Gödel's First Incompleteness Theorem

The following result is a cornerstone of modern logic:

Self-referential Lemma. For any formula $\psi(x)$, there is a sentence ϕ such that $(\phi \leftrightarrow \psi([\ulcorner\phi\urcorner]))$ is a consequence of Q.

Proof: You would hope that such a deep theorem would have an insightful proof. No such luck. I am going to write down a sentence ϕ and verify that it works. What I won't do is give you a satisfactory explanation for why I write down the particular formula I do. I write down the formula because Gödel wrote down the formula, and Gödel wrote down the formula because, when he played the logic game he was able to see seven or eight moves ahead, whereas you and I are only able to see one or two moves ahead. I don't know anyone who thinks he has a fully satisfying understanding of why the Self-referential Lemma works. It has a rabbit-out-of-a-hat quality for everyone.

We begin by defining a function f, as follows:

If m is the code of a formula $\theta(x,y)$, then f(m) is $\ulcorner(\exists y)(\theta([m],y) \wedge \psi(y))\urcorner$.

Otherwise, f(m) = 0.

[I should really be using "x_0" and "x_1" in place of "x" and "y" here, but too many subscripts are annoying.].

The set of formulas is Δ, the function that takes m to [m] is Σ, and the arithmetical version of the operation of substituting a term for free occurrences of a variable in a formula is Σ. Consequently, when we write out the definition of f, we get a Σ formula. Consequently, there is a formula $\mu(x,y)$ that functionally represents f in Q, so that, for each m, we have:

$$Q \vdash (\forall y)(\mu([m].y) \leftrightarrow y = [f(m)]).$$

Where k is the Gödel number for $\mu(x,y)$ (that is, $k = \ulcorner\mu(x,y)\urcorner$), let ϕ be the sentence:

$$(\exists y)(\mu([k],y \wedge \psi(y))$$

Then $\ulcorner\phi\urcorner = f(k)$, and we have:

$$Q \vdash (\forall y)\mu([k],y) \leftrightarrow y = [\phi]).$$

Consequently,

$$Q \vdash (\exists y)((\mu([k],y) \wedge \psi(y)) \leftrightarrow \psi([\phi])).$$

that is,

$$Q \vdash (\phi \leftrightarrow \psi([\ulcorner\phi\urcorner]).\boxtimes$$

Generalized Self-Referential Lemma. For any formula $\psi(x,z_1,z_2,...,z_n)$, there is a formula $\phi(z_1,z_2,...,z_n)$ such that:

$$Q \vdash (\forall z_1)(\forall z_2)...(\forall z_n)(\phi(z_1,z_2,...,z_n) \leftrightarrow \psi(\ulcorner\phi\urcorner,z_1,z_2,...,z_n)).$$

Proof: In the proof of the Self-Referential Lemma, the extra variables quietly go along for the ride.⊠

First Incompleteness Theorem. If Γ is a consistent Σ set of axioms that includes Q, then there is a true sentence that isn't provable in Γ.

Proof: By Craig's Theorem, we may assume Γ is Δ. Use the Self-referential Lemma to find a sentence ϕ such that

$$Q \vdash (\phi \leftrightarrow \neg\, \mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner])).$$

If ϕ were provable in Γ, then $\mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner])$ would be a true Σ sentence, hence provable in Q, hence provable in Γ. But also, since ϕ and $(\phi \leftrightarrow \neg\, \mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner]))$ are both provable in Γ, $\neg\, \mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner])$ is provable in Γ. This contradicts the consistency of Γ.

Since ϕ isn't provable in Γ, $\mathrm{Bew}_\Gamma(\ulcorner\phi\urcorner)$ is false. Hence $\neg\, \mathrm{Bew}_\Gamma(\ulcorner\phi\urcorner)$ is true, and ϕ is true.⊠

Corollary. Any Σ set of axioms that includes Q and in ω-consistent is incomplete, that is, there are sentences that are neither provable nor refutable in the theory.

Proof: As before, we may take our set Γ of axioms to be Δ. Let ϕ be the sentence constructed in the proof of the first incompleteness theorem. We saw already that ϕ isn't provable in Γ. Hence, for each m, m is not the code of a proof in Γ of ⌜ϕ⌝. Since the formula B_Γ strongly represents {<x,y>: x is the code of a proof of y in Γ} in Q, ¬ [m] B_Γ [⌜ϕ⌝] is provable in Q, hence provable in Γ. Since Γ is ω-consistent, (∃y) y B_Γ [⌜ϕ⌝] isn't provable in Γ. That is, Bew_Γ([⌜ϕ⌝]) isn't provable in Γ and so ¬ ϕ isn't provable in Γ.⊠

Because ϕ isn't provable in Γ, ¬ ϕ is consistent with Γ, and hence (∃y)y B_Γ [⌜ϕ⌝] is consistent with Γ, even though, for each m, ¬ [m] B_Γ [⌜ϕ⌝] is a consequence of Γ. Consequently, Γ ∪ {(∃y)y B_Γ [⌜ϕ⌝]} is an example of a consistent, ω-inconsistent theory.

Shortly after Gödel's proof, Barclay Rosser recognized that the hypothesis of ω-consistency was stronger than needed.

Stronger Corollary (Rosser). Any Σ set of axioms that includes Q and is consistent is incomplete.

Rosser proved this by constructing a sentence σ that is provably equivalent to:

(∀y)(y B_Γ [⌜σ⌝] → (∃z < y) z B_Γ [⌜ ¬ σ ⌝]).

The proof looks a lot like the proof that every Δ set is strongly representable and the proof that every Σ total function is functionally representable. Indeed, the idea of these proofs originated with Rosser's proof. Since we have used Rosser's idea to prove the every Δ set is strongly representable, we can prove a

Still Stronger Corollary (Tarski, Mostowski, and Robinson). There isn't any Δ set that includes the sentences provable in Q and excludes the sentences refutable.

Note that this implies Rosser's result, since, if Γ were complete, then the set of consequences of Γ would be Σ, and the complement of the set of consequences of Γ, which is {nonsentences} ∪ {sentences θ: ¬θ is a consequence of Γ}, would be Δ.

Proof: Suppose D were a Δ set that includes the sentences provable in Q and excludes the sentences refutable. Let δ(x) strongly represent D in Q, and use the Self-Referential Lemma to find a sentence η with

$$Q \vdash (\eta \leftrightarrow \neg\, \delta([\ulcorner \eta \urcorner])).$$

If $\ulcorner \eta \urcorner$ is in D, then $\delta([\ulcorner \eta \urcorner])$ is provable in Q, and so η is refutable in Q, contrary to the hypothesis that D excludes the sentences refutable in Q. So $\ulcorner \eta \urcorner$ must not be in D. Then $\neg\delta([\ulcorner \eta \urcorner])$ is provable in Q, and so η is provable in Q, contrary to the hypothesis that D includes the sentences provable in Q. Contradiction.⊠

A theory is said to be *decidable* iff the set of it's consequences is Δ. This usage is confusing. If you say that Peano Arithmetic is decidable, you might be making the true statement that there is an algorithm for determining whether a sentence is an axiom of PA, or you might be making the false statement that there's an algorithm for determining whether a sentence is a consequence of the axioms of PA. The established practice is to accept the latter reading, but it's a practice that makes for easy mix-ups. It's the result of a common failure to make it clear whether by a "theory" one means a set of axioms or the set of consequences of a set of axioms. Indulging in the unfortunate usage, we have the following:

Theorem. No decidable theory is consistent with Q.

Proof: This is where we use the fact that Q, unlike PA, can be written down as a single sentence. If Ω were a decidable theory consistent with Q, then {sentences ϕ: (Q → ϕ) is a consequence of Ω} would be a Δ set that includes the consequences of Q and excludes the sentences refutable in Q.⊠

> **Church's Theorem.** The set of sentences valid in the predicate calculus isn't Δ.

Proof: The set of valid sentences is consistent with Q, so it better not be a decidable theory.⊠

If ϕ is the Gödel sentence for PA – the sentence that asserts, "I am not provable in PA" – then we can recognize ϕ as true, even though ϕ isn't provable in PA. Consequently, PA doesn't include everything we can recognize as true. There is nothing special about PA in this. Replace PA by your favorite true Σ theory, and you'll get the same answer.

Let's see what happens if we take Γ to be the set of sentences of the language of arithmetic that we can recognize as true. By this, I don't mean merely the sentences we are able to prove in formal system. I mean the sentences we are capable of recognizing as true by any cognitive methods available to us. One of those cognitive methods is proof, and indeed we'll count a sentence as recognizably true if it is in principle derivable from other sentences that are recognizably true, even if the derivation is too complicated for us to carry it out in practice.

Assuming that Γ is Σ, we can form the Gödel sentence for Γ. We can recognize that ϕ is true, even though ϕ isn't derivable from Γ. But wait a minute. Γ was supposed to include everything we could recognize as true, yet ϕ is a sentence we can recognize as true, even though it's not a consequence of Γ.

The conclusion J. R. Lucas[1] wants us to draw from this is that the set of arithmetical sentences we can recognize as true isn't Σ. This is a philosophically important conclusion. It's fatal for the computational model of mind, which has it that the way to understand the human mind is to regard it as a gigantic computing machine. But the consequences go farther than that. The workings of the human mind can't even be simulated by a Turing machine. Now the operation of any ordinary mechanical device that takes symbolic inputs and yields symbolic outputs can be simulated by a Turing machine. This includes mechanical devices made of flesh and blood, with a carbon-based central processing unit, as well as devices made of steel and plastic with a silicon-based CPU. The human mind has within it some spark of divinity that cannot be mimicked by any merely mechanical system.

Most philosophers have wanted to resist Lucas's conclusion, but there has been no consensus what part of the argument to reject. One countervailing idea is this: The mere existence of the Gödel sentence has no surprising consequences. What makes the Lucas argument go is that we can explicitly write down the Gödel sentence, and once we have written it down, we can recognize its truth. In order for us to write down such a Gödel sentence, it is not enough that there exist a Σ set of sentences whose consequences are all the arithmetical sentences we can recognize as true. We have to be able to explicitly specify the Σ set. The conclusion to be drawn from the Lucas argument isn't that there isn't a computer program that simulates the operation of the human mind (or, at least, that part of human mental activity that is

---

1. "Minds, Machines, and Gödel," *Philosophy* 36 (1961): 120-24. The argument is taken up by Roger Penrose, *The Emperor's New Mind* (New York and Oxford: Oxford University Press, 1989).

concerned with arithmetic). It's that, if there is such a program, we can't say what it is, or can't say with enough precision to write down the program and its Gödel sentence.

While this response resists Lucas's conclusion, it nonetheless takes us some distance down the path Lucas has pointed us. If you take an ordinary mechanical device, like a clock or an adding machine, we see that it' possible to find out exactly how the de vice works. Simply unscrew the back and examine it closely. The practical difficulties that stand in the way of doing the same thing fora hum being are immense. But before Gödel's theorem we wouldn't have thought that *in principle* it was impossible for a human being to know her own program. It turns out, however, that a human being is fundamentally different from a mere mechanical device in that it isn't possible even in theory for a human being to know her own program, whereas it is possible, at least in theory, to know in detail how a mechanical device works. The spark of divinity is still there, albeit in embers.

One think to say in response is that, to know what sentences an agent is able to recognize as true, it is not enough to know in detail how the agent's mind works. We also have to know which arithmetical sentences are true, since in order to recognize a sentence as true, it has to be true. The most we can hope to determine just by examining an agent's mental state is what sentences the agent regards as true, that is, which sentences she believes. To say which of these regardings of a sentence as true ought to count as recognitions of true, we have to know about the natural number system as well as about the agent's inner states. For the Lucas argument to even get started, we have to take Γ be the set of sentences the agent's belief-forming processes permit her to regard as true, rather than those she is able to recognize as true. But with that emendation, does not the Lucas argument show that, even though it is in principle possible to

specify the outputs of a purely mechanican system by examining it closely, it isn't possible to do the same for the outputs of the human mind?

Perhaps not. What the First Incompleteness Theorem shows is that, if Γ is consistent, then the Gödel sentence for Γ is true. But how do we know that the set of arithmetical sentences we regard as true is consistent? Of course, we'd like to hope it's consistent, but you can't expect to tell by examining the belief-forming mechanism whether it ever generates a contradiction, for the same reason that you can't tell by examining it's program whethe av given Turing machine will halt. As we shall see in detail when we turn to the Second Incompleteness Theorem, consisntent arithmetical theories can't prove their own consistency. We can't be sure that our Gödel sentence is true because we can't be sure that our beliefs are consistent. What we get in the end is not a spark of divinity but a counsel to humility.

Another place that Gödel's theorem has caused philosophical consternation is more directly concerned with the foundations of mathematics. Before the nineteenth century, mathmaticians weren't insecure about what they were doing. What geometers studied was the structure of space (although the extend to which that structure was independent of our ways of representing space was controversial). After the advent of non-Euclidean geometry, this way of thinking was no longer tenable. Geometers studies many different, mutually incompatible systems, and they can't all describe the structure of reality.

The traditional attitude toward geometry was a version of Aristotelean, as opposed to Platonic, realism. According to Plato, mathematical entities exist eternally in a pure realm all to themselves, free of the vicissitudes of bodies and sensations. Before birth, our uncorrupted minds could perceive them directly, but, now that we are embodied, our mathematical understanding

consists in recollecting what before we could plainly see. Modern thinkers find this account of how mathematical knowledge is acquired implausible, so a central difficulty for mathematical Platonists — people who believe that mathematicians study actually existing things that are beyond the reach of space and time — is, How can we know about such things, when they don't affect us? Also, why are they so useful scientifically, when they are causally inert?

Aristotle's version or realism avoided this difficulty. The sense in which Aristotle's mathematical objects were "abstract' was different from the sense in which Plato's were. According to Aristotle, geometers studied ordinary physical things, regarding them from a "abstract"point of view that pays attention to size, shape, and position, but ignores color, texture, weight, smell, and taste. The attraction of such a viewpoint faded dramatically with the advent of non-Euclidean geometry.

Toward the end of the nineteenth century, an alternative conception of what mathematicians were doing became prominent. Mathematics isn't "about" anything. What mathematicians do is develop the consequences of systems of axioms. Which of those axioms are actually useful in describing material reality isn't a question for the mathematician; it's a question for the physicist. If there are any ways of interpreting the mathematical language so as to make the axioms true, we can be assured that the same interpretation will also satisfy the theorems. But whether there are such interpretations is not the mathematician's concern.

This "formalist" perspective accurately describes what algebraists do. A "group" is anything that satisfies the axioms of group theory, and what a group theorist does is to discover the properties true of everything that satisfies the axioms. The picture doesn't fit what number theorists do. Whatever we take the axioms of number theory to be, there will be further

statements we can recognize as true — true, that is, in the "intended model" of the language — that aren't consequences of the axioms. By showing this is so, the First Incompleteness Theorem makes the formalist position difficult to maintain.