4 Undecidability and incompleteness

In Chapter 1, we introduced the very idea of a negation-incomplete, effectively axiomatized, formal theory T.

We noted that if we are aiming to construct a theory of basic arithmetic, we would 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 (at least as far as statements of basic arithmetic are concerned). But Gödel's First Incompleteness Theorem tells us that that's impossible: roughly, a nice enough theory T will always be negation incomplete for basic arithmetic.

Now, as we noted in Chapter 2, 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 which uses enough of the language of arithmetic*, or (ii) the idea of T's being a *consistent theory which proves enough arithmetic.* Then we saw in Chapter 3 that Gödel's own proofs, of either flavour, go via the idea of numerically coding up inside arithmetic itself 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 of Chapter 3, 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?

It certainly is, as should become quite clear over the coming chapters. But I think it might well go a little way towards calming the worry that some illegitimate trick is being pulled, and it is certainly of intrinsic interest, if we first give a somewhat different sort of proof of incompleteness, one that doesn't go via any explicitly self-referential construction. This proof will, however, introduce the idea of a *diagonalization argument.* And as we will later see that it is in fact 'diagonalization' rather than self-reference which is really the key to Gödel's own proof.

So now read on ...

4.1 Negation completeness and decidability

Let's start with another definition:

Defn. 17. *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 $T \vdash \varphi$.*

(Reality check: a theory T formally *decides* a sentence φ if T proves either φ or $\neg\varphi$; a theory T is *decidable* if for any φ we can effectively determine whether $T \vdash \varphi$. Two different notions then with similar terminology: in practice, though, you shouldn't get confused![1])

We can now easily to show:

Theorem 6. *Any consistent, negation-complete, effectively axiomatized formal theory is decidable.*

Proof For convenience, we can assume our theory T's proof system is a Frege/ Hilbert axiomatic logic, where proofs are just linear sequences of wffs (but it should be pretty obvious how to generalize the argument to other kinds of proof systems, where proof arrays are arranged e.g. as trees of some kind).

Recall, we stipulated (in Defns. 2, 3) that if T is a properly formalized theory, its formalized language L has a finite number of basic symbols. 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 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. to 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 in a properly formalized theory, by Defns. 2, 3). If the sequence is indeed a kosher, well-constructed, proof, finishing with a sentence, then list this last wff φ as a T-theorem.

We can in this way start mechanically generating a list that will eventually contain any T-theorem (since any T-theorem is the last sentence in a proof).

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 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 can declare that φ is *not* a theorem.

Hence, there *is* a dumbly mechanical 'wait and see' procedure for deciding whether φ is a T-theorem, a procedure which (given our assumptions about T) is guaranteed to deliver a verdict in a finite number of steps. □

[1] To fix ideas, note that a theory can be decidable without deciding every wff. For example, the toy propositional theory T of §1.3 is decidable because a truth-table test will determine whether $T \vdash \varphi$ for any wff φ of T's language. In particular, we see that $T \nvdash \mathsf{q}$ and $T \nvdash \neg\mathsf{q}$. Therefore T doesn't decide whether q, so T doesn't decide every wff.

We are, of course, relying here on a *very* relaxed notion of effective decidability-in-principle, where we aren't working under any practical time constraints or constraints on available memory etc. (so note, '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 our official generous sense (see the comments again on Defn. 1, or the further explanation in *IGT2*, §3.1).

4.2 Capturing numerical properties in a theory

Here's an equivalent way of rewriting part of an earlier definition:

Defn. 13. *A numerical property P is* expressed *by the open wff $\varphi(\mathsf{x})$ with one free variable in a language L which contains the language of basic arithmetic iff, for every n,*

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

(Recall, $\overline{\mathsf{n}}$ indicates L's standard numeral for n.) And now we want a new companion definition:

Defn. 18. *The theory T captures the numerical property P by the open wff $\varphi(\mathsf{x})$ iff, for any n,*

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

Note the contrast: 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. (To be honest, 'represents' is much more commonly used than my 'captures', but I'll stick here to the slightly idiosyncratic but memorable jargon adopted in *IGT2*.)

Just as a theory can express two-place relations (say) as well as monadic properties, a theory can capture relations as well as properties. So (for future reference) we expand our definition in the obvious way like this:

Defn 18. *(continued) The theory T captures the two-place numerical relation R by the open wff $\varphi(\mathsf{x},\mathsf{y})$ iff, for any m, n,*

i. if m has the relation R to n, then $T \vdash \varphi(\overline{\mathsf{m}},\overline{\mathsf{n}})$,
ii. if m does not have the relation R to n, then $T \vdash \neg\varphi(\overline{\mathsf{m}},\overline{\mathsf{n}})$.

But for the moment, let's concentrate on the case of capturing properties.

Ideally, of course, we'll want any theory that aims to deal with arithmetic not just to express but to capture lots of numerical properties, i.e. to prove which particular numbers have or lack these properties. But what particular 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 again). So we can, in principle, run the procedure to decide whether n has this property P. 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 a property P. Formalization aims at regimenting what we can in principle already do: it isn't supposed to hobble our efforts. So while we might have some passing interest in more limited theories, we will ideally aim for a formal theory T which at least (a) is able to frame some open wff $\varphi(\mathsf{x})$ which expresses the decidable property P, and (b) is such that if n has property P, $T \vdash \varphi(\overline{\mathsf{n}})$, and if n does not have property P, $T \vdash \neg\varphi(\overline{\mathsf{n}})$. In short, we will want T to capture P in the sense of our definition.

The 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. 19. *A formal theory T is sufficiently strong iff it captures all decidable numerical properties.*

(It would be equally natural, of course, to require the theory also capture all decidable relations and all computable functions – but for present purposes we don't need to worry about that.)

In sum: it seems a reasonable and desirable condition on an ideal formal theory of the arithmetic of the natural numbers that it be sufficiently strong – when *we* can (or at least, given world enough and time, *could*) decide whether a particular number has a certain property, the *theory* can do it.

4.3 Sufficiently strong theories are undecidable

We now prove a lovely theorem (take it slowly, savour it!):

Theorem 7. *No consistent, effectively axiomatized and sufficiently strong formal theory is decidable.*

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

If T is sufficiently strong, it must have a supply of open wffs (for expressing numerical properties). And by Defn 2, it must in fact be decidable what strings of symbols are T-wffs with the free variable 'x'. And we can use the dodge in the proof of Theorem 6 to start mechanically listing such wffs

$$\varphi_0(\mathsf{x}), \varphi_1(\mathsf{x}), \varphi_2(\mathsf{x}), \varphi_3(\mathsf{x}), \ldots.$$

For we can just churn out all the strings of symbols of T's language 'in alphabetical order', and then mechanically select out the wffs with free variable 'x'.

So now we can introduce the following definition:

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

That's a perfectly coherent stipulation. Of course, property D isn't presented in the familiar way in which we ordinarily present properties of numbers: but our definition tells us what has to be the case for n to have the property D, and that's all we will need.

Now for the key observation: our 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, $\varphi_n(\mathsf{x})$. Then it is a mechanical matter to form the numeral $\overline{\mathsf{n}}$, substitute it for the variable, and then 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 resulting wff $\neg\varphi_n(\overline{\mathsf{n}})$ 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 $\varphi(\mathsf{x})$. Let's suppose the d-th wff does the trick: that is to say, property D is captured by $\varphi_d(\mathsf{x})$.

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

if n has the property D, $T \vdash \varphi_d(\overline{\mathsf{n}})$,
if n doesn't have the property D, $T \vdash \neg\varphi_d(\overline{\mathsf{n}})$.

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

i. if d has the property D, $T \vdash \varphi_d(\overline{\mathsf{d}})$,
ii. if d doesn't have the property D, $T \vdash \neg\varphi_d(\overline{\mathsf{d}})$.

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

iii. d has the property D if and only if $T \vdash \neg\varphi_d(\overline{\mathsf{d}})$.

From (ii) and (iii), it follows that whether d has property D or not, the wff $\neg\varphi_d(\overline{\mathsf{d}})$ is a theorem either way. So by (iii) again, d does have property D, hence by (i) the wff $\varphi_d(\overline{\mathsf{d}})$ 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* is decidable leads to contradiction. □

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 lovely result a *non-trivialization theorem.* We can't trivialize an interesting area of mathematics which contains enough arithmetic by regimenting it into an effectively axiomatized theory T, and then just pass T over to a computer program to tell us what's a theorem and what isn't. There can't be such a program.

4.4 Diagonalization

Let's highlight the key construction here. In defining the property D, for each n, we take the n'th wff $\varphi_n(\mathsf{x})$, and plug in the standard numeral for the index n (before taking the negation of the result). This sort of thing is called *diagonalization.* Why?

Well, just imagine the square array you get by writing $\varphi_0(\overline{0})$, $\varphi_0(\overline{1})$, $\varphi_0(\overline{2})$, etc. in the first row, $\varphi_1(\overline{0})$, $\varphi_1(\overline{1})$, $\varphi_1(\overline{2})$, etc. in the next row, $\varphi_2(\overline{0})$, $\varphi_2(\overline{1})$, $\varphi_2(\overline{2})$ etc. in the next row, and so on. *Then the wffs of the form $\varphi_n(\overline{\mathsf{n}})$, including $\varphi_d(\overline{\mathsf{d}})$, lie down the diagonal through the array.*

We'll be meeting other instances of this sort of construction. And it is a diagonalization of this kind that is really at the heart of Gödel's incompleteness proof.[2]

4.5 Incompleteness again!

So we have now shown:

Theorem 6. *Any consistent, negation-complete, effectively axiomatized formal theory is decidable.*

Theorem 7. *No consistent, effectively axiomatized and sufficiently strong formal theory is decidable.*

It immediately follows that

Theorem 8. *A consistent, effectively axiomatized, sufficiently strong, formal theory cannot be negation complete.*

Wonderful! A seemingly remarkable theorem, proved remarkably quickly (this time without having to simply assume unproved lemmas along the way).[3]

Note, though, that – unlike Gödel's own proof strategy – Theorem 8 doesn't actually yield a specific undecidable sentence for a given theory T.

And more importantly, the interest of the theorem depends on the still-informal notion of a sufficiently strong theory being in good order. Theorem 2

[2] For the grandfather of all diagonalization arguments, due to Georg Cantor, see http://en.wikipedia.org/wiki/Cantor's_diagonal_argument (as well as *IGT2*, §2.5).

[3] I learnt the argument in this chapter as a student – so decades ago! – from lectures by Timothy Smiley.

claimed incompleteness on the assumption that T can prove a certain as-yet-unspecified amount of arithmetic. Our new Theorem 8 claims incompleteness on the more specific basis that, for any decidable property of numbers, T can case-by-case determine which numbers have the property. Now, I wouldn't have written up the argument in this chapter if this notion of T's being 'sufficiently strong' were intrinsically problematic. Still, we are left with a project here: we will want to give a sharper account of what makes for an effectively decidable property in order to (i) clarify the notion of sufficient strength, while (ii) still making it plausible that we want sufficiently strong theories in this clarified sense.

That can indeed be done, and it turns out that a surprisingly weak theory called Robinson Arithmetic which we meet in the next chapter is already sufficiently strong. However, supplying and defending the needed sharp account of the notion of effective decidability in order to pin down the notion of sufficient strength takes some effort! And it arguably takes at least as much effort compared with the task of filling in the needed details for proving incompleteness by Gödel's original method as partially sketched in Chapter 3. So over the next chapters, we are going to revert to exploring something closer to Gödel's route to the incompleteness theorems.

Still, our argument in this present chapter is highly suggestive and well worth knowing about.