Gödel's Second Incompleteness Theorem Explained in Words of One Syllable
Author(s): George Boolos
Source: *Mind*, Jan., 1994, New Series, Vol. 103, No. 409 (Jan., 1994), pp. 1-3
Published by: Oxford University Press on behalf of the Mind Association
Stable URL: https://www.jstor.org/stable/2253954

Gödel's Second Incompleteness Theorem Explained in Words of One Syllable

GEORGE BOOLOS

First of all, when I say "proved", what I will mean is "proved with the aid of the whole of math". Now then: two plus two is four, as you well know. And, of course, *it can be proved* that two plus two is four (proved, that is, with the aid of the whole of math, as I said, though in the case of two plus two, of course we do not need the *whole* of math to prove that it is four). And, as may not be quite so clear, it can be proved that it can be proved that two plus two is four, as well. And it can be proved that it can be proved that it can be proved that two plus two is four. And so on. In fact, if a claim can be proved, then it can be proved that the claim can be proved. And *that* too can be proved.

Now: two plus two is not five. And it can be proved that two plus two is not five. And it can be proved that it can be proved that two plus two is not five, and so on.

Thus: it can be proved that two plus two is not five. Can it be proved as well that two plus two *is* five? It would be a real blow to math, to say the least, if it could. If it could be proved that two plus two is five, then it could be proved that five is not five, and then there would be *no* claim that could *not* be proved, and math would be a lot of bunk.

So, we now want to ask, can it be *proved* that it can't be proved that two plus two is five? Here's the shock: no, it can't. Or to hedge a bit: *if* it can be proved that it can't be proved that two plus two is five, *then* it can be proved as well that two plus two is five, and math is a lot of bunk. In fact, if math is not a lot of bunk, then no claim of the form "claim *X* can't be proved" can be proved.

So, if math is not a lot of bunk, then, though it can't be proved that two plus two is five, it can't be proved *that* it can't be proved that two plus two is five.

By the way, in case you'd like to know: yes, it *can* be proved that if it can be proved that it can't be proved that two plus two is five, then it can be proved that two plus two is five.

"I wish he would explain his explanation"

If, as we shall assume, the whole of mathematics can be formalized as a formal theory of the usual sort (no small assumption), then there is a formula Proof (x,y)

Mind, Vol. 103 . 409. January 1994

of the language (of that theory) obtainable from a suitable description of the theory ("as" a formal theory) that meets the following three conditions:

(i) if $\vdash p$, then $\vdash \Box p$,
(ii) $\vdash (\Box (p \to q) \to (\Box p \to \Box q))$, and
(iii) $\vdash (\Box p \to \Box \Box p)$

for all sentences p, q of the language. We have written: $\Box p$ to abbreviate: $\exists x\mathrm{Proof}(x, \ulcorner p \urcorner)$, where $\ulcorner p \urcorner$ is a standard representation in the language for the sentence p. ($\ulcorner p \urcorner$ might be the numeral for the Gödel number of p.) "$\vdash$" is a preposed verb phrase (of our language) meaning "is provable in the theory". "Proof(x,y)" is a noun phrase (of our language) denoting a formula (of the theory's language) whose construction parallels any standard definition of "... is a proof of ___ in the theory". Thus, for any sentence p of the language, $\Box p$ is another sentence of the language that may be regarded as saying that p is provable in the theory.[1]

Conditions (i), (ii), and (iii) are called the Hilbert-Bernays-Löb derivability conditions; they are satisfied by all reasonable formal theories in which a certain small amount of arithmetic can be proved.

Since the theory is standard, all tautologies in its language are provable in the theory, and all logical consequences in its language of provable statements are provable.

It follows that for all sentences p, q,

(iv) if $\vdash (p \to q)$, then $\vdash (\Box p \to \Box q)$.

For: if $\vdash (p \to q)$, then by (i) $\vdash \Box (p \to q)$; but by (ii), $\vdash (\Box (p \to q) \to (\Box p \to \Box q))$, and then $\vdash (\Box p \to \Box q)$ by modus ponens.

$\bot$ is the zero-place truth-functional connective that is always evaluated as false. Of course $\bot$ is a contradiction. We shall need to observe later that $(\neg q \to (q \to \bot))$ is a tautology. If $\bot$ is not one of the primitive symbols of the language, it may be defined as any refutable sentence, e.g., one expressing that two plus two is five.

With the aid of $\bot$, there is an easy way to say that the theory is consistent: $\nvdash \bot$, i.e., $\bot$ is not provable in the theory. The sentence of the language stating that the theory is consistent can thus be taken to be $\neg \Box \bot$, which is identical with $\neg \exists x\mathrm{Proof}(x, \ulcorner \bot \urcorner)$.

We may prove Gödel's second incompleteness theorem, as well as the theorem that the second incompleteness theorem is provable in the theory ("the formalized second incompleteness theorem"), as follows.

Via the technique of diagonalization, introduced by Gödel (1931) in "On formally undecidable propositions... ", a sentence p can be found that is equivalent in the theory to the statement that p is unprovable in the theory, i.e. a sentence such that

1. $\vdash p \leftrightarrow \neg \Box p$
2. $\vdash p \to \neg \Box p$ truth-functionally from 1

[1] For an extended account of the application of modal logic to the concept of provability in formal theories, see Boolos (1993).

3. $\vdash \Box p \rightarrow \Box \neg \Box p$ by (iv) from 2

4. $\vdash \Box p \rightarrow \Box \Box p$ by (iii)

5. $\vdash \neg \Box p \rightarrow (\Box p \rightarrow \bot)$ a tautology

6. $\vdash \Box \neg \Box p \rightarrow \Box (\Box p \rightarrow \bot)$ by (iv) from 5

7. $\vdash \Box (\Box p \rightarrow \bot) \rightarrow (\Box \Box p \rightarrow \Box \bot)$ by (ii)

8. $\vdash \Box p \rightarrow \Box \bot$ truth-functionally from 3, 6, 7, and 4

9. $\vdash \neg \Box \bot \rightarrow p$ truth-functionally from 8 and 1

10. $\vdash \Box \neg \Box \bot \rightarrow \Box p$ by (iv) from 9

11. $\vdash \neg \Box \bot \rightarrow \neg \Box \neg \Box \bot$ truth functionally from 8 and 10.

(We have omitted outermost parentheses in (1) through (11).)

Thus if $\vdash \neg \Box \bot$, then both $\vdash \neg \Box \neg \Box \bot$, by (11), and $\vdash \Box \neg \Box \bot$, by (i), whence $\vdash \bot$, by the propositional calculus. So if $\nvdash \bot$, then $\nvdash \neg \Box \bot$.

Department of Linguistics and Philosophy
M.I.T.
Cambridge, MA 02139
USA

GEORGE BOOLOS

REFERENCES

Boolos, George 1993: *The Logic of Provability*. Cambridge: Cambridge University Press.

Gödel, Kurt 1931: "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I". *Monatshefte für Mathematik und Physik* 38, pp. 173-198, translated in his *Collected Works*, Volume I, ed. Solomon Feferman et al. Oxford: Oxford University Press, 1986, pp. 145-95.