# GOD, THE DEVIL, AND GÖDEL

> Although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others.
>
> Descartes, *Discourse on Method*

## Introduction

Descartes believed that (nonhuman) animals were essentially machines, but that humans were obviously not; for a machine

> could never use speech or other signs as we do when placing our thoughts on record for the benefit of others. . . . it never happens that (a beast) arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do.[1]

Descartes argued that Beasts were machines, because, like machines, they lacked certain capabilities of humans. La Mettrie accepted Descartes' reasoning.[2] But he disagreed about what machines could do. He felt that a machine could be constructed to do anything that a man could do—and thus agreed that man too was a machine. Descartes' argument took an obvious, proven fact about men (their speaking ability) and argued that machines did not have this ability. It's not obvious that he established his point about machines—for it's not obvious that machines *must* fail where it is indeed obvious that men succeed. And, as we saw, La Mettrie, for one, was unconvinced—unconvinced to the extent that he turned

---

[1] René Descartes *(Discourse on Method)*, in *The Philosophical Works of Descartes*, Elizabeth Haldane and G.R.T. Ross, translators (Cambridge: Cambridge University Press, 1912), Vol. I, p. 116.

[2] For an excellent account of the controversy, see Gunderson, K.R., "Descartes, La Mettrie, Language and Machines," *Philosophy*, 39 (1964).

The brief comments in this paper derive largely from Gunderson's discussion.

Descartes' arguments back upon him and, accepting Descartes' test, argued that man was a machine, for he believed that machines could be constructed to pass the test. So things were once more at a standstill.

In this paper, I concern myself with a similar claim—but one placed in a different philosophical climate. Like Descartes, John Lucas[3] (and others) has argued that man couldn't be a machine —for (he argues) man can do something which has been shown by Gödel to be beyond the reach of machines. The case is slightly different from Descartes'—since Descartes had, of course, nothing resembling a *proof* that machines were incapable of man's linguistic behavior. Remember La Mettrie. Perhaps our latter-day Cartesian has finally struck the mark—for now it has been shown that certain machines are *logically* precluded from succeeding at certain tasks. If it can be shown that man can perform these tasks, then the issue is settled once and for all.

## I

Paul Rosenbloom attributes to André Weil the saying that "God exists, since mathematics is consistent, and the Devil exists, since we cannot prove it."[4] Gödel, however, is the missing link, for he supposedly proved, very roughly speaking, that if mathematics *is* consistent we cannot prove it. It might therefore be said that he's clinched the case for Satan's existence. For, surely, if mathematics *isn't* consistent we can hardly prove that it *is*. So Satan wins either way. God, on the other hand, may not fare so well; for should mathematics not be consistent, we would have to look to another aspect of His infinite bounty for proof of His existence. These are heady matters, dark doings, and I do not propose to discourse on the present state of mathematical theology. I only raise them to give an indication of how far-reaching the philosophical consequences of Gödel's incompleteness theorems might be.

---

[3] Lucas, J.R., "Minds, Machines, and Gödel," reprinted in *Minds and Machines*, edited by A. R. Anderson (Englewood Cliffs: Prentice-Hall Inc., 1964). (Page references will be to this reprinting.)

[4] The Elements of Mathematical Logic. (New York: Dover Publications Inc., 1950), p. 72.

Yet, the example is not entirely frivolous. For though it might fail to illustrate the actual implications of Gödel's theorems, it does illustrate rather neatly the mechanism by which philosophical implications get alleged. The formula is simple. Take a view about what Gödel proved, i.e. what some theorem states (in this case, that we cannot prove the consistency of mathematics); add a more clearly philosophical view (in this case, what it takes to conjure up the Devil); mix, and you have, ready-made, a philosophical implication (in this case, Satan's existence). But to take a more realistic example, based on the same quotation, we can analyze an argument that might be given for the view that Gödel proved that *we* cannot prove that mathematics is consistent. The first ingredient in *this* case might be the second incompleteness theorem. For present purposes we can say that Gödel proved that any consistent formal system *S* for ordinary arithmetic containing Peano's familiar axioms or their equivalent contains certain formulas which express the consistency of *S* but none of which are theorems of *S*. To extract the consequence that the consistency of mathematics could not be proved, we must add the second ingredient: a philosophical view concerning what constitutes mathematics and what constitutes proof. It would suffice to identify provability with derivability in some particular formal system, and mathematics with the body of propositions expressible in that system, with 'expressible' suitably understood. These are, of course, not the only things that would suffice. Perhaps there are some *plausible* assumptions which, when conjoined to the second incompleteness theorem yield the desired result. That's not the point. The point is that you need some further premises—and in this case, clearly philosophical ones. For it is hardly a *mathematical* fact (if a fact at all) that absolute provability can be identified with formal derivability in some particular formal system (though to give a mathematically precise definition of provability is to make it a mathematical question *for provability so defined*). And I assert this without being in a position to argue it by presenting a neat distinction between what constitutes mathematics and what constitutes philosophy. I am making a sort of Duhemian point about philosophical implications. You need some to get some. At least any interesting ones. But this is a point which, once made should be forgotten. For, to take another example, I do *not* mean to imply that the following could not be a philosophical view:

(1) All mathematical propositions are expressible as closed formulas of *Principia Mathematica.*

(2) Of any pair [$A$,$-A$] of closed formulas of *Principia* exactly one is true under the intended interpretation.

(3) A closed formula of *Principia* is true if and only if it is a theorem of *Principia.*

and therefore

(4) A mathematical proposition is true if and only if some formula expressing it in *Principia* is a theorem of *Principia.*

I only wish to indicate that once Rosser, extending Gödel's first incompleteness theorem, showed that (2) and (3) are incompatible (because they jointly imply that of any such pair exactly one is a theorem), it is still possible consistently to maintain any combination of the above which doesn't include both (2) and (3). One may, as some are inclined to do, deny (2) (presumably weakening 'exactly one' to 'at most one'). Or one might deny (3), asking with Alan Ross Anderson "If the proof does not show that truth outruns provability in PM (provided the system is consistent, then what of importance does it show?"[5] The truth is, as these examples illustrate, that in a typical case, what is shown by Gödel's theorem (s) to be false is the conjunction of two (or more) philosophical views, say ($p \cdot q$), such that there always remain adherents of $p$ and adherents of $q$ (and alas, sometime also adherents of ($p \cdot q$) as well). However, in such a case what is usually *alleged* to have been disproved by Gödel is either that $p$ or that $q$. It therefore requires not only the metamathematical result, but also considerable philosophical argument to establish the desired conclusion. I wish in this paper to examine in detail one such alleged implication and to show that it conforms to the above pattern—that what is alleged to have been disproved by Gödel's incompleteness theorems has *not* been disproved. Rather it is a conjunction of the allegedly faulty principle with some other, possibly more dubious principles, that must be rejected. Considerable further argument must be given to show that some particular member of this conjunction must be discarded.

In our first example, we saw that one of the Satanic consequences

---

[5] "Mathematics and the Language Game," reprinted in *Philosophy of Mathematics,* eds. P. Benacerraf and H. Putnam (Englewood Cliffs: Prentice-Hall Inc., 1964), p. 486.

being drawn from Gödel's theorem is that it demonstrated a limitation on man's powers—in this case the power to prove things. Not everyone, however, has thought that Gödel's theorems established a limitation on human powers.[6] John Lucas, for one, takes Gödel's theorem to prove the falsity of Mechanism: "Gödel's theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines."[7] It is this view, and the arguments Lucas presents in its support, that I wish to examine here. But first I would like to lay the groundwork for the argument by explaining briefly what Gödel proved and how he went about it. I trust that the following exposition will prove too elementary to be of any interest to those who are familiar with the logical facts, and too compressed for those who are not. For the sake of future reference, however, it must be done. First a word about the relation of Gödel to machines.

Lucas identifies a (Turing) machine as the instantiation of a formal system. This is legitimate, since for any system which is formal in the required sense, there exists a theorem-proving machine which 'proves' all and only the theorems of that system. And similarly, for any theorem-proving Turing machine, there exists a formal system whose theorems are all and only the theorems the machine prints on its tape. Consequently, many theorems (such as Gödel's) concerning what can and cannot be done with formal systems have exact analogues concerning what can and cannot be done with Turing machines. This connects Gödel's theorem with *Turing machines.* But, if we are to believe its name, Mechanism is a thesis having to do with *machines, tout court.* What connection is there between Turing machines and honest machines? (For Turing machines are mathematically defined objects and need not have

[6] There is, of course, an obvious way in which they do establish such a limitation—and for the record we might note it here. I alluded above to Rosser's extension of the first incompleteness theorem. In that extension he showed that the systems Gödel discussed were either incomplete or inconsistent. It follows that no human can construct such a system which is both complete and consistent. Just as it follows from the nonexistence of a construction trisecting the angle using only a straightedge and compass that no human can carry out such a construction. Anything that has been shown to be impossible is impossible—even for humans.

[7] Lucas, J. R., *op. cit.*, p. 43.

much to do with real ones.) There is at least this much. Theoretically, one of our standard digital computers, properly programmed and given enough tape, could do the work of any Turing machine. It is an open question whether certain things which do not satisfy Turing's specifications might also count as machines (for the purpose of Mechanism). If so, then to prove that it is impossible to "explain the mind" as a Turing machine (whatever that might involve) would not suffice to establish Lucas's thesis—which is that it is impossible "to explain the mind as a *machine*." I mention this here to establish the link between Gödel's theorems and Lucas' thesis. I will not return to this point and, in what follows, I will not distinguish between Turing machines and real live ones.

## II

The first incompleteness theorem (hereafter 'Gödel I') states that any *formal system* which contains the usual operations on positive integers $(+,\times)$, quantification, identity, and truth functions is, if $\omega$-consistent, then incomplete. Now, a system *contains the operations of addition and multiplication* in the desired sense if it contains the usual axioms on them ($x+0=x$, $x+Sy=S(x+y)$, etc. . . .). It is *$\omega$-consistent* if no formula of the form $(\exists x)F(x)$ is a theorem, while each of $-F(0)$, $-F(1)$, $-F(2)$, etc., are also theorems. It is *incomplete* if there is some formula $F$, without free variables, such that neither $F$ nor $-F$ is a theorem; such a sentence would be called an undecidable sentence of the system. A system is (simply) *consistent* if there is no formula $F$ such that both $F$ and $-F$ are theorems. Since the class of systems we will be considering contain the familiar laws of the propositional calculus, this definition of consistency is equivalent to one on which a system is consistent if and only if there is some formula which is not a theorem. This follows easily from the remark that '$(p \,\&\, -p) \supset q$' is a tautology. Hence, if each of $F$ and $-F$ are theorems, then so is $(F \,\&\, -F)$ and by *modus ponens*, so is '$q$', for which you may substitute any formula whatever. It follows that an $\omega$-consistent system is also simply consistent.

So, in this vocabulary, one can formulate the usual laws of arithmetic. The standard number-theoretic concepts and operations, such as prime number, exponentiation, division, etc. can also be

defined. Finally, a system of this sort is a *formal system* if it satisfies the following requirement: there is an algorithm for deciding, given any sequence $F_1 \ldots F_n$ of formulas of the system, whether that sequence is a proof of $F_n$ in the system—i.e., whether every member of the sequence is either an axiom, or follows from previous formulas in the sequence by one of the rules of inference. Gödel showed that numbers can be paired with the formulas and finite sequences of formulas of a formal system in such a way that a sequence is a proof of its last member if and only if a certain relation holds between the number corresponding to the sequence and the number corresponding to its last formula. Calling these numbers, appropriately enough, 'Gödel numbers', and letting '$g(F)$' denote the Gödel number of $F$ the following can be done: it can be shown in the metalanguage that for any formal system $Z$ containing the apparatus previously mentioned, there exists a relation $R_Z$ among numbers, depending on $Z$, such that a sequence $F_1, \ldots, F_n$ is a proof in $Z$ of $F_n$ if and only if $R_Z\,[g(F_1, \ldots, F_n), g(F_n)]$. The next, and possibly most startling and difficult step, was to show that the systems of arithmetic under discussion are sufficiently rich to be able to define the relation $R$ for themselves. I.e., that for any such $Z$, there exists a predicate '$B(x,y)$' in the vocabulary of $Z$ such that for any two numbers $m$ and $n$, $R_Z(m,n)$ if and only if '$B(m,n)$' is a theorem of $Z$ (where '$m$' and '$n$' are the numerals of $Z$ representing the numbers $m$ and $n$).[8] This formalized the syntax of $Z$ in $Z$. Gödel then constructed a sentence $H$ with number $g(H)$ which had the form

(1) $(x)\text{—}B(x,g(H))$.

Under the intended interpretation, $H$ says that no number is the Gödel number of a proof (in $Z$) of the formula whose Gödel number is $g(H)$. So, under this interpretation, $H$ is true if and only if it is not provable in $Z$. Furthermore, whenever any formula $F$ is a theorem of $Z$,

(2) $\text{—}(x)\text{—}B(x,g(F))$

is also a theorem of $Z$—i.e., it is provable in $Z$ that $F$ is a theorem of $Z$. This follows from the above plus some elementary properties

[8] This formulation, using the bi-conditional, assumes the simple consistency, of Z.

of $Z$. Putting '$H$' for '$F$' in (2), we get, on the assumption that $H$ is a theorem, that

(3) $-(x)-B(x,g(H))$

is also a theorem. But (3) is the negation of (1). Therefore, if $H$ is a theorem, so is $-H$, since it expresses the statement that $H$ is a theorem: this makes $Z$ inconsistent. Therefore, if $Z$ is consistent, then $H$ is not among its theorems. This is the first half of *Gödel I*.

The other half of Gödel's argument establishes that if $Z$ is $\omega$-consistent, $-H$ is not a theorem of $Z$. But since, as we remarked above, $\omega$-consistency implies consistency, it follows that if $Z$ is $\omega$-consistent, neither $H$ nor $-H$ is provable in $Z$. This completes the sketch of the proof of *Gödel I*.

To sketch the proof of *Gödel II*, it suffices to make the following remarks. Consider any theorem of $Z$. Say '2=2'. Now, certainly if the *negation* of '2=2' is also a theorem, $Z$ is inconsistent. Similarly, if the negation of '2=2' is *not* a theorem, $Z$ is consistent. For if $Z$ were inconsistent, *every* formula would be a theorem. Hence the statement that the negation of '2=2' is not a theorem is equivalent with the assertion of the consistency of $Z$. But, letting $k$ be the Gödel number of '$-(2=2)$',

(4) $(x)-B(x,k)$

is a formula of $Z$ which says that no number is the Gödel number of a proof of the formula whose Gödel number is $k$—or, less cumbersomely, that '$-(2=2)$' is not a theorem of $Z$. Since, as we have seen, this is equivalent with the assertion that $Z$ is consistent, (4) is a formula of $Z$ expressing the consistency of $Z$. But then there must be infinitely many such formulas, since one may with equal effect put for $k$ the Gödel number of the negation of any theorem of $Z$.

Now, *Gödel II* states that if $Z$ is consistent, then no member of a certain class of formulas expressing in $Z$ the consistency of $Z$ is a theorem of $Z$, where that class is the one described in the previous paragraph. Gödel proved this by showing that the argument presented above for the first half of *Gödel I* is formalizable within $Z$. But that argument has as its conclusion that if (a) $Z$ is consistent then (b) $H$ is not a theorem of $Z$. Now, (a) is expressible in $Z$ by any one of a class of formulas. Let us pick one, say the one corresponding to (4), and abbreviate it 'Con($Z$)'. Also, as we have seen, $H$ itself expresses (b) in $Z$. Gödel therefore showed that '$\text{Con}(Z) \supset H$' *was a theorem of* $Z$. Therefore, if 'Con($Z$)' were also

a theorem, so would *H* be, by *modus ponens,* and, by the first half of *Gödel I, Z* would be inconsistent. So, *Z* is consistent if and only if 'Con(*Z*)' is not a theorem of *Z*.

## III

This is what Gödel proved. How do we go from here to the falsity of Mechanism? Well, let us first replace all talk of formal systems with the equivalent talk about machines; i.e., instead of speaking of formal systems let us speak of their corresponding Turing machines. Lucas wants to show that Gödel's theorems imply that no such machine could match the deductive output of a mind: that the mind can outstrip any machine in deductive prowess. We must take care here, for there are some trivial ways in which the thesis could be false, and Lucas is well aware of many of these and does not intend it in these ways. For example, it is clear that present digital computers can add, multiply, etc., much faster and more efficiently than any human. But this does not count. By the deductive output of a device, Lucas means simply the set of theorems that device is capable of proving, in the weak sense of 'capable' in which a whole life spent doing nothing but proving theorems, producing a meager total of seventeen, does not prove that there were not infinitely many theorems that I was capable of proving. So, to establish that I can outstrip machines in this sense, it would suffice to show that the set of theorems I am capable of proving properly includes that of any machine. Lucas's paper contains several statements of this view and arguments for it, all essentially similar, though I must confess that some are more puzzling than others. The rest of his paper is taken up with matters which are peripheral to the main point (discussions of inductive machines, comments on the self-reflective nature of consciousness, the re-establishment of moral responsibility, etc.). In the present section I will outline and briefly discuss the highlights of his view and arguments. It should come as no surprise that I will claim to find them inadequate, though, as I have said, interesting and puzzling. In the following section, I will reconstruct what I take to be the best case that one can make for a view such as Lucas's and extract from *that* what would seem to be the import of the Gödel theorems for the philosophical thesis of Mechanism.

After a brief (but somewhat faulty) exposition of Gödel's argument, Lucas presents his case:

> Gödel's 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 an adequate model of the mind, that minds are essentially different from machines [p. 44, my italics].

And

> 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 in this one respect they cannot do so well: in that for every machine there is a truth which it cannot produce as being true, but which a mind can . . . The Gödelian formula is the Achilles heel of the cybernetical machine. And therefore we cannot hope 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 [p. 47].

As further support, Lucas refutes some objections that might be raised against him. He correctly argues that the fact that for every consistent machine there is another equally consistent machine that can outstrip it in deductive capacity is no refutation of his point, which is that the mind can outstrip *every* machine, and hence couldn't be one itself.

The general drift of the argument should be clear, although its detailed form might not be. *Gödel I* establishes that any $\omega$-consistent formal system (or machine) adequate for arithmetic contains undecidable propositions. In particular, if a machine is consistent, then its Gödel formula $H$ cannot be among its theorems. But, Lucas claims, the mind can see it to be true, or "produce it as true." Therefore, $H$ is in the output of the mind but not of that machine.

Before discussing this argument in detail, let me mention some issues that I will *not* discuss. I won't delve into what Lucas might mean by "the mind" as opposed to particular people. Nor will I delve into the extremely interesting question of what general empirical conditions of adequacy one wishes to place on explanations of

mental phenomena: in order to make out a reasonable case for a kind of psycho-physiological Mechanism is it necessary to produce a model which duplicates the deductive output, *in Lucas's sense,* of any particular person? Possibly not, since no matter how infinite or even non-effective Lucas might argue that this would need to be, the actual output of any actual person is bound to be finite, *very* finite. These matters are all relevant to Lucas's claims that no Mechanistic explanation is possible—for no amount of evidence you could gather could strictly refute a strict finitist: although the range and diversity of the abilities displayed by humans may make some of his explanations unduly complicated if he dispenses with the simplifying assumption that humans internalize certain (infinite) recursive devices. These questions have been receiving a good deal of discussion in recent years, particularly in connection with developments in generative linguistics. I will bypass all these issues, assuming with Lucas that there is some empirical sense to supposing that someone has internalized certain rules and has the total output of those rules somehow in his repertory. So let us return to Lucas.

Two things must be noticed about this argument. The first is that it is not obviously valid—in the sense that it is not at all clear that the prowess *claimed* for the mind is one that *Gödel I* precludes for machines. For the argument seems to commit an equivocation:

> The conclusions it is possible for the machine to produce as being true will therefore correspond to the theorems that can be proved in the corresponding formal system. We now construct a Gödelian formula in this formal system. This formula cannot be *proved-in-the-system.* Therefore the machine cannot produce the corresponding formula as being 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 [p. 47].

Ignoring for the time being the respects in which this hangs on the unproved assumption of the consistency of the system, what is it that *Gödel I* precludes the machine (let's call her 'Maud') from doing? Evidently, it is to prove *H* (her Gödel formula) from *her* axioms according to *her* rules. But can Lucas do *that?* Just as evidently not. But what then *can* Lucas do which Maud cannot? One thing he might be able to do is give an *in*formal proof of *H:* informal in the sense that it cannot be formalized in Maud's system. But is

it clear that Maud cannot do this too? Maud is limited in the things for which she can offer formal$_{\text{Maud}}$ proofs. But does this limit the *informal* (i.e. not formalizable in Maud) proofs she can conjure up? It is not clear that it does. For Maud can carry out the Gödel argument on herself: by *Gödel II* she can prove '$\text{Con(Maud)} \supset H$'. Couldn't she thus ". . . convince [herself] that [$H$], although unprovable-in-[Maud], was nonetheless—in fact for that very reason—true . . ."? I don't see what in Gödel's arguments precludes this. To be sure, one might reply that no machine, not even one named Maud, can be said to convince itself that formulas are true. But of course, if that's why she can't, then we hardly need Gödel's theorems to establish it. And they don't help. As far as Gödel's theorems are concerned, provided that Maud doesn't delude herself into thinking that just because she has convinced herself that $H$ is true she has *proved* it, she can go on convincing herself of that, and of many other things besides. *Of course,* if convincing herself in this way were to count as *proof,* she would indeed be inconsistent. I shall return to this point in an appendix.

But let us try to construe the argument in such a way that it does not equivocate. There is such a construal, although Lucas never makes it clear that he would opt for it. He seems content to allow the sense in which he can prove things that Maud cannot to remain an informal one (cf., e.g., p. 56). But this can be made more precise. Let us interpret Lucas as claiming that he can prove $H_{\text{Maud}}$ in some formal system which is consistent and includes axioms for elementary arithmetic. So, to prove a formula is to derive it as a formal theorem of a consistent system which includes the postulates of arithmetic. But 'derive . . .' here cannot mean 'show that . . . is a theorem of'. For the relation '$T$ is a theorem of the formal system $S$' is one which has its analogue in Maud, under a suitable numbering of formal systems and of their vocabularies. We can limit our attention here to formal systems whose vocabulary is the same as Maud's. The set of all such formal systems (machines) can be enumerated by Maud, by assigning a number to each. Identifying the $i$-th formal system in the enumeration as $W_i$, it follows from Gödel's work that the relation '$F_1, \ldots, F_n$ is a proof of $F_n$ in $W_i$' has its counterpart in Maud: i.e. whenever a statement of that form is true, its translation into Maud is provable by Maud. Calling '$\text{Maud}_1$' the result of adding $H$ as an axiom to Maud, then '$H$ is a

theorem of Maud$_1$' is provable by Maud. So, thus interpreted, what Lucas can 'prove' does not differentiate him from Maud. If, however, we add the requirement that the axioms of each such system be themselves 'theorems' for Lucas, then certainly Maud cannot, in *that* sense, 'prove' *H*. But what *is* that sense? It cannot be formal derivability. It must be some sense of absolute provability in which everything Maud can prove is provable, but in which some things beyond Maud's reach are also provable. Whether or not there is such a sense Lucas never makes clear. Nor does he indicate what he thinks its properties are—though it seems obvious from his arguments that whatever is thus provable must also be true. In each case, the system must be not only consistent, but *correct* (i.e. have only true theorems). Or so it seems from what little he says. Clearly, in this sense, there are things that Maud cannot prove, and similarly for any other consistent machine. But now, Lucas is claiming that given any such machine, he can prove (in the same sense of 'prove') formulas which that machine cannot prove.

Thus interpreted, of course, the argument does not equivocate on 'prove'. And now, every proof that he produces will be a formal one in a suitable system—just like Maud's. But there will be no formal system in which every proof that he produces will be a proof. The union of all the formal systems he produces is not a formal system. So, certainly, *if* he can do this, he is not a machine, and, if he is not a machine, then Gödel's theorems do not preclude his being able to do it. But, of course, very little reason has been given for thinking that Lucas *can*.

Secondly, as Lucas notes later in his paper, in order to conclude that the Gödel sentence under consideration is true, one must conclude, or presuppose, that the machine in question is consistent. For *all* that follows from Gödel's theorems is that *if* the machine is consistent, *then H*. In order to conclude that *H*, one must be able to conclude that the machine is consistent. That, of course, does not appear in the cited arguments, which are apparently supposed to be valid in their own right. But Lucas is aware of this objection and discusses the general problem. He concludes that we can know that arithmetic and certain formal systems embodying it are consistent. For, as we have already seen, the mind can, in the relevant sense, "enunciate truths of arithmetic" [p. 47]. Add to this that "It . . . seems both proper and reasonable for the mind to assert

its own consistency . . . Not only can we fairly say simply that we *know* that we are consistent, apart from our mistakes, but we must in any case *assume* that we are, if thought is to be possible at all . . ." [p. 56]. And of course, we know this on the basis of informal arguments. But there is nothing wrong with that. There is "no objection to producing informal arguments either for the consistency of a formal system or of something less formal and less systematized. Such informal arguments will not be able to be completely formalized: but then the whole tenor of Gödel's results is that we ought not to ask, and cannot obtain, complete formalization" [p. 56].

Lucas presents one more argument which we should examine, one which is considerably more seductive than the last: Lucas imagines a sort of fencing match between himself and the Satanic Mechanist. As this is portrayed the Mechanist is challenged to produce a mechanical model of the mind—i.e. a blueprint for a Turing machine. He produces one, say Sam. Lucas then finds something that Sam cannot do, but which the mind can. The Mechanist may modify his example, and Lucas has shot at the revised model. If the Mechanist can produce a machine for which Lucas cannot find and prove a Gödel sentence, then, Mechanism, in the form of the Prince of Darkness, triumphs.

> . . . if he cannot, then . . . [the Mechanist's thesis] . . . is not proven: and since—as it turns out—he necessarily cannot, it is refuted. To succeed he must be able to produce some definite mechanical model of the mind . . . But since he cannot, in principle cannot, produce any mechanical model that is adequate, even though the point of failure is a minor one, he is bound to fail and mechanism must be false [p. 50].

This 'proof' of the falsity of mechanism rests on the claim that Lucas could find a flaw in any mechanical model of the mind that the Mechanist might conjure up, the flaw being that Lucas could find and prove some statement which was not a theorem of the machine of which this was a model. Let us give Lucas his due and more. Let us grant that he can, in this sense, find a flaw in any machine that the Mechanist constructs. *This does not prove his point.* For it is conceivable that a machine could do that as well. By shifting to this fencing match, Lucas abandons the claim that he can find a flaw *in any machine,* for the claim that he can find

one *in any machine the Mechanist can construct.* But the mechanist, despite his Satanic purpose, is but a man, and therefore probably a machine of relatively low order of complexity, given what's possible. There might well be a low limit on what kinds of machines his mind can develop and encompass. So, getting to a point where he can't produce anything Lucas can't fault proves nothing at all. But this argument is more convincing—precisely because Lucas is being pitted against a mere man, and we have fresh in our memories the example of Lucas tilting at all logically possible machines and laying them low one by one. Finally, as to the claim that the Mechanist *necessarily* cannot bring forth a faultless machine, it must be pointed out that even if it is a necessary truth that the Mechanist cannot produce a consistent and complete machine, *it is hardly necessary that Lucas should be able to make up the deficiency*. That is what is at issue.

But I think we have sufficiently labored the point. Lucas certainly does not present sufficiently cogent arguments for his case. Yet, I don't think we should abandon it all without making a serious effort to see what, if anything, does follow from Gödel's theorems concerning the possible existence of a mechanistic model for the mind. I will do this by trying to make more rigorous the vague arguments Lucas presents. Possibly something of interest will emerge.

## IV

I wish now to present an argument which contains the assumption that the mind is *at best* a Turing machine, employs both Gödel Theorems, and ends in a contradiction. I think that this argument fairly represents what underlies the vague ones that Lucas presents, and has the virtue that it is spelled out in detail. On the basis of it, I come to a rather different conclusion from Lucas concerning what implications Gödel's incompleteness theorems have for mechanistic philosophy. I will present the argument, discussing each step.

1. Let $S = [x \mid \text{I can prove } x]$

*S* represents my deductive output. It may be viewed as consisting of sentences under an interpretation. The sense of 'prove' involved is *not* one which limits *S* to the output of a machine, but it *does*

involve the assumption of correctness which we saw Lucas making and without which his arguments don't go through. It must be a sense of 'prove' which Lucas can use.

2. Let $S^* = [x \mid S \vdash x]$
$S^*$ is the closure of $S$ under the rules of first order logic with identity. I.e., anything derivable from $S$ by first order logic with identity is in $S^*$. I do not assume that $S$ is thus closed, for that would be gratuitous, and anyway, seems false. Some deductive claim might be too long or too complicated for me to follow.

3. $S^*$ is consistent.
Since every member of $S$ is true—I can't prove what is false—and first order logic preserves truth, every member of $S^*$ is true. $S^*$ is therefore highly consistent.

4. '$Con(S^*)$' $\epsilon$ $S$
Since 1-3 above constitutes a proof that $S^*$ is consistent, and since I produced that proof, it follows that I can prove that $S^*$ is consistent. Therefore, by the definition of $S$, '$Con(S^*)$' $\epsilon$ $S$.

5. '$Con(S^*)$' $\epsilon$ $S^*$
Since $S \subseteq S^*$, by 4. Note that this corresponds roughly to Lucas's statement on page 56 that he knows he is consistent. Actually, that would be the statement that '$Con(S)$' $\epsilon$ $S$. But 5 could then be derived. Note also that we haven't yet run afoul of Gödel. For it is only formal systems (i.e. machines) that are precluded by Gödel from proving their own consistency.

6. $(x)\,(W_x \subseteq S^* \supset Con\,(W_x))$
From 3. Since $S^*$ is consistent, so are all of its recursively enumerable subsets (see step 9 below for an explanation of the '$W$' notation).

7. '$(x)\,(W_x \subseteq S^* \supset Con\,(W_x))$' $\epsilon$ $S$
Since 1-6 is a proof of it which I produced. (I am being a bit cavalier here—but only to avoid the tedium of actually producing such a proof.)

8. '$(x)\,(W_x \subseteq S^* \supset Con\,(W_x))$' $\epsilon$ $S^*$
Since, by its definition, $S \subseteq S^*$, and by 7.

9. Suppose there is a recursively enumerable set $W_j$ such that
   a) '$Q \subseteq W_j$' $\epsilon$ $S^*$
   b) '$W_j \subseteq S^*$' $\epsilon$ $S^*$
   c) $S^* \subseteq W_j$

This will be the only assumption of the proof and c) will correspond to Lucas's assumption that I am a Turing machine.

That $W_j$ be recursively enumerable is simply the condition that it be the output of some theorem-proving Turing machine. For a fixed alphabet there exists an enumeration of all such machines by the device of Gödel-numbering all possible programs in some normal form. Letting $W_x$ be the machine (whose program is) numbered $x$, then this numbering lets us effectively recover the program from the number. (We will arbitrarily extend the enumeration by assigning the empty set to $W_z$ if $z$ is not the number of a program.) In this way, for each integer $i$, $W_i$ is a Gödel number of a recursively enumerable set of theorems; the theorems which the $i$-th machine can prove. It is important here that the machines be named in a 'transparent' way—that the program for the machine should be effectively recoverable from the name of the machine. Why this is important will emerge later in the proof. We should simply note it here as the reason for the '$W$' notation.

The first condition on $W_j$ is that I should be able to prove that it is adequate for arithmetic, in the usual sense. For $W_j$ to contain the first order closure of axioms as weak as those of theory $Q$ of Tarski, Mostowski, and Robinson[9] would suffice. Although Lucas obviously assumes that there is at least one formal system adequate for arithmetic all of whose theorems (and more) he can prove, he does not seem to distinguish between that assumption, call it '$A$', and the further assumption: that he can prove that $A$. It is the *latter* which he needs. The reason why will emerge more clearly in steps 15 and 17.

Similarly, the second is that I can prove of $W_j$ that it is a subset of my output. It will not suffice that $W_j$ merely *be* a subset of my output. *I must be able to identify it as such and by that name,* i.e. by a name which reveals its program. Again, the reason for the distinction will emerge later—in this case in step 14. Finally, the

[9] Tarski, A., Mostowski, A., and Robinson, R. M., *Undecidable Theories* (Amsterdam: North Holland, 1953), p. 51.

third condition on $W_j$ is that I be a subset of it. Or more precisely, that my closure under first order logic is a subset of a Turing machine. To assume this is to assume the negation of what Lucas wants to prove: for if $S^* \subseteq W_j$, then there is a machine ($W_j$) which can prove everything I can prove, and possibly more. Note that 9c is not equivalent to the supposition that I *am* a Turing machine. It fails to be equivalent in two ways. The most obvious, of course, is that it does not assert that $S^* = W_j$, though the identity follows from 9b and the fact that only truths can be proved. The less obvious way is that for *me* to be a Turing machine would be [not for $S^*$, but] for $S$ to be identical with $W_j$. So, 9c might be true and I might not be a Turing machine–but of course not in the way in which Lucas would have me fail to be one. Lucas argues that the mind is not a Turing machine on the grounds that it can prove *more* than any Turing machine. But it may still fail to be a Turing machine by being able to prove *less* than any given machine adequate for arithmetic and satisfying 9a and 9b. I don't know if this would be of much solace to Lucas, but the possibility is worth noting here. If his argument for the non-machinehood of the mind based on the supposition that the mind can prove more than any machine should fail, he might like to avail himself of the view that minds are limited to proving what turn out to be nonrecursively enumerable subsets of what perfectly sound machines can prove. In any event, I use 9c in this reconstruction, because this is the way that Lucas argues. He thinks that the reason we are not machines is that we can prove more than any machine.

The way in which 9 will be used is that what follows will be proved for any $W_j$ satisfying it, should there be any. If 9 leads to a contradiction, then there is no Turing machine satisfying all three parts of 9. What this would establish will, of course, require some discussion. But to continue the argument:

10. $Q \subseteq W_j$
This follows from 9a, since whatever can be proved must be true.

11. There is a formula $H$ having the Gödel properties $((x)\text{—}B(x,g(H)))$ such that if $H \in W_j$, so does $\text{—}H$, and $W_j$ is inconsistent. This is the version of *Gödel I* applicable to $W_j$, since by 10, $W_j$ is adequate for arithmetic, and by 9, it is (equivalent to) a formal system.

12. '$\mathrm{Con}(W_j) \supset H' \in W_j$
Again, since by 9 and 10 $W_j$ is formal and adequate for arithmetic. This is *Gödel II* for $W_j$. Note that 'Con' in the antecedent of the sentence mentioned in 12 is not italicized, while it is italicized wherever it appears previously in this proof (steps 3-8). This is because '$\mathrm{Con}(W_j)$' must be in the form appropriate to mirror the first half of *Gödel I* in $W_j$ itself. It is an abbreviation for some sentence in which the program for $W_j$ is coded in the proof predicate. In particular, we are not entitled to assume without proof that if '$Con(W_j)' \in S^*$ then necessarily so does '$\mathrm{Con}(W_j)$'. We will therefore keep them at least typographically distinct.

13. '$W_j \subseteq S^* \supset Con(W_j)' \in S^*$
This follows from 8, since $S^*$ is closed under first order logic.

14. '$Con(W_j)' \in S^*$
From 9b, 13, and the fact that $S^*$ is closed under *modus ponens*. The use made here of 9b illustrates the fact that not only need I be able to prove everything the machine $W_j$ can prove, but also I need to be able to prove *that*, even to obtain the consistency of $W_j$ in this weaker form. It would, of course have sufficed to assume 14 directly, instead of 9b. But we must be a bit circuitous if we are to give Lucas's argument the appearance of an argument. It is important to dig out some reasons why he might think that he can prove the consistency of $W_j$.

15. '$(x)(Q \subseteq W_x \supset (Con(W_x) \equiv \mathrm{Con}(W_x)))' \in S^*$
For ease of reference, let us call the quoted part '*B*'. *B* states that any recursively enumerable set containing $Q$ is consistent in the more general sense if and only if its Gödel consistency formula holds: more accurately, that the two consistency predicates are co-extensive for formal systems containing $Q$. To establish *B* it suffices to show that any such system has a formula which expresses in it the consistency of the system. This is, of course, part of *Gödel II*—and a part which I sketched in the second section of this paper. Given any system satisfying the antecedent of 15, I can construct the formula which states that '—(2=2)' is not a theorem of that system and show that formula to be true under the intended interpretation if and only if the system is consistent, in the more general sense applicable to nonformal systems. Note that I can only

show that I can do this for systems given to me in a canonical way: in such a way that the program for the machine is recoverable from the name. Lucas often relies on what he takes to be his ability to show of any formal system adequate for arithmetic that its Gödel sentence is true. But he fails to notice that it depends very much on *how* he is given that system. If given a black box and told not to peek inside, then what reason is there to suppose that Lucas or I can determine its program by watching its output? But I must be able to determine its program (if this makes sense) if I am to carry out Gödel's argument in connection with it. Knowing (and having proved) that any $\omega$-consistent machine contains undecidable sentences does not distinguish me from a machine (unless, of course, the very fact of knowing or proving something does). If the machine is not designated in such a way that there is an effective procedure for recovering the machine's program from the designation, one may well know that one is presented with a machine but yet be unable to do anything about finding the Gödel sentence for it. The problem becomes even more acute if one supposes the machine to be oneself—for in that case there may in fact be no way of discovering a relevant index—of finding out one's own program. It is for this reason that the antecedent of $B$ makes $Q$ a subset of $W_j$ thus designated. For then I don't need any additional assumptions to be able to construct the sentence '$\mathrm{Con}(W_j)$' which appears as the right half of the consequent. But since the antecedent is in that form, in order eventually to show that '$\mathrm{Con}(W_j)$' $\in S^*$ I must assume not only that is $Q$ a subset of $W_j$, but also that I can prove it. This is the reason for 9a. For now, from 15 and the closure of $S^*$ it follows that

16. '$Q \subseteq W_j \supset (Con(W_j) \equiv \mathrm{Con}(W_j))' \in S^*$,

and by 9a, 14 and the closure of $S^*$ under truth functions, we have

17. '$\mathrm{Con}(W_j)' \in S^*$.

*Now* $S^*$ contains the consistency of $W_j$ in a usable form. For now we can feed it into the antecedent of the formula of step 12, provided, of course, we can show that

18. '$\mathrm{Con}(W_j)' \in W_j$.

But this is where assumption 9c enters in. If $S^*$ is part of the output

of some Turing machine, then that Turing machine can 'prove' its own consistency and is therefore inconsistent, so we have

19. $H$, $\sim H$, are in $W_j$, and $W_j$ is inconsistent.
But it follows from 9b that $W_j \subset S^*$ therefore

20. $H$, $\sim H$ belong to $S^*$ and $S^*$ is inconsistent, contradicting 3.
$S$ is also inconsistent, but perhaps only semantically so, depending on its closure properties, about which nothing has been assumed. It would seem to follow that there is no machine satisfying 9.

If we review the above argument, we can see that the contradiction was derived from our tripartite assumption—9—plus definitions 1 and 2. Assuming that the definitions are not to be questioned, then it seems that we must reject 9; Lucas argues that Gödel's theorems imply the negation of 9c: that I can prove more than any Turing machine adequate for arithmetic, though, as I have it, the negation of 9c amounts to my closure under the rules of first order logic being able to prove more than any Turing machine. In any event, I suggest that they imply no such thing. At best Gödel's theorems imply the negation of the conjunction of 9a, 9b, and 9c. They imply that given any Turing machine $W_j$, either I cannot prove that $W_j$ is adequate for arithmetic, or if I am a subset of $W_j$, then I cannot prove that I can prove everything $W_j$ can. *It seems to be consistent with all this that I am indeed a Turing machine, but one with such a complex machine table (program) that I cannot ascertain what it is. In a relevant sense, if I am a Turing machine, then perhaps I cannot ascertain which one.* In the absence of such knowledge, I can cheerfully go around 'proving' my own consistency, but not in an arithmetic way—not using my own proof predicate. Ignorance is bliss. Of course, I might be an inconsistent Turing machine. Lucas's protestations to the contrary are not very convincing.

Actually, the result that we have obtained is a bit stronger than I make out in the last paragraph. If I am a Turing machine not only can I not ascertain which one, but neither can I ascertain of any instantiation of the machine that I happen to be that it is an instantiation of that machine. I need in no way identify it with myself. For if presented with a machine with my program in such a way that I can decipher its program, then by the argument I gave,

if everything else holds, I can 'prove' its consistency and we are both inconsistent, contradicting 3. This is represented in my argument by the fact that for the contradiction I needed only 9c, and not the stronger assumption that '$S^* \subseteq W_j$' belongs to $S^*$ (I can prove that it can prove everything I can). Together with 9b, that would have required me to know (be able to prove) that I am the Turing machine that I am. As things now stand, if there is a Turing machine that can prove everything my first order closure can, *then I cannot show of (any instantiation of) that machine both that it is adequate for arithmetic and that I can prove everything it can.* This result, though considerably weaker than what Lucas claimed, seems still significant. One person to whom I explained it concluded that Psychology as we know it is therefore impossible. For, if we are not at best Turing machines, then it is impossible, and if we are, then there are certain things we cannot know about ourselves or any others with the same output as ourselves. I won't take sides.

But if we ignore the possibility that 9a might be false, then we can restate the import of Gödel's theorems as follows. If I am a Turing machine, then I am barred by my very nature from obeying Socrates' profound philosophic injunction: KNOW THYSELF.

## Appendix

Lest the above prove too convincing, I should like to present one more argument, very briefly: In step 3 we used the principle that whatever I could prove was true. This might be formally stated, with the aid of the Gödel numbering apparatus of arithmetic as the schema (letting $S = [x \mid x$ is a Gödel number of something I can prove]).

A: $n \in S \supset N$

where '$n$' is replaced by the numeral representing the Gödel number of the formula which replaces '$N$'. By the usual diagonalization function, I can construct a sentence with Gödel number $k$ which has the form $-(k \in S)$[10]. Since $K$ is '$-(k \in S)$', the instance of A corresponding to this sentence is of course

[10] This is not strictly accurate, as the matter is a bit more complicated; but the complication is of no essential interest to us here.

(a) $k \in S \supset -(k \in S)$
which is truth functionally equivalent with
(b) $-(k \in S)$

But (b) is derivable using only arithmetic, the principle A, and first order logic (and in fact has been so derived by me). Therefore, by the same principles which enabled me to get 4, 7 and 15, since the Gödel number of (b) is $k$, it follows that
(c) $k \in S$,
contradicting (b).

This sobering little derivation should make us look respectfully at Maud, who, if you recall, convinced herself of $H$ and of her own consistency—though she was warned not to take too seriously being so convinced. In particular, she shouldn't go around blabbing it.

For our purposes, the contradiction derived above casts serious doubt on the meager results that we had been able to salvage for Lucas—for it appears that the principles used in the derivation of the contradiction themselves lead to contradiction, without the courteous help either of the Gödel theorems or of the special assumptions.

More precisely (and accurately), the following principles seem to underline the above derivation:
(A') if $n$ is the Gödel number of a sentence, then $\vdash n \in S \supset N$, and
if $n$ is not the Gödel number of a sentence, then $\vdash \in S \supset (1=1)$
A' is therefore certainly a formal axiom schema. Now let $Q'=Q \cup A'$. $Q'$ is a formal system and (b) is certainly a theorem of $Q'$. To obtain (c), however, something more is required. The following might do; it certainly represents the intuition that whatever I have proved I can prove *plus*, of course, that whatever I have derived as a theorem of $Q'$ I have proved. Without some sufficient conditions on provability, we can get nowhere. Letting '$B$' be the proof predicate for $Q'$, form $Q''$ by adding the following rule to $Q'$
(P) if $\vdash B(n,m)$, then $\vdash m \in S$,
i.e., $Q''$ is the closure of $Q'$ under P.

$Q''$ should, if our intuitions are correct, represent a subset at least of what I can prove. But now $Q''$ is inconsistent, for both (b) and (c) are theorems. Since (b) is a theorem of $Q'$ and $Q'$ is formal, letting $n$ be the Gödel number of a proof of (b) (the

sequence [(a),(b)] might constitute such a proof), we have, also as a theorem of $Q'$,

($b_1$) $B(n,k)$

But since $Q''$ includes $Q'$, both ($b$) and ($b_1$) are theorems of Q''. Now, by P

(c) $k \in S$

is a theorem of $Q''$, and $Q''$ is inconsistent.

Unfortunately, the detailed analysis of the ingredients that went into this contradiction would take us too far afield.[11] I hope and trust that the principles used in my reconstruction of Lucas's argument will survive sufficiently intact to preserve that argument and the implications we have drawn from it. But that is the topic of another paper.[12]

PAUL BENACERRAF

PRINCETON UNIVERSITY

---

[11] For an illuminating discussion of a similar paradox, in connection with knowledge, see Kaplan, D., and Montague, R., "A Paradox Regained," *Notre Dame Journal of Formal Logic*, 1 (1960).

[12] I would like to record my thanks to Hilary Putnam for his criticism of an earlier version of this paper.