

Incompleteness and undecidability

Peter Smith

October 26, 2009

In Episode 1, we introduced the very idea of a negation-incomplete formalized theory T . We noted that if we aim to construct a theory of basic arithmetic, we'll ideally like the theory to be able to prove *all* the truths expressible in the language of basic arithmetic, and hence to be negation complete. But Gödel's First Incompleteness Theorem says, very roughly, that a nice theory T containing enough arithmetic will always be negation incomplete.

Now, the Theorem comes in two flavours, depending on whether we cash out the idea of being 'nice enough' in terms of (i) the semantic idea of T 's being a *sound* theory, or (ii) the idea of T 's being a *consistent theory which proves enough arithmetic*. And we noted that Gödel's own proofs, of either flavour, go via the idea of numerically coding up inside arithmetic syntactic facts about what can be proved in T , and then constructing an arithmetical sentence that – via the coding – in effect 'says' *I am not provable in T*.

We ended by noting that, at least at the level of arm-waving description that of Episode 1, the Gödelian construction might look a bit worrying. After all, we all know that self-reference is dangerous – think Liar Paradox! So is Gödel's construction entirely legitimate?

Later, we'll see it that it certainly is. But first I think it might well go a little way towards calming anxieties that some illegitimate trick is being pulled, and is certainly of intrinsic interest, if we give a different sort of proof of incompleteness that doesn't go via any worryingly self-referential construction. So now read on . . .

4 Negation completeness and decidability

Let's start with another definition (sections, definitions and theorems will be numbered consecutively through these notes, to make cross-reference easier):

Defn. 13. *A theory T is decidable iff the property of being a theorem of T is an effectively decidable property – i.e. iff there is a mechanical procedure for determining, for any given sentence φ of T 's language, whether $T \vdash \varphi$.*

It's then easy to show:

Theorem 3. *Any consistent, negation-complete, axiomatized formal theory T is decidable.*

Proof For convenience, we'll assume T 's proof-system is a Frege/Hilbert axiomatic logic, where proofs are just linear sequences of wffs (it will be obvious how to generalize the argument to other kinds of proofs systems, e.g. where proof arrays are trees).

Recall, we stipulated that T 's formal language L has a finite number of basic symbols (of course that's no real restriction, since given two symbols, e.g. 'p' and "'", we can construct an infinite supply of composite symbols: p, p', p'', p''', etc.). Now, we can evidently put those basic symbols in some kind of 'alphabetical order', and then start mechanically listing off all the possible strings of symbols in some kind of order – e.g. the one-symbol strings, followed by the finite number of two-symbol strings in 'dictionary' order, followed by the finite number of three-symbol strings in 'dictionary' order, followed by the four-symbol strings, etc., etc.

Now, as we go along, generating sequences of symbols, it will be a mechanical matter to decide whether a given string is in fact a sequence of wffs. And if it is, it will be a mechanical matter to decide whether the sequence of wffs is a T -proof, i.e. check whether each wff is either an axiom or follows from earlier wffs in the sequence by one of T 's rules of inference. (That's all effectively decidable by Defns 2, 3). If the sequence is a kosher, well-constructed, proof, then list its last wff φ , i.e. the theorem proved.

So, we can in this way, start mechanically generating a list of all T -theorems (since any T -theorem has a proof, and by churning through all possible strings of symbols, we churn through all possible proofs).

And that enables us to decide, of an arbitrary sentence φ of our consistent, negation-complete T , whether it is indeed a T -theorem. Just start dumbly listing all the T -theorems. Since T is negation complete, eventually either φ or $\neg\varphi$ turns up (and then you can stop!). If φ turns up, declare it to be a theorem. If $\neg\varphi$ turns up, then since T is consistent, we know that φ is *not* a theorem.

Hence, there *is* a dumbly mechanical ‘wait and see’ procedure for deciding whether φ is a T -theorem. \square

We are, of course, relying here on a relaxed notion of effective decidability-in-principle where we aren’t working under time constraints (‘effective’ doesn’t mean ‘practically efficacious’ or ‘efficient’!). We might have to twiddle our thumbs for an immense time before one of φ or $\neg\varphi$ turns up. Still, our ‘wait and see’ method is guaranteed in this case to produce a result in finite time, in an entirely mechanical way – so this counts as an effectively computable procedure in the official generous sense (explained more in *IGT*, §2.2).

5 Capturing numerical properties in a theory

Here’s an equivalent way of rewriting the earlier Defn. 12:

Defn. 14. A property P is expressed by the open wff $\varphi(x)$ with one free variable in an arithmetical language L iff, for every n ,

- i. if n has the property P , then $\varphi(n)$ is true,
- ii. if n does not have the property P , then $\neg\varphi(n)$ is true.

A relation R is expressed by the open wff $\psi(x, y)$ with two free variables iff, for every m, n ,

- i. if m is R to n , then $\psi(m, n)$ is true,
- ii. if m is not R to n , then $\neg\psi(m, n)$ is true.

Now we want a new companion definition:

Defn. 15. The theory T captures the property P by the open wff $\varphi(x)$ iff, for any n ,

- i. if n has the property P , then $T \vdash \varphi(n)$,
- ii. if n does not have the property P , then $T \vdash \neg\varphi(n)$.

The theory T captures the relation R by the open wff $\psi(x, y)$ iff, for any m, n ,

- i. if m is R to n , then $T \vdash \psi(m, n)$,
- ii. if m is not R to n , then $T \vdash \neg\psi(m, n)$,

So: what a theory can *express* depends on the richness of its language; what a theory can *capture* (mnemonic: case-by-case prove) depends on the richness of its axioms and rules of inferences.

Ideally, of course, we’ll want any theory that aims to deal with arithmetic not just to express but to capture lots of arithmetical properties, i.e. to prove which particular numbers have or lack which properties.

But what sort of properties do we want to capture? Well, suppose that P is some effectively decidable property of numbers, i.e. one for which there is a mechanical procedure for deciding, given a natural number n , whether n has property P or not (see Defn. 1). Now, when we construct a formal theory of the arithmetic of the natural numbers, we will surely want deductions inside our theory to be able to track, case by case, any mechanical calculation that we can already perform informally. We don’t want going formal to *diminish* our ability to determine whether n has this property P . Formalization aims at regimenting what we can already do: it isn’t supposed to hobble our efforts. So while we might have some passing interest in more limited theories, we might naturally aim for a formal theory T which at least (a) is able to frame some open wff $\varphi(x)$ which expresses the decidable property P , and (b) is such that if n has property P , $T \vdash \varphi(n)$, and if n does not have property P , $T \vdash \neg\varphi(n)$. In short, we want T to capture P in the sense of our definition.

The working suggestion therefore is that, if P is any effectively decidable property of numbers, we ideally want a competent theory of arithmetic T to be able to capture P . Which motivates the following definition:

Defn. 16. *A formal theory T including some arithmetic is sufficiently strong iff it captures all decidable numerical properties.*

And it seems a reasonable and desirable condition on a formal theory of the arithmetic of the natural numbers that it be sufficiently strong.

6 Sufficiently strong theories are undecidable

We now prove a lovely theorem:

Theorem 4. *No consistent, sufficiently strong, axiomatized formal theory is decidable.*

Proof We suppose T is a consistent and sufficiently strong axiomatized theory yet also decidable, and derive a contradiction.

If T is sufficiently strong, it must have a supply of open wffs. And by Defn 2, it must in fact be decidable what strings of symbols are open T -wffs with the free variable ‘ x ’. And we can use the dodge in the proof of Theorem 3 to start mechanically listing such wffs

$$\varphi_0(x), \varphi_1(x), \varphi_2(x), \varphi_3(x), \dots$$

For we can just churn out all the strings of symbols of T ’s language, and mechanically select out the wffs with free variable ‘ x ’.

Now we can introduce the following definition:

$$n \text{ has the property } D \text{ if and only if } T \vdash \neg\varphi_n(n).$$

The supposition that T is a decidable theory entails that D is an effectively decidable property of numbers.

Why? Well, given any number n , it will be a mechanical matter to start listing off the open wffs until we get to the n -th one, $\varphi_n(x)$. Then it is a mechanical matter to form the numeral n , substitute it for the variable and prefix a negation sign. Now we just apply the supposed mechanical procedure for deciding whether a sentence is a T -theorem to test whether the wff $\neg\varphi_n(n)$ is a theorem. So, on our current assumptions, there is an algorithm for deciding whether n has the property D .

Since, by hypothesis, the theory T is sufficiently strong, it can capture all decidable numerical properties. So it follows, in particular, that D is capturable by some open wff. This wff must of course eventually occur somewhere in our list of the $\varphi(x)$. Let’s suppose the d -th wff does the trick: that is to say, property D is captured by $\varphi_d(x)$.

It is now entirely routine to get out a contradiction. For, just by definition, to say that $\varphi_d(x)$ captures D means that for any n ,

$$\begin{aligned} &\text{if } n \text{ has the property } D, T \vdash \varphi_d(n), \\ &\text{if } n \text{ doesn't have the property } D, T \vdash \neg\varphi_d(n). \end{aligned}$$

So taking in particular the case $n = d$, we have

- i. if d has the property D , $T \vdash \varphi_d(d)$,
- ii. if d doesn’t have the property D , $T \vdash \neg\varphi_d(d)$.

But note that our initial definition of the property D implies for the particular case $n = d$:

- iii. d has the property D if and only if $T \vdash \neg\varphi_d(d)$.

From (ii) and (iii), it follows that whether d has property D or not, the wff $\neg\varphi_d(d)$ is a theorem either way. So by (iii) again, d does have property D , hence by (i) the wff $\varphi_d(d)$ must be a theorem too. So a wff and its negation are both theorems of T . Therefore T is inconsistent, contradicting our initial assumption that T is consistent.

In sum, the supposition that T is a consistent and sufficiently strong axiomatized formal theory of arithmetic *and* decidable leads to contradiction. \square

So, if T is properly formalized, consistent and can prove enough arithmetic, then there is no way of mechanically determining what's a T -theorem and what isn't. We could, I suppose, call this result a *non-trivialization theorem*. We can't trivialize an interesting area of mathematics which contains enough arithmetic by regimenting it into a theory T , and then passing T over to a computer to tell us what's a theorem and what isn't.

It's worth remarking on the key construction here. We take a sequence of wffs $\varphi_n(x)$ (for $n = 0, 1, 2, \dots$) and then considering the (negations of) the wffs $\varphi_0(0)$, $\varphi_1(1)$, $\varphi_2(2)$, etc. This sort of thing is called a *diagonalizing*. Why?

Well just imagine the square array you get by writing $\varphi_0(0)$, $\varphi_0(1)$, $\varphi_0(2)$, etc. in the first row, $\varphi_1(0)$, $\varphi_1(1)$, $\varphi_1(2)$, etc. in the next row, $\varphi_2(0)$, $\varphi_2(1)$, $\varphi_2(2)$ etc. in the next row, and so on. Then the wffs of the form $\varphi_n(n)$ lie down the diagonal!

As we'll see, it is diagonalization (harmless and non-paradoxical) and not any worrying kind of self-reference that is really at the heart of Gödel's incompleteness proof.

7 A corollary about the decidability of logic

Defn. 17. *A formalized logic is decidable if the property of being a theorem of the logic – i.e. a sentence deducible from no premisses – is decidable.*

It is familiar that standard propositional logic is decidable (doing a truth-table test or a tree test decides what's a tautology, and the theorems are all and only the tautologies). It is familiar too that there's no obvious analogue for deciding of an arbitrary sentence whether it is theorem of standard first-order logic (a.k.a. the predicate calculus). But is there some other decision procedure?

Well, Theorem 4 now has an interesting corollary:

Theorem 5. *If there is a consistent theory with a first-order logic which is sufficiently strong and has a finite number of axioms, then first-order logic is undecidable.*

Proof Suppose Q is a consistent finitely axiomatized theory with a first-order logic and which is sufficiently strong. Since it is finitely axiomatized, we can wrap all its axioms together into one long conjunction, Q . And then, trivially, $Q \vdash \varphi$ if and only if $\vdash Q \rightarrow \varphi$; i.e. we can prove φ inside Q if and only if a certain related conditional is logically provable from no assumptions. So if the logic were decidable, and (1) we could mechanically tell whether the conditional $Q \rightarrow \varphi$ is a logical theorem, then (2) we could mechanically decide whether φ is a Q -theorem. But since Q is a consistent sufficiently strong formalized theory (2) is impossible. So (1) is impossible – the logic must be undecidable. \square

Much later, we'll find that there is indeed a consistent, finitely axiomatized, weak arithmetic with a first-order logic, which is sufficiently strong – the so-called Robinson Arithmetic Q fits the bill. So that will settle it: first-order logic really is undecidable.

8 Incompleteness again

Theorem 3 says: any consistent, negation-complete, axiomatized formal theory is decidable. Theorem 4 says: no consistent, sufficiently strong, axiomatized formal theory is decidable. It immediately follows that

Theorem 6. *A consistent, sufficiently strong, axiomatized formal theory cannot be negation complete.*

Wonderful! A seemingly remarkable theorem proved remarkably quickly. But what can we learn from it?

Well, note that – unlike Gödel’s own result – Theorem 6 doesn’t actually yield a specific undecidable sentence for a given theory T . And more importantly, it doesn’t tell us that T must have an undecidable *arithmetic* sentence.

So suppose we start off with a consistent ‘sufficiently strong’ theory T couched in some language which just talks about arithmetic matters: then this theory T is incomplete, and will have arithmetical formally undecidable sentences. But now imagine that we extend T ’s language (perhaps it now talks about sets of numbers as well as about numbers), and we give it richer axioms, to arrive at an expanded consistent theory U . Now, U will still be sufficiently strong if T is, and so Theorem 6 will still apply. Note, however, that as far as Theorem 6 is concerned, it could be that U repairs the gaps in T and proves every truth statable in T ’s language, while the incompleteness has now ‘moved outwards’, so to speak, to claims involving U ’s new vocabulary. Gödel’s result is a lot stronger: he shows that some incompleteness will always remain *even in the theory’s arithmetical core*.

Still, the current theorem is surprising enough. Set down a purely arithmetical theory. Either it won’t be sufficiently strong (will fail to prove some things you’d want a formalized arithmetic to prove) or it is incomplete (so still will fail to prove some arithmetic truths).

Finally, though, we should stress that the interest of Theorem 6 really depends on the notion of a sufficiently strong theory – defined in terms of the informal notion of a decidable property of numbers – being in good order. Well, obviously, I wouldn’t have written this Episode if the notion of sufficient strength was intrinsically problematic. However, making good that claim by given a sharper account of the notion of decidability takes quite a lot of effort! And it takes much more effort than we need to prove incompleteness by Gödel’s original method. So over the next Episodes, we are going to revert to exploring Gödel’s route to the incompleteness theorems.

At this point, you can usefully read Chs 4 and 6 of *IGT*. You might also skim Ch. 5 – but proof details there are perhaps only for real enthusiasts: in fact the arguments are about as tricky as any in the book, so I don’t want you to get fazed by them!