

Gödel's Incompleteness Theorems of 1931*

We are now in a position to discuss the theorems that Gödel proved in 1931 and the consequences they have for Hilbert's program. Throughout, variables in italics (x, y , etc.) will be meaningful metamathematical variables, while boldface variables (\mathbf{x}, \mathbf{y} , etc.) will be variables of the first-order system \mathcal{N} . Most of the work needed to prove Gödel's results (the "99% perspiration" part) goes into establishing the following lemma.

Lemma. *The following predicate $A(x, y)$ is numeralwise expressible in \mathcal{N} :*

$A(x, y)$: x is the number of wf $\mathcal{F}_x(\mathbf{x}_1)$ and y is the number of a deduction whose last line is $\mathcal{F}_x(0^x)$ (where " $\mathcal{F}_x(0^x)$ " is $\text{sub}(\mathcal{F}_x, 0^x \rightarrow \mathbf{x}_1)$). ■

Let $\mathcal{A}(\mathbf{x}_1, \mathbf{y})$ be a wf which numeralwise expresses $A(x, y)$.¹ make the wf

$$(\forall \mathbf{y}) \sim \mathcal{A}(\mathbf{x}_1, \mathbf{y}); \tag{\dagger}$$

Let p be the Gödel number of wf (\dagger) , so that (\dagger) is $\mathcal{F}_p(\mathbf{x}_1)$. Use this number p to create the "Gödel wf" \mathcal{G} :

$$\mathcal{G} := \text{sub}(\mathcal{F}_p, 0^p \rightarrow \mathbf{x}_1) = (\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y}). \tag{*}$$

Although it will *prove* nothing, interpreting wf \mathcal{G} via the informal predicate $A(x, y)$ will make clear what motivated Gödel to write it down in the first place. The informal statement that corresponds to \mathcal{G} is

For all y , $A(p, y)$ is false.

Now, p is definitely the number of a wf, namely (\dagger) , so, for all y , it must be the second assertion of $A(p, y)$ that is false:

For all y , y is not the number of a deduction
whose last line is $\text{sub}(\mathcal{F}_p, 0^p \rightarrow \mathbf{x}_1)$.

Furthermore, $\text{sub}(\mathcal{F}_p, 0^p \rightarrow \mathbf{x}_1)$ is the wf \mathcal{G} itself. So, when interpreted, the meaning of \mathcal{G} is

For all y , y is not the number of a deduction
whose last line is \mathcal{G} ;

or, more succinctly,

wf \mathcal{G} cannot be proved.

Thus, when interpreted, \mathcal{G} asserts its own unprovability. To reiterate: this fact proves nothing about the formal system. But the fact that $\mathcal{A}(\mathbf{x}_1, \mathbf{y})$ numeralwise expresses $A(x, y)$ provides the way forward.

Gödel's First Theorem, Part 1. *If \mathcal{N} is consistent, then not $\vdash_{\mathcal{N}} \mathcal{G}$.*

Proof. I will prove the contrapositive. Suppose that $\vdash_{\mathcal{N}} \mathcal{G}$; let k be the number of a deduction whose last line is \mathcal{G} . It follows that $A(p, k)$ is true.² Because $\mathcal{A}(\mathbf{x}_1, \mathbf{y})$ numeralwise expresses $A(x, y)$, it then follows that

$$\vdash_{\mathcal{N}} \mathcal{A}(0^p, 0^k).$$

* This handout draws extensively on the account in S. C. Kleene's *Introduction to Metamathematics* (Van Nostrand, 1950).

¹ Note that $A(x, y)$ has built into it arithmetic that tracks the operation $\text{sub}(\mathcal{F}_x, 0^x \rightarrow \mathbf{x}_1)$. It is important—and easy—to arrange that the same formal variable \mathbf{x}_1 that is used in this subbing operation also be used to express the variable x in $A(x, y)$.

² Check this assertion carefully in your own mind.

Append to this deduction the following.

1.	$\mathcal{A}(0^p, 0^k)$	as above
2.	$((\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y})) \rightarrow \sim \mathcal{A}(0^p, 0^k)$	$\mathcal{K}5$
3.	$((\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y}) \rightarrow \sim \mathcal{A}(0^p, 0^k)) \rightarrow (\mathcal{A}(0^p, 0^k) \rightarrow \sim (\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y}))$	Tautology
4.	$\mathcal{A}(0^p, 0^k) \rightarrow \sim (\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y})$	MP(2,3)
5.	$\sim (\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y})$	MP(1,4)

Now, line 5 above is exactly $\sim \mathcal{G}$. We thus have a deduction of $\sim \mathcal{G}$ (as well as of \mathcal{G}) in \mathcal{N} ; so \mathcal{N} is not consistent. ■

To prove the second part of Gödel's First Theorem, I must introduce a stronger notion of consistency.

Definition. \mathcal{N} is said to be ω -consistent if there is no variable \mathbf{y} and no wf $\mathcal{B} = \mathcal{B}(\mathbf{y})$ for which **all** of the following are true:

$$\begin{array}{l}
 \vdash \mathcal{B}(0) \\
 \overset{\mathcal{N}}{\vdash} \mathcal{B}(0') \\
 \overset{\mathcal{N}}{\vdash} \mathcal{B}(0'') \quad \text{and} \quad \overset{\mathcal{N}}{\vdash} \sim (\forall \mathbf{y}) \mathcal{B}(\mathbf{y}) \\
 \overset{\mathcal{N}}{\vdash} \mathcal{B}(0''') \\
 \vdots
 \end{array}$$

It is important to note that if \mathcal{N} is ω -consistent, then it is also simply consistent—that is, consistent in the ordinary sense. Here is why. Take any wf \mathcal{B} and any variable \mathbf{y} —it doesn't matter which ones. If \mathcal{N} is ω -consistent, then either $\sim (\forall \mathbf{y}) \mathcal{B}(\mathbf{y})$ is not a theorem of \mathcal{N} , or there is at least one $k \in \mathbf{N}$ for which $\mathcal{B}(0^k)$ is not a theorem of \mathcal{N} . Either way, we have identified a wf that is not deducible in \mathcal{N} , so (since \mathcal{N} extends L) \mathcal{N} is consistent.

It is also important to note that in general, simple consistency does **not** imply ω -consistency. I will provide a counterexample below.

Gödel's First Theorem, Part 2. *If \mathcal{N} is ω consistent, then not $\overset{\mathcal{N}}{\vdash} \sim \mathcal{G}$.*

Proof. Suppose that \mathcal{N} is ω consistent. Then, as explained above, \mathcal{N} is also simply consistent, so by Gödel's First Theorem, Part 1, not $\overset{\mathcal{N}}{\vdash} \mathcal{G}$.) That means, for all $k = 0, 1, 2, \dots$, that $\mathcal{A}(p, k)$ is false. (Recall that p is the Gödel number of the wf (\dagger) .) Because $\mathcal{A}(\mathbf{x}_1, \mathbf{y})$ numeralwise expresses $A(x, y)$, we have

$$\begin{array}{l}
 \vdash \sim \mathcal{A}(0^p, 0) \\
 \overset{\mathcal{N}}{\vdash} \sim \mathcal{A}(0^p, 0') \\
 \overset{\mathcal{N}}{\vdash} \sim \mathcal{A}(0^p, 0'') \\
 \overset{\mathcal{N}}{\vdash} \sim \mathcal{A}(0^p, 0''') \\
 \vdots
 \end{array}$$

so again by ω -consistency,

$$\text{not } \overset{\mathcal{N}}{\vdash} \sim (\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y});$$

that is,

$$\text{not } \overset{\mathcal{N}}{\vdash} \sim \mathcal{G}. \quad \blacksquare$$

Counterexample. Assume that \mathcal{N} is consistent. Then (by Part 1 of Gödel's First Theorem) \mathcal{G} is a closed wf that is not a theorem of \mathcal{N} , so that if we adjoin $\sim \mathcal{G}$ as an axiom to \mathcal{N} , the resulting system—call it \mathcal{N}' —**will**

be consistent (see class notes from March 24 or Prop. 4.37 in the text). But \mathcal{N}' is **not** ω -consistent, since

$$\begin{array}{l} \vdash \sim \mathcal{A}(0^p, 0) \\ \mathcal{N}' \\ \vdash \sim \mathcal{A}(0^p, 0') \\ \mathcal{N}' \\ \vdash \sim \mathcal{A}(0^p, 0'') \\ \mathcal{N}' \\ \vdash \sim \mathcal{A}(0^p, 0''') \\ \mathcal{N}' \\ \vdots \end{array}$$

and also

$$\mathcal{N}' \vdash \sim (\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y}). \blacksquare$$

There is more to say about the concept of ω -consistency. First: Suppose that \mathcal{N} is consistent but not ω -consistent, so that not $\vdash \mathcal{G}$, but possibly $\vdash \sim \mathcal{G}$. (Keep in mind that we do not know for a fact that \mathcal{N} boasts either form of consistency.) Could the system possibly be complete in this case? The answer turns out to be No; in 1936, a mathematician named Barkley Rosser constructed a closed wf \mathcal{H} , analogous to \mathcal{G} but slightly more complicated, and proved that if \mathcal{N} is simply consistent, then neither \mathcal{H} nor $\sim \mathcal{H}$ is deducible in \mathcal{N} .

Second: if \mathcal{N} is consistent, then (by Part 1, again) we have

$$\text{not } \mathcal{N} \vdash (\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y});$$

but (as in proof of Part 2), we also have

$$\begin{array}{l} \vdash \sim \mathcal{A}(0^p, 0) \\ \mathcal{N} \\ \vdash \sim \mathcal{A}(0^p, 0') \\ \mathcal{N} \\ \vdash \sim \mathcal{A}(0^p, 0'') \\ \mathcal{N} \\ \vdash \sim \mathcal{A}(0^p, 0''') \\ \mathcal{N} \\ \vdots \end{array}$$

This is a type of *incompleteness* (called ω -incompleteness).

The consequences of the First Theorem for Hilbert's program are serious but not fatal. The formalists were attempting to construct a system that was provably both "omnipotent" (complete) and "benevolent" (consistent). The First Theorem shows that it is impossible for \mathcal{N} to have both of these properties. Indeed—if \mathcal{N} is consistent—it misses completeness by a huge amount. Here is why. One can form either of two consistent extensions of \mathcal{N} by adjoining either \mathcal{H} or $\sim \mathcal{H}$ (to get either \mathcal{N}'_1 or \mathcal{N}'_2 , say). Then for either of these, one can perform the whole construction again.³ One starts by making a predicate $A'(x, y)$ (analogous to $A(x, y)$) that will be numeralwise expressed by $\mathcal{A}'(\mathbf{x}_1, \mathbf{y})$, and one eventually ends up with a wf \mathcal{H}' —actually either \mathcal{H}'_1 or \mathcal{H}'_2 , depending upon whether one is working over \mathcal{N}'_1 or \mathcal{N}'_2 —such that neither \mathcal{H}' nor $\sim \mathcal{H}'$ is deducible in the extension \mathcal{N}' . Then, adjoining \mathcal{H}'_1 or $\sim \mathcal{H}'_1$ (respectively \mathcal{H}'_2 or $\sim \mathcal{H}'_2$) to \mathcal{N}'_1 (respectively \mathcal{N}'_2), one can make any of four inequivalent extensions extensions \mathcal{N}'' of \mathcal{N} . By iterating this process forever, one can find **uncountably many** inequivalent consistent extensions of \mathcal{N} , one for each of the uncountably many possible sequences of choices (of which wf, \mathcal{H} or $\sim \mathcal{H}$, to append at each stage). Each of these uncountably many extensions is consistent; so each of them has a model (see handout discussed March 24 and March 30). In other words *there are uncountably many inequivalent models of \mathcal{N} !*

³ Analogously: after applying Cantor's diagonal construction to construct a decimal that does not appear in a given sequence of decimals, you can append the new decimal to the top of the sequence and apply the construction again.

So \mathcal{N} is much, much weaker than the classical Peano system on which it is based. (The classical Peano system can be shown (with classical reasoning) to be categorical—that is, to have only one model.)

On the other hand, the system \mathcal{N} can prove quite a lot. So the Formalists still have some hope: if someone should show that \mathcal{N} is benevolent, then both the proof methods built into \mathcal{N} and the large parts of number theory that \mathcal{N} can deduce will have been shown to be trustworthy. This will be a vast improvement on what the Intuitionists permit.

Unfortunately, Gödel's Second Theorem puts paid to this hope. Let me begin to explain the Second Theorem by noting that, *via* the Gödel numbering, the assertion “ \mathcal{N} is consistent” is equivalent to a number-theoretical statement constructed from numeralwise expressible predicates.⁴ Start with the following two predicates.

$B(x, y)$: x is the number of wf \mathcal{F}_x , and y is the number of a deduction in \mathcal{N} whose last line is \mathcal{F}_x .

$C(x, z)$: x is the number of wf \mathcal{F}_x , and z is the number of a deduction in \mathcal{N} whose last line is $\sim\mathcal{F}_x$. The assertion “ \mathcal{N} is consistent” is then equivalent to the number theoretic assertion:

$$\text{For all } x: \begin{array}{l} \text{either there is no } y \text{ for which } B(x, y) \text{ is true} \\ \text{or there is no } z \text{ for which } C(x, z) \text{ is true.} \end{array} \quad (1)$$

Furthermore, if $B(x, y)$ and $C(x, z)$ are numeralwise expressed by $\mathcal{B}(\mathbf{x}, \mathbf{y})$ and $\mathcal{C}(\mathbf{x}, \mathbf{z})$ respectively—and if $\mathcal{B}(\mathbf{x}, \mathbf{y})$ and $\mathcal{C}(\mathbf{x}, \mathbf{z})$ *mean* $B(x, y)$ and $C(x, z)$ (*via* the usual interpretation of \mathcal{N})—then assertion (1) corresponds to the wf, below, which I will dub “CONSIS”:

$$(\forall \mathbf{x})((\forall \mathbf{y}) \sim \mathcal{B}(\mathbf{x}, \mathbf{y}) \vee (\forall \mathbf{z} \sim \mathcal{C}(\mathbf{x}, \mathbf{z}))) \quad (\text{CONSIS})$$

Now consider again the statement and proof of Gödel's First Theorem, Part 1. The statement is:

If \mathcal{N} is consistent, then \mathcal{G} is undeducible.

Since the proof uses Intuitionistic reasoning, there is every reason to believe that \mathcal{N} can reproduce this proof within itself; and—since \mathcal{G} , when interpreted, is the statement that \mathcal{G} cannot be deduced in \mathcal{N} —it should be possible for this to take the form of a deduction of the wf $\text{CONSIS} \rightarrow \mathcal{G}$:⁵

$$\frac{}{\mathcal{N}} \vdash \text{CONSIS} \rightarrow \mathcal{G} \quad (2)$$

So: now suppose that someone manages to prove that \mathcal{N} is consistent—either Intuitionistically or classically—with methods that have been built into \mathcal{N} . Then \mathcal{N} should be able to simulate this proof; that is, we should have

$$\frac{}{\mathcal{N}} \vdash \text{CONSIS}. \quad (3)$$

This would mean that \mathcal{N} could put (2) together with (3) to get

$$\frac{}{\mathcal{N}} \vdash \mathcal{G}, \quad (4)$$

so that (as in the proof of Gödel's First Theorem, Part 1) also

$$\frac{}{\mathcal{N}} \vdash \sim \mathcal{G}, \quad (5)$$

so that \mathcal{N} is not consistent. We have proved

Gödel's Second Theorem. *There is no proof of the consistency of \mathcal{N} by means that \mathcal{N} can simulate. ■*

⁴ This is only one of many ways to do this.

⁵ Gödel did not supply the \mathcal{N} -deduction of $(\text{CONSIS} \rightarrow \mathcal{G})$ in his original paper, but in 1939, Hilbert and Bernays did this.

Exercise. The ideas and methods here are delicate and subtle. Test your own understanding of them by finding the flaw in the following argument.

Suppose that \mathcal{N} is ω -consistent. Let \mathcal{N}' be the system obtained from \mathcal{N} by adjoining \mathcal{G} as an axiom. Then obviously

$$\vdash_{\mathcal{N}'} \mathcal{G},$$

so that by Gödel's Theorem, Part 1,

$$\vdash_{\mathcal{N}'} \sim \mathcal{G}.$$

Thus, \mathcal{N}' is inconsistent. But since

$$\text{not } \vdash_{\mathcal{N}} \sim \mathcal{G}$$

(by Gödel's Theorem, Part 2), \mathcal{N}' also has to be consistent. We thus have a contradiction. Conclusion: \mathcal{N} is not ω -consistent.

These two theorems of Gödel's have one final consequence for the Formalist program. (Assume throughout that \mathcal{N} is consistent.) Let $B(y)$ be a meaningful number-theoretic predicate (of one variable y) that is numeralwise expressed by the formal predicate $\mathcal{B}(\mathbf{y})$. Suppose also, for each natural number k , that $\mathcal{B}(0^k)$ (when interpreted in the standard model of \mathcal{N}) means the same thing that $B(k)$ does. Then: for each natural number k ,

$$B(k) \text{ is true} \implies \vdash_{\mathcal{N}} \mathcal{B}(0^k);$$

and

$$B(k) \text{ is false} \implies \vdash_{\mathcal{N}} \sim \mathcal{B}(0^k),$$

so by the consistency of \mathcal{N} ,

$$B(k) \text{ is false} \implies \text{not } \vdash_{\mathcal{N}} \mathcal{B}(0^k).$$

In short:

$$B(k) \text{ is true} \iff \vdash_{\mathcal{N}} \mathcal{B}(0^k).$$

I want to consider the wfs

$$(\forall \mathbf{y})\mathcal{B}(\mathbf{y})$$

and

$$(\exists \mathbf{y})\mathcal{B}(\mathbf{y}).$$

If $\vdash_{\mathcal{N}} (\forall \mathbf{y})\mathcal{B}(\mathbf{y})$, then (by suitable uses of $\mathcal{K}5$), $B(y)$ is true for all y . However, $B(y)$ can be true for all y without $(\forall \mathbf{y})\mathcal{B}(\mathbf{y})$ being deducible in \mathcal{N} . As was pointed out on page 2, an example is \mathcal{G} itself: because \mathcal{G} is not deducible in \mathcal{N} , “not $A(p, y)$ ” is true for all y ; but the wf $(\forall \mathbf{y}) \sim \mathcal{A}(0^p, \mathbf{y})$ is \mathcal{G} itself and hence nondeducible. In other words, we have only one implication:

$$\text{for all } y, B(y) \text{ is true} \iff \vdash_{\mathcal{N}} (\forall \mathbf{y})\mathcal{B}(\mathbf{y}).$$

For $(\exists \mathbf{y})\mathcal{B}(\mathbf{y})$, the situation is partially reversed—and perhaps even completely reversed. Consider: If $B(k)$ is true for some k , then for that k , $\vdash_{\mathcal{N}} \mathcal{B}(0^k)$; so by a suitable contrapositive of $\mathcal{K}5$, $\vdash_{\mathcal{N}} (\exists \mathbf{y})\mathcal{B}(\mathbf{y})$. However, the deducibility of $(\exists \mathbf{y})\mathcal{B}(\mathbf{y})$ might not guarantee the truth of the statement “There exists a y that makes $B(y)$ true.” The example here is $(\exists \mathbf{y})\mathcal{A}(0^p, \mathbf{y})$, which is $\sim \mathcal{G}$. We know that the statement

“There exists a y that makes $A(p, y)$ true”

is a false statement (assuming the consistency of \mathcal{N}); but if \mathcal{N} is not ω -consistent, there is no way to be sure that $(\exists \mathbf{y})\mathcal{A}(0^p, \mathbf{y})$ is not deducible in \mathcal{N} . In other words, we have only opposite implication:

$$\text{for some } y, B(y) \text{ is true} \implies \underset{\mathcal{N}}{\vdash} (\exists \mathbf{y})\mathcal{B}(\mathbf{y}).$$

This, to my mind, is a further blow to the hope of using metamathematics to prove to the Intuitionists that classical logic is trustworthy. We have just raised the possibility that *even if \mathcal{N} is consistent*, it might still be capable of deducing a false statement. So, even if we could show that \mathcal{N} is consistent, the Intuitionists would nevertheless be justified in viewing its theorems with suspicion. And indeed, long before the advent of the incompleteness theorems, Brouwer had rejected a proof of the consistency of \mathcal{N} , should it be found, as an acceptable answer to his reservations about classical mathematics. In 1923, in a heated exchange with Hilbert on precisely this issue, Brouwer said, “An incorrect theory which is not stopped by a contradiction is nonetheless incorrect, just as a criminal policy unchecked by a reprimanding court is nonetheless criminal.”

One final comment seems in order. Some people claim that Gödel’s First Theorem is evidence that what humans can do is not coextensive with what computers can do, because, for \mathcal{N} or for any consistent extension of it⁶, human beings can see that its Gödel *wf* (or Rosser *wf*) is true, whereas the system cannot deduce its own Rosser *wf*. This seems to me to be a flawed argument, because it misstates the First Theorem. By appealing to the First Theorem, human beings can see that *if the formal system is consistent*, then the system cannot deduce its own Rosser *wf* (which would then express a true statement); but (according to the Second Theorem), humans cannot prove the consistency of the formal system. But this is precisely coextensive with what the formal system itself can do, if it is consistent: it can deduce $\text{CONSIS} \rightarrow \mathcal{G}$, but it cannot deduce CONSIS .

⁶ To be precisely honest: any consistent extension with a decidable set of axioms.