleged proof is a proof. Presumably this means that the notion of proof must be general recursive, since there is no known effective check which is not general recursive. Even if we were to add an axiom schema to the logistic system $A^2$, the set of instances of this axiom schema must be general recursive, if not indeed primitive recursive, in terms of their Gödel numbers. One need only show that the notion of mathematical proof which is obtained by adding this axiom schema to $A^2$ is expressible in $A^2$ by means of Gödel numbering in order to carry through an incompleteness proof along the lines given above, and this should be possible in virtue of theorems connecting general recursion and primitive recursion (for example, that any general recursive relation can be expressed by means of quantifiers and primitive recursive relations).
>
> Thus in a general way, the Gödel proof is not only a proof of incompleteness, but also a proof of incompletability. Since the only known way of making precise the notion of mathematical proof is the logistic system, the usual conclusion drawn

from the Gödel proof is that any precise formulation of arithmetic cannot be complete—a conclusion which shatters one of the hopes of the Hilbert program.

A genuine mathematical proof is not merely a sequence of formulae satisfying certain purely syntactic conditions (viz., every element of the sequence is either an element of the recursively specified set of "axioms," or else follows from formulae occurring earlier in the sequence by means of one of the recursively specified set of "rules of inference"). Rather, a genuine proof is what such a sequence of formulae semantically expresses: a line of reasoning, consisting of propositions, that conclusively demonstrates a proposition. Church's argument that since a proof must "carry final conviction" the notion of proof must be effective, if sound, applies directly to authentic proofs, and only derivatively, by extension, to their syntactic expression within a logistic system. If sound, the argument supports the broad conclusion that there must be an effective procedure that enables one to decide of any mathematical proposition whether it is or is not an element of **Ax**. In fact, assuming Church's thesis (in the form indicated above), his argument, if sound, supports the conclusion that the elements of **Ax** must *constitute a recursive set of propositions*, in the sense that there is a recursive numerical function that exactly characterizes a set of Gödel numbers of sentences of a possible formal language expressing each of the elements of **Ax**—e.g., a recursive function that yields 1 for the Gödel number of any axiomatic sentence and 0 for the Gödel number of any other sentence of the language in question. (See note 19.)

Church's argument, however, does not itself carry conviction. First, the fact that an auditor may justifiably doubt whether a purported proof is correct (and thus a genuine proof) does not entail that the line of reasoning in question does not after all conclusively demonstrate its conclusion with mathematical certainty (i.e., is not a genuine proof). A proof provides potential epistemic justification for conviction; the carrying of conviction is a horse of a different color. Whether the horse drinks from the water to which it is led is up to the horse. It is not unusual for a theorem to be proved before it is confirmed that the reasoning is thoroughly sound—sometimes well before this is confirmed even to the original author's satisfaction. In such cases, a potential epistemic justification for conviction is provided before conviction is carried—perhaps even before conviction is actually justified by its potential justification. Church's concern is with the auditor who questions whether a purported proof that has been spelled out in full, with a justification provided for each step, is correct. Often one can know that a given object has a given property even in the absence of an effective test for the property in question. Often one can even prove this. In particular, a given proof's correctness can be verified without applying any general test capable of verifying the correctness of any proof whatsoever. It is typically sufficient to re-check each step of the particular proof in question, and to verify that those particular steps are legitimate. One can do this by applying certain sufficient conditions for the justification of a step, even in the absence of a

complete set of such conditions, let alone a complete set of effectively decidable necessary and sufficient conditions. Where one auditor may doubt whether a particular piece of reasoning is a proof, another auditor may correctly see, without the benefit of an effective test, that the reasoning is perfectly sound. In that case, the reasoning can decisively establish its conclusion, at least for the second auditor.[31]

For that matter, even if there is an effective test, its mere existence does not put an end to the infinite regress of demanding a proof, then demanding a proof that the first proof is correct, then demanding a proof that the second proof is correct, and so on.[32] Nor does the existence of an effective test for proofs eliminate the possibility of justified doubt in a given case. To quell such doubt the test has to be applied to the proof in question. One may then question whether the test has been applied correctly. And even if one is satisfied that it was, one may justifiably doubt whether the purported test itself is correct. If an auditor wonders whether a particular proposition employed as an axiom in the proof is indeed antecedently known, it is no answer to point out that the formula expressing the proposition in question was written under the heading "AXIOMS" in setting up the primitive basis of a particular logistic system for mathematics (or is generated by an effective procedure for producing the logistic system's "axioms"). The auditor's question is not whether the formula is playing the role of an axiom in the purported proof, or whether it is called an 'axiom'; the question concerns the proposition expressed, whether it is genuinely known with mathematical certainty without independent mathematical proof. The so-called test simply assumes it is so, as it were, by stipulative fiat. The prospects are dim for an effective procedure for deciding whether it really is so. If such a procedure is required for there to be proof, mathematical ignorance is considerably wider than is currently realized.

On the contrary, the general issue of whether the entire line of reasoning in question is a proof is separate from the issue the proof itself is intended to settle: whether the theorem in question is true. The reasoning, if it is correct, enables an auditor to know the theorem with mathematical certainty. This is the purpose of the proof, its *raison d'etre*. To ask whether the purported proof is correct is to raise a separate, further question, an epistemological metaquestion related to the issue of whether one knows that one knows—a question that the auditor need not consciously consider in order to gain knowledge of the theorem with mathematical certainty on the basis of the proof. If the assumptions employed in the line of reasoning are *in fact* already known with mathematical certainty, and the inference rules are of the right character (so as to transfer mathematical certainty to the inferred conclusion), the reasoning can be of the right sort to establish its theorem conclusively, and to confer mathematical certainty for an auditor, even if the question of whether it does so is never raised—perhaps even if the question is raised and answered incorrectly, as long as the auditor continues to believe the theorem on the basis of the proof.

Church's argument is fundamentally Cartesian in character. It assumes that knowledge with mathematical certainty precludes the possibility of a certain

kind of justifiable doubt. Church supposes that in order genuinely to know something in mathematics one must be able to prove it beyond all possible justifiable doubt, and in order to do this one must be able to prove beyond justifiable doubt that one has done so, by applying an effective test. Descartes took this assumption a step further, requiring that all knowledge, mathematical or otherwise, be obtained by proof that is not subject to doubt of this sort. But the same mistake occurs even when the assumption is restricted to knowledge in mathematics with mathematical certainty. Despite the astounding feats of its champions, the assumption inexorably leads to skepticism. One may legitimately wonder, for example, how one knows (and in particular, whether by direct mathematical insight) that if integers $n$ and $m$ have the same successor then $n = m$. It is doubtful that anything other than Descartes's *Cogito* is completely immune from the kind of doubt raised by demanding indubitable proof that one's proof is a proof. One may even doubt whether the *Cogito* is.

None of this diminishes the epistemological power of mathematical proof. That power is awesome. Though not immune from Cartesian doubt, mathematical proof provides a way—indeed, the only way—to extend human knowledge with mathematical certainty beyond the severely narrow confines of **Ax**. Few epistemological mechanisms can achieve the kind of certainty that mathematical proof confers. In any event, even if Church's argument is not cogent, it does not follow that his conclusion is incorrect.

Though severely narrow, **Ax** may be remarkably diverse. As noted, **Ax** includes fundamental mathematical truths that no one in the entire history of human life will have ever apprehended—let alone believed, let alone known. Some elements of **Ax** involve concepts that are humanly apprehendible but of which no one will have ever formed a grasp. Some elements of **Ax** may be knowable only through modes of thought which are humanly possible but in which no one will have ever engaged. It may be that, though each element of **Ax** taken individually is humanly knowable with mathematical certainty, no possible human mind could apprehend all of them—let alone believe all of them, let alone know all of them with mathematical certainty. As far as Gödel's theorems go, the question is left open whether **Ax** is effectively decidable, or at least effectively enumerable, or enumerable at all—even whether the elements constitute a set. Hao Wang has reported that, though Gödel derives only a disjunction from his second incompleteness theorem, he believed Hilbert was correct in rejecting the second disjunct.[33] In light of Gödel's first theorem (and Church's thesis), Hilbert's optimism that the human mind is capable of solving any purely mathematical problem carries with it the view that the axioms of human mathematics are not effectively decidable. If every purely mathematical problem is humanly solvable in principle, then there is no effective procedure for listing the axioms of human mathematics. This would not mean that the human brain is not (among other things) an organic machine. It does mean that, insofar as Hilbert's optimism is correct, the theorem-proving capacity of the human mind far exceeds that of any theorem-proving mechanism whose deductive basis is effectively enumerable—a restatement of Gödel's disjunction. But one does not

have to be optimistic to appreciate that if the human brain is a machine, then it is a remarkable one—or else Gödel was not human (or both).

**Notes**

*The present essay grew out of meetings of the Santa Barbarians Discussion Group, organized by C. Anthony Anderson. I am indebted to the participants for encouraging my thoughts on the topic and for their comments on an early draft, and especially to Anderson for his valuable assistance.

1. Nagel and Newman, *Gödel's Proof* (New York University Press, 1958, 1967), at pp. 100–102.
2. In *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, 1989), at p. 416. Penrose revisits some of the issues in *Shadows of the Mind: A Search for the Missing Science of Consciousness* (Oxford University Press, 1994), and "Beyond the Doubting of a Shadow: A Reply to Commentaries on *Shadows of the Mind*," *Psyche*, 2, 23 (1996).
3. Lucas, "Minds, Machines and Gödel," *Philosophy*, XXXVI (1961); reprinted in A. R. Anderson, ed., *Minds and Machines* (Englewood Cliffs, NJ: Prentice-Hall, 1964), pp. 43–59, at 44, 47. A conclusion opposite in thrust from that of Lucas, Nagel and Newman, and Penrose is urged by Judson Webb, *Mechanism, Mentalism and Metamathematics* (Dordrecht: D. Reidel, 1980).
4. Putnam, "Minds and Machines," in Sidney Hook, ed., *Dimensions of Mind: A Symposium* (New York: New York University Press, 1960); reprinted in A. R. Anderson, ed., *Minds and Machines* (Englewood Cliffs, NJ: Prentice-Hall, 1964), pp. 72–97, at 77.
5. Lucas has replied, in "Minds, Machines, and Gödel: A Retrospect," in P. J. R. Millican and A. Clark, eds, *Machines and Thought: The Legacy of Alan Turing, Volume 1* (Oxford University Press, 1996), that the mechanist's claim that the proposed logistic system captures human mathematical reasoning is otiose unless the mechanist concedes that the system is consistent, and it is from this premise that Lucas derives the Gödelian sentence (p. 117). But unless the premise is itself proved mathematically, Lucas's derivation does not constitute a proof, or anything close to a proof. Given Lucas's objective, it is not sufficient for him to argue merely that mechanism cannot be proved.

   An assessment of the arguments and assertions of Lucas and Penrose is provided in Stewart Shapiro, "Incompleteness, Mechanism, and Optimism," *The Bulletin of Symbolic Logic*, 4 (September 1998), pp. 273–302.
6. "Some Basic Theorems on the Foundations of Mathematics and Their Implications," in Gödel's *Collected Works, III: Unpublished Essays and Lectures*, S. Feferman, J. W. Dawson, Jr., W. Goldfarb, C. Parsons, and R. N. Solovay, eds (Oxford University Press, 1995), pp. 304–323, at 310. See also Hao Wang, *A Logical Journey: From Gödel to Philosophy* (Cambridge, Ma.: MIT Press, 1996), especially chapters 6 and 7, pp. 183–246.
7. If the second alternative obtains—that there are purely mathematical questions of a certain sort that the human intellect is in principle unable to prove or disprove—this would seem to indicate that truth in pure mathematics is not reducible to provability (demonstrability), since the two are not even co-extensional. This conclusion relies on the assumption that if there are humanly undecidable purely mathematical

propositions, at least some have truth value. In fact, the propositions that are produced in Gödel's proof as undecidable in the logistic system in question are true (their negations false) provided the system is consistent, and are otherwise false. Any analogous propositions that are undecidable in human mathematics are likewise truth valued, so that truth in pure mathematics would provide no guarantee of certainty, or even potential certainty.

8. *Ibid.*, pp. 309–310.
9. Boolos, "Introductory Note to *1951," in Gödel's *Collected Works, III: Unpublished Essays and Lectures*, S. Feferman, J. W. Dawson, Jr., W. Goldfarb, C. Parsons, and R. N. Solovay, eds (Oxford University Press, 1995), pp. 290–304, at 294.
10. Boolos misformulates his conclusion by saying that if there is a Turing machine whose output is the set of sentences expressing just those mathematical propositions provable by a mind capable of understanding all propositions expressed by any sentence of the form $\ulcorner(\forall x)(\exists y)\phi(x, y) = 0\urcorner$, where $x$ and $y$ are sequences of integer variables and $\phi(x, y)$ is a polynomial with integer coefficients, then there is a true mathematical proposition *of this same technical sort* that cannot be proved by that same mind. It is evident that the conclusion Boolos intends is, rather, that if there is a Turing machine that produces exactly the mathematical truths provable by a mind with such comprehension, then there is a mathematical truth *that such a mind understands* (never mind what technical sort it is) but cannot prove. The latter carries with it the suggestion that the mind's incapacity, under the envisaged circumstances, does not result from a lack of understanding.

    The suggestion, however, is misleading. It is built into the case that the mind's theorem-proving capacity, by hypothesis, does not exceed the output of some theorem-enumerating Turing machine or other. This in itself says nothing about *why* the mind's mathematical prowess is thus limited. No logical inconsistency results by adding that the mind's limitations do not result from any lack of understanding. But neither has it been argued that the prospect of such a human mind—whose theorem-proving capacity coincides exactly with the output of Turing machine but nevertheless capable of fully understanding that which, as a consequence of the second theorem, it therefore cannot prove—is a real psychological possibility. These issues are in any case irrelevant. Boolos's intended conclusion follows from the second incompleteness theorem in the same way as the misformulated conclusion. Any possible generating activity whose potential output happens to coincide with the actual output a Turing machine—human or alien, animate or inanimate, with understanding or without—cannot in the relevant sense prove its own consistency.
11. **HuMath** is a proper subclass of the class of propositions, purely mathematical or otherwise, humanly knowable with mathematical certainty (i.e., with the same degree of certainty attainable in pure mathematics). It is not assumed that **Math**, or **HuMath**, is a set in the classical sense. Rather, the use of these terms in bold typeface in a sentence is to be regarded as an abbreviation for statements employing predicates that apply, respectively, to all purely mathematical truths and all purely mathematical truths humanly knowable with mathematical certainty. To say, for example, that a proposition $p$ is one of (or an "element of," or "belongs to") **Math** is to say no more (or less) than that $p$ is one of *these* propositions [the truths of pure mathematics], and to say that **HuMath** is a proper subclass of **Math**, is to say that all of *these* propositions [the truths of pure mathematics that are humanly knowable with mathematical certainty] are among *those* propositions [the truths of pure mathematics] but not vice versa. From the former it follows that *if* there is a set $M$ of all

truths of pure mathematics then $p \in M$, from the latter that again if there is such a set as $M$ then the subset $HM$ consisting of those elements humanly knowable with mathematical certainty is proper. Neither the antecedent of these conditionals nor its negation is presupposed.

12. One may take Heisenberg's Uncertainty Principle to entail this.
13. There may be other epistemic mechanisms, or potential epistemic mechanisms, that are, by contrast, precluded by a mind's being human, i.e. by the nature and biology of humanity. One such may be the $\omega$-rule of inference, which licenses the inference from premises, $\phi(0)$, $\phi(1)$, $\phi(2)$, and so on, to their generalization in $\ulcorner(\forall n)\phi(n)\urcorner$. Unless the human mind can reason with infinitely premises in a finite time span, it may be unable to draw inferences in accordance with this rule.
14. *Op. cit.*, p. 305.
15. This observation is to be taken in a sense in which it is beyond reasonable dispute. Some writers have mistakenly taken the incompleteness results to cast doubt on it. Thus Penrose writes: "Gödel's theorem...established...that the powers of human reason could not be limited to any accepted preassigned system of formalized rules" (*op. cit.*). It is incumbent on one who denies the observation to specify how the phenomenon of proof in mathematics might be otherwise understood while avoiding mathematical mysticism.

    Contemporary holistic empiricism holds that even knowledge of mathematical axioms is inextricably interconnected with all human knowledge taken as a whole, and thus ultimately empirical and fallible. Epistemological holism, however, is not inconsistent (as suggested by Shapiro, *op. cit.*) with the observation—well confirmed by actual practice—that knowledge in mathematics, unlike other disciplines, is furthered by an epistemologically special tool: mathematical proof from axioms, themselves humanly knowable with certainty without proof. Certainty, even mathematical certainty, does not entail immunity from error, let alone the absolute impossibility of human fallibility. (Some holists have proved their own fallibility on the very point in question.) Holistic empiricism maintains that the principles governing mathematical reasoning are ultimately judged, and conceivably might be revised, on ordinary empirical grounds. Whatever the shortcomings of this epistemological stance, it is not committed to denying the obvious role of mathematical proof in extending knowledge with certainty.
16. Axioms may be regarded as special rules of inference permitting inferences *ex nihilo*. On this conception, the deductive basis of a logistic system consists entirely of primitive inference rules. It is common, on the other hand, to minimize the set of primitive (non-axiom) inference rules by taking *modus ponens* as the only such rule, replacing every other inference rule,

    From $\phi_1$, $\phi_2$,..., and $\phi_n$, to infer $\psi$

    with all instances of the corresponding axiom schema $\ulcorner(\phi_1 \supset (\phi_2 \supset (...(\phi_n \supset \psi))...)\urcorner$.
17. Notice that the resulting axiom set is defined by a precise, finite rule. See note 23 below.
18. Under this restriction, the deductive system that takes all sentences expressing truths of arithmetic as axioms (though it exists) is disqualified as a logistic or formal system or theory. Thus Wang—the expositor who more than any other brought Gödel's philosophical views into the public domain—gives the following informal statement of the first incompleteness theorem (*op. cit.*, p. 3): *No formal system of mathematics can be both consistent and complete*; or alternatively, *Any consistent formal*

*theory of mathematics must contain undecidable propositions*. Similarly, C. Smorynski, "The Incompleteness Theorems," in J. Barwise, ed., *Handbook of Mathematical Logic* (Amsterdam: North Holland, 1977, 1983), pp. 821–865, states the theorem as follows: *Let **T** be a formal theory containing arithmetic. Then there is a sentence $\varphi$ which asserts its own unprovability and is [undecidable by **T** if **T** is $\omega$-consistent]* (p. 825).

19. The possible formal language in question should satisfy certain minimal constraints. As a matter of clarity, for example, ambiguity is precluded. The language is assumed to contain denumerably many expressions, to be bivalent (i.e., every sentence is either true or false and never both), and also such that a version of Tarski's theorem about truth holds for it. The language must also include the resources to express any mathematical concept that figures in any element of **HuMath**—including such concepts that have not yet been, or will never be, discovered or apprehended. (It is not assumed that the language contains only a finite number of logical or mathematical primitive constants.)
20. Gödel showed how to construct a formula along the lines of $\phi_{Cons}$ roughly for any logistic system suitable for arithmetic that includes the resources to designate any recursive function of integers and whose primitive deductive basis is recursive. For details, see Elliot Mendelson, *Introduction to Mathematical Logic* (New York: D. Van Nostrand, 1979), chapter 3, especially pp. 161–162. (See also the following note.) The notion of a mathematical *axiom*, in the sense of a fundamental, purely mathematical truth that is humanly knowable with mathematical certainty without independent mathematical proof, is not itself a purely mathematical notion and is not directly expressible in the language in question. Instead, assuming the elements of **Ax** are recursively enumerable, those propositions may be indirectly specified within the formula $\phi_{Proof}$, and hence within $\phi_{Cons}$, by means of a direct, purely mathematical specification of the recursive function $f$ that enumerates the Gödel numbers of sentences expressing those very propositions. As a corollary of Gödel's first incompleteness theorem, there can be no expression of the language that extensionally specifies **Math** in an analogous manner. (This is Tarski's theorem about truth; see the preceding note.)

    The formulae $\phi_{Proof}$ and $\phi_{Cons}$ do not strictly speaking semantically express the notions of proof from such-and-such axioms (those generated by recursive function $f$) and the consistency of such-and-such axioms and inference rules, respectively. The mathematical notions that are semantically expressed are, however, provably equivalent to these meta-theoretic notions. Indeed, the relationship is closer than mere provable equivalence; in a sense, the formulae are a code for the meta-theoretic notions. It is useful in the present context to think of the language of $\phi_{Proof}$ and $\phi_{Cons}$ as consisting of integers (Gödel numbers) functioning directly as expressions, and of the expression of a proof within the language—i.e., of a "proof" in the syntactic sense of a sequence of formulae—as a sequence of such integers-*qua*-formulae (rather than as its encoded representation by a single integer *via* the integer's prime factorization). Then $\phi_{Cons}$ semantically expresses that there are no such proof-sequences of integers culminating in integers one of which is the number-theoretic negation of the other (or something trivially equivalent to this).
21. Rosser, "Extensions of Some Theorems of Gödel and Church," *Journal of Symbolic Logic*, 1 (1936), pp.87–91. It follows from the result obtained by William Craig in "Axiomatizability Within a System," *Journal of Symbolic Logic*, 18, 1 (March 1953), pp. 30–32, that if **Ax** is recursively enumerable, then even if **Ax** is not itself recur-

sive, **HuMath** is primitive recursively axiomatizable. (Thanks to C. Anthony Anderson for calling my attention to the relevance of Craig's result.)

22. Shapiro, *op. cit.*, explains the first disjunct of Gödel's disjunction as the denial of the thesis that "all human arithmetic procedures are effective algorithms," and says that Gödel inclined instead to hold (with Lucas and Penrose) that "some of the routines and procedures that humans can employ...cannot be simulated on a Turing machine. There are inherently *non-computational* human arithmetic *procedures*" (pp. 277, 290, emphasis Shapiro's).
23. Gödel says in the passage quoted that his second incompleteness theorem "*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] no well-defined system of correct axioms can contain...all true mathematical propositions..." (The thrust of this remark is evidently better conveyed if the italicized phrase 'I believe' is deleted.) A similar remark is reported by Wang (*op. cit.*, p. 187): "There is a vague idea that we can find a set of axioms such that (1) all these axioms are evident to us; (2) the set yields all of mathematics. It follows from my incompleteness theorem that it is impossible to set up an axiom system satisfying (1) and (2), because, by (1), the statement expressing the consistency of the system should also be evident to me.—All this is explicitly in my Gibbs lecture." In order for someone to "set up" (i.e., fully specify) an infinite system of axioms, there would have to be an effective procedure for enumerating them. The term 'well-defined' is evidently a synonym in this context for 'recursively enumerable'.
24. "For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent" (*op. cit.*, in the passage quoted above from p. 309). Trivially, no contradiction is validly deducible from a set of truths. The casual manner of Gödel's remark creates the impression that this triviality is sufficient for the thesis, whereas strictly speaking, this justification is incomplete. Given a class of putative inference rules, one must know with mathematical certainty that every element of the class is valid in order to know with the same certainty that no falsehood, and hence no contradiction, is derivable from truths by their means. The validity of each inference rule of human mathematical reasoning is humanly knowable with mathematical certainty. Assuming the inference rules constitute an effectively decidable set, it is reasonable to suppose further that those very rules can be known with mathematical certainty to be one and all valid. Gödel's thesis then follows.
25. The argument I attribute to Gödel is significantly different from that to which Boolos's criticisms apply. Still other interpretations have been proposed. Wang (*op. cit.*, p. 185) apparently construes Gödel as arguing that if the axioms and inference rules of human mathematics were finite in number, then we could not know those very propositions and rules to be the basis of human mathematical knowledge, since otherwise we could know something about that basis (by confirming each element individually) that is not deducible from it—its consistency—and hence they would not be *all* the axioms and rules of human mathematics.

    I believe for a variety of reasons that this cannot be Gödel's argument. Curiously, Wang notes that the same line of argument yields another conclusion—one that is, in fact, significantly stronger—namely, that the basis of human mathematics

is infinite. Wang might mean to attribute to Gödel a somewhat different argument: We cannot know any finite basis to be the basis of human mathematics; for otherwise we could prove (by individual confirmation) something mathematical that is not deducible from that basis: that the basis (and hence all) of human mathematical knowledge is consistent. But this will not do either. That the basis of human mathematical knowledge—whatever it is, and whatever its size—is internally consistent is trivial and as certain as any mathematical theorem. This, however, is not reducible to a mathematical truth. It is an epistemological truism.

26. This probably yields the intent behind the following remark of Gödel's, reported by Wang (*op. cit.*, p. 186): "The incompleteness results do not rule out the possibility that there is a theorem-proving computer which is in fact equivalent to mathematical intuition. But they imply that, in such a—highly unlikely for other reasons—case, either we do not know the exact specification of the computer or we do not know that it works correctly." If **Ax** is recursively enumerable, so that a computer program might be written for proving theorems from it, then even if we were to write such a program, we could not know that its product is correct; otherwise we would also know what, according to the second incompleteness theorem, we cannot prove: its consistency.
27. Lucas, *op. cit.* (p. 44 of the reprinting in Anderson, *Minds and Machines*), declares that Gödel's results obtain for any formal system that is consistent and contains the natural numbers and the operations of addition and multiplication. In a later footnote (p. 52*n*6), he explicitly mentions the restriction that the primitive deductive basis be recursively enumerable.
28. Gödel evidently believed that the human mind does possess some self-validating insight of this sort. Cf. Wang, *op. cit.*, pp. 187–189.
29. Church, *Introduction to Mathematical Logic, I* (Princeton University Press, 1956), at pp. 50–51. See note 18 above.
30. *Ibid.*, pp. 53–54.
31. C. Anthony Anderson makes a related objection in "Alonzo Church's Contributions to Philosophy and Intensional Logic," *Bulletin of Symbolic Logic*, 4, 2 (June 1998), pp. 129–171, at 130–131.
32. Cf. Lewis Carroll, "What the Tortoise Said to Achilles," *Mind*, N.S. IV, 14 (April 1895), pp. 278–280.
33. Wang, *From Mathematics to Philosophy* (London: Routledge and Kegan Paul, 1974), at pp. 324–326.