Cambridge University Press
978-1-107-02284-3 - An Introduction to Gödel's Theorems: Second Edition
Peter Smith
Excerpt
More information

1 What Gödel's Theorems say

1.1 Basic arithmetic

It is child's play to grasp the fundamental notions involved in the arithmetic of addition and multiplication. Starting from zero, there is a sequence of 'counting' numbers, each having exactly one immediate successor. This sequence of numbers – officially, *the natural numbers* – continues without end, never circling back on itself; and there are no 'stray' natural numbers, lurking outside this sequence. Adding n to m is the operation of starting from m in the number sequence and moving n places along. Multiplying m by n is the operation of (starting from zero and) repeatedly adding m, n times. It's as simple as that.

Once these fundamental notions are in place, we can readily define many more arithmetical concepts in terms of them. Thus, for any natural numbers m and n, $m < n$ iff there is a number $k \neq 0$ such that $m + k = n$. m is a factor of n iff $0 < m$ and there is some number k such that $0 < k$ and $m \times k = n$. m is even iff it has 2 as a factor. m is prime iff $1 < m$ and m's only factors are 1 and itself. And so on.[1]

Using our basic and defined concepts, we can then frame various general claims about the arithmetic of addition and multiplication. There are obvious truths like 'addition is commutative', i.e. for any numbers m and n, $m + n = n + m$. There are also some very unobvious claims, yet to be proved, like Goldbach's conjecture that every even number greater than two is the sum of two primes.

That second example illustrates the truism that it is one thing to understand what we'll call *the language of basic arithmetic* (i.e. the language of the addition and multiplication of natural numbers, together with the standard first-order logical apparatus), and it is quite another thing to be able to evaluate claims that can be framed in that simple language.

Still, it is extremely plausible to suppose that, whether the answers are readily available to us or not, questions posed in the language of basic arithmetic do *have* entirely determinate answers. The structure of the natural number sequence, with each number having a unique successor and there being no repetitions, is (surely) simple and clear. The operations of addition and multiplication are (surely) entirely well-defined; their outcomes are fixed by the school-room rules. So what more could be needed to fix the truth or falsity of propositions that – perhaps via a chain of definitions – amount to claims of basic arithmetic?

To put it fancifully: God lays down the number sequence and specifies how the operations of addition and multiplication work. He has then done all he needs

[1] 'Iff' is, of course, the standard logicians' shorthand for 'if and only if'.

to do to make it the case e.g. that Goldbach's conjecture is true (or false, as the case may be).

Of course, that way of putting it is rather *too* fanciful for comfort. We may indeed find it compelling to think that the sequence of natural numbers has a definite structure, and that the operations of addition and multiplication are entirely nailed down by the familiar basic rules. But what is the real content of the thought that the truth-values of all basic arithmetic propositions are thereby 'fixed'?

Here's one appealing way of giving non-metaphorical content to that thought. The idea is that we can specify a bundle of fundamental assumptions or *axioms* which somehow pin down the structure of the natural number sequence, and which also characterize addition and multiplication (after all, it is pretty natural to suppose that we *can* give a reasonably simple list of true axioms to encapsulate the fundamental principles so readily grasped by the successful learner of school arithmetic). So now suppose that φ is a proposition which can be formulated in the language of basic arithmetic. Then, the appealing suggestion continues, the assumed truth of our axioms 'fixes' the truth-value of any such φ in the following sense: either φ is logically deducible from the axioms, and so φ is true; or its negation $\neg\varphi$ is deducible from the axioms, and so φ is false. We may not, of course, actually stumble on a proof one way or the other: but the proposal is that such a proof is always possible, since the axioms contain enough information to enable the truth-value of any basic arithmetical proposition to be deductively extracted by deploying familiar step-by-step logical rules of inference.

Logicians say that a theory T is *negation-complete* if, for every sentence φ in the language of the theory, either φ or $\neg\varphi$ can be derived in T's proof system. So, put into that jargon, the suggestion we are considering is this: we should be able to specify a reasonably simple bundle of true axioms which, together with some logic, give us a *negation-complete* theory of basic arithmetic – i.e. we could in principle use the theory to prove or disprove any claim which is expressible in the language of basic arithmetic. If that's right, truth in basic arithmetic could just be equated with provability in this complete theory.

It is tempting to say rather more. For what will the axioms of basic arithmetic look like? Here's one candidate: 'For every natural number, there's a unique next one'. This is evidently true; but evident *how*? As a first thought, you might say 'we can just see, using mathematical intuition, that this axiom is true'. But the idea of mathematical intuition is obscure, to say the least. Maybe, on second thoughts, we don't need to appeal to it. Perhaps the axiom is evidently true because it is some kind of definitional triviality. Perhaps it is just part of what we *mean* by talk of the natural numbers that we are dealing with an ordered sequence where each member of the sequence has a unique successor. And, plausibly, other candidate axioms are similarly true by definition.

If those tempting second thoughts are right, then true arithmetical claims are *analytic* in the philosophers' sense of the word; that is to say, the truths of basic arithmetic will all flow deductively from logic plus axioms which are trivially true-by-definition.[2] This so-called 'logicist' view would then give us a very neat explanation of the special certainty and the necessary truth of correct claims of basic arithmetic.

1.2 Incompleteness

But now, in headline terms, *Gödel's First Incompleteness Theorem shows that the entirely natural idea that we can give a complete theory of basic arithmetic with a tidy set of axioms is wrong.*

Suppose we try to specify a suitable axiomatic theory T to capture the structure of the natural number sequence and pin down addition and multiplication (and maybe a lot more besides). We want T to have a nice set of true axioms and a reliably truth-preserving deductive logic. In that case, everything T proves must be true, i.e. T is a *sound* theory. But now Gödel gives us a recipe for coming up with a corresponding sentence G_T, couched in the language of basic arithmetic, such that – assuming T really is sound – (i) we can show that G_T can't be derived in T, and yet (ii) we can recognize that G_T must be true.

This is surely quite astonishing. Somehow, it seems, the truths of basic arithmetic must elude our attempts to pin them down by giving a nice set of fundamental assumptions from which we can deduce everything else. So how does Gödel show this in his great 1931 paper which presents the Incompleteness Theorems?

Well, note how we can use numbers and numerical propositions to encode facts about all sorts of things. For a trivial example, students in the philosophy department might be numbered off in such a way that the first digit encodes information about whether a student is an undergraduate or postgraduate, the next two digits encode year of admission, and so on. Much more excitingly, Gödel notes that we can use numbers and numerical propositions to encode facts about *theories*, e.g. facts about what can be derived in a theory T.[3] And what he then did is find a general method that enabled him to take any theory T strong enough to capture a modest amount of basic arithmetic and construct a corresponding arithmetical sentence G_T which encodes the claim 'The sentence G_T itself is unprovable in theory T'. So G_T is true if and only if T can't prove it.

[2] Thus Gottlob Frege, writing in his wonderful *Grundlagen der Arithmetik*, urges us to seek the proof of a mathematical proposition by 'following it up right back to the primitive truths. If, in carrying out this process, we come only on general logical laws and on definitions, then the truth is an analytic one.' (Frege, 1884, p. 4)

[3] By the way, it is absolutely standard for logicians to talk of a theory T as *proving* a sentence φ when there is a logically correct derivation of φ from T's assumptions. But T's assumptions may be contentious or plain false or downright absurd. So, T's proving φ in this logician's sense does not mean that φ is proved in the sense that it is established as true. It is far too late in the game to kick against the logician's usage, and in most contexts it is harmless. But our special concern in this book is with the connections and contrasts between being true and being provable in this or that theory T. So we need to be on our guard. And to help emphasize that proving-in-T is not always proving-as-true, I'll often talk of 'deriving' rather than 'proving' sentences when it is the logician's notion which is in play.

Suppose then that T is sound. If T were to prove G_T, G_T would be false, and T would then prove a falsehood, which it can't do. Hence, if T is sound, G_T is unprovable in T. Which makes G_T *true*. Hence $\neg\mathsf{G}_T$ is false. And so that too can't be proved by T, because T only proves truths. In sum, still assuming T is sound, neither G_T nor its negation will be provable in T. Therefore T can't be negation-complete.

But in fact we don't need to assume that T is *sound*; we can make do with very significantly less. Gödel's official version of the First Theorem shows that T's mere *consistency* is enough to guarantee that a suitably constructed G_T is true-but-unprovable-in-T. And we only need a little more to show that $\neg\mathsf{G}_T$ is not provable either (we won't pause now to fuss about the needed extra assumption).

We said: the sentence G_T encodes the claim that that very sentence is unprovable. But doesn't this make G_T rather uncomfortably reminiscent of the Liar sentence 'This very sentence is false' (which is false if it is true, and true if it is false)? You might well wonder whether Gödel's argument doesn't lead to a cousin of the Liar paradox rather than to a theorem. But not so. As we will see, there really is nothing at all suspect or paradoxical about Gödel's First Theorem as a technical result about formal axiomatized systems (a result which in any case can be proved without appeal to 'self-referential' sentences).

'Hold on! If we locate G_T, a Gödel sentence for our favourite nicely axiomatized sound theory of arithmetic T, and can argue that G_T is true-though-unprovable-in-T, why can't we just patch things up by adding it to T as a new axiom?' Well, to be sure, if we start off with the sound theory T (from which we can't deduce G_T), and add G_T as a new axiom, we will get an expanded sound theory $U = T + \mathsf{G}_T$ from which we *can* quite trivially derive G_T. But we can now just re-apply Gödel's method to our improved theory U to find a new true-but-unprovable-in-U arithmetic sentence G_U that encodes 'I am unprovable in U'. So U again is incomplete. Thus T is not only incomplete but, in a quite crucial sense, is *incompletable*.

Let's emphasize this key point. There's nothing at all mysterious about a theory's failing to be negation-complete. Imagine the departmental administrator's 'theory' D which records some basic facts about the course selections of a group of students. The language of D, let's suppose, is very limited and can only be used to tell us about who takes what course in what room when. From the 'axioms' of D we'll be able, let's suppose, to deduce further facts – such as that Jack and Jill take a course together, and that ten people are taking the advanced logic course. But if there's currently no relevant axiom in D about their classmate Jo, we might not be able to deduce either J = 'Jo takes logic' or $\neg\mathsf{J}$ = 'Jo doesn't take logic'. In that case, D isn't yet a negation-complete story about the course selections of students.

However, that's just boring: for the 'theory' about course selection is no doubt completable (i.e. it can readily be expanded to settle every question that can be posed in its very limited language). By contrast, what gives Gödel's First Theorem its real bite is that it shows that any nicely axiomatized and sound theory of basic arithmetic must *remain* incomplete, however many new true axioms we give it.[4] (And again, we can weaken the soundness condition, and can – more or less – just require consistency for incompletability.)

1.3 More incompleteness

Incompletability does not just affect theories of basic arithmetic. Consider set theory, for example. Start with the empty set $\varnothing$. Form the set $\{\varnothing\}$ containing $\varnothing$ as its sole member. Then form the set containing the empty set we started off with plus the set we've just constructed. Keep on going, at each stage forming the set of all the sets so far constructed. We get the sequence

$$\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}, \ldots$$

This sequence has the structure of the natural numbers. We can pick out a first member (corresponding to zero); each member has one and only one successor; it never repeats. We can go on to define analogues of addition and multiplication. Now, any standard set theory allows us to define this sequence. So if we could have a negation-complete and sound axiomatized set theory, then we could, in particular, have a negation-complete theory of the fragment of set theory which provides us with an analogue of arithmetic. Adding a simple routine for translating the results for this fragment into the familiar language of basic arithmetic would then give us a complete sound theory of arithmetic. But Gödel's First Incompleteness Theorem tells us there can't be such a theory. So there cannot be a sound negation-complete set theory.

The point evidently generalizes: any sound axiomatized mathematical theory T that can define (an analogue of) the natural-number sequence and replicate enough of the basic arithmetic of addition and multiplication must be incomplete and incompletable.

1.4 Some implications?

Gödelian incompleteness immediately challenges what otherwise looks to be a really rather attractive suggestion about the status of basic arithmetic – namely the logicist idea that it all flows deductively using simple logic from a simple bunch of definitional truths that articulate the very ideas of the natural numbers, addition and multiplication.

But then, how *do* we manage somehow to latch on to the nature of the unending number sequence and the operations of addition and multiplication in a way that outstrips whatever rules and principles can be captured in definitions? At this point it can begin to seem that we must have a rule-transcending cognitive grasp of the numbers which underlies our ability to recognize certain 'Gödel

[4]What makes for being a 'nicely' axiomatized theory is the topic of Section 4.3.

sentences' as correct arithmetical propositions. And if you are tempted to think so, then you may well be further tempted to conclude that minds such as ours, capable of such rule-transcendence, can't be machines (supposing, reasonably enough, that the cognitive operations of anything properly called a machine can be fully captured by rules governing the machine's behaviour).

So already there's apparently a quick route from reflections about Gödel's First Theorem to some conclusions about the nature of arithmetical truth and the nature of the minds that grasp it. Whether those conclusions really follow will emerge later. For the moment, we have an initial idea of what the First Theorem says and why it might matter – enough, I hope, already to entice you to delve further into the story that unfolds in this book.

1.5 The unprovability of consistency

If we can derive even a modest amount of basic arithmetic in theory T, then we'll be able to derive $0 \neq 1$.[5] So if T *also* proves $0 = 1$, it is inconsistent. Conversely, if T is inconsistent, then – since we can derive anything in an inconsistent theory[6] – it can prove $0 = 1$. But we said that we can use numerical propositions to encode facts about what can be derived in T. So there will in particular be a *numerical* proposition Con_T that encodes the claim that we can't derive $0 = 1$ in T, i.e. encodes in a natural way the claim that T is consistent.

We know, however, that there is a numerical proposition which encodes the claim that G_T is unprovable: we have already said that it is G_T itself.

So this means that (half of) the conclusion of Gödel's official First Theorem, namely the claim that if T is consistent then G_T is unprovable, can *itself* be encoded by a numerical proposition, namely $\mathsf{Con}_T \to \mathsf{G}_T$. And now for another wonderful Gödelian insight. It turns out that the informal reasoning that we use, outside T, to show 'if T is consistent, then G_T is unprovable' is elementary enough to be mirrored by reasoning inside T (i.e. by reasoning with numerical propositions which encode facts about T-proofs). Or at least that's true so long as T satisfies conditions just a bit stronger than the official First Theorem assumes. So, again on modest assumptions, we can derive $\mathsf{Con}_T \to \mathsf{G}_T$ inside T.

But the official First Theorem has already shown that if T is consistent we can't derive G_T in T. So it immediately follows that if T is consistent it can't prove Con_T. *And that is Gödel's Second Incompleteness Theorem.* Roughly interpreted: nice theories that include enough basic arithmetic can't prove their own consistency.[7]

[5] We'll allow ourselves to abbreviate expressions of the form $\neg\sigma = \tau$ as $\sigma \neq \tau$.

[6] There are, to be sure, deviant non-classical logics in which this principle doesn't hold. In this book, however, we aren't going to say much more about them, if only because of considerations of space.

[7] That *is* rough. The Second Theorem shows that, if T is consistent, T can't prove Con_T, which is certainly *one* natural way of expressing T's consistency inside T. But couldn't there be some *other* sentence, Con'_T, which also in some good sense expresses T's consistency, where

1.6 More implications?

Suppose that there's a genuine issue about whether T is consistent. Then even before we'd ever heard of Gödel's Second Theorem, we wouldn't have been convinced of its consistency by a derivation of Con_T inside T. For we'd just note that if T were in fact inconsistent, we'd be able to derive any T-sentence we like in the theory – including a false statement of its own consistency!

The Second Theorem now shows that we would indeed be right not to trust a theory's announcement of its own consistency. For (assuming T includes enough arithmetic), if T entails Con_T, then the theory must in fact be *in*consistent.

However, the real impact of the Second Theorem isn't in the limitations it places on a theory's proving its own consistency. The key point is this. If a nice arithmetical theory T can't even prove *itself* to be consistent, it certainly can't prove that a *richer* theory T^+ is consistent (since if the richer theory is consistent, then any cut-down part of it is consistent). Hence we can't use 'safe' reasoning of the kind we can encode in ordinary arithmetic to prove that other more 'risky' mathematical theories are in good shape. For example, we can't use unproblematic arithmetical reasoning to convince ourselves of the consistency of set theory (with its postulation of a universe of wildly infinite sets).

And *that* is a very interesting result, for it seems to sabotage what is called Hilbert's Programme, which is precisely the project of trying to defend the wilder reaches of infinitistic mathematics by giving consistency proofs which use only 'safe' methods. A great deal more about this in due course.

1.7 What's next?

What we've said so far, of course, has been extremely sketchy and introductory. We must now start to do better. After preliminaries in Chapter 2 (including our first example of a 'diagonalization' argument), we go on in Chapter 3 to introduce the notions of effective computability, decidability and enumerability, notions we are going to need in what follows. Then in Chapter 4, we explain more carefully what we mean by talking about an 'axiomatized theory' and prove some elementary results about axiomatized theories in general. In Chapter 5, we introduce some concepts relating specifically to axiomatized theories of arithmetic. Then in Chapters 6 and 7 we prove a pair of neat and relatively easy results – first that any sound and 'sufficiently expressive' axiomatized theory of arithmetic is negation incomplete, and then similarly for any consistent and 'sufficiently strong' axiomatized theory. For reasons that we will explain in Chapter 8, these informal results fall some way short of Gödel's own First Incompleteness Theorem. But they do provide a very nice introduction to some key ideas that we'll be developing more formally in the ensuing chapters.

T *does* prove Con'_T (and we avoid trouble because T doesn't prove $\mathsf{Con}'_T \to \mathsf{G}_T$)? We'll return to this question in Sections 31.6 and 36.1.

2 Functions and enumerations

We start by fixing some entirely standard notation and terminology for talking about functions (worth knowing anyway, quite apart from the occasional use we make of it in coming chapters). We next introduce the useful little notion of a 'characteristic function'. Then we explain the idea of enumerability and give our first example of a 'diagonalization argument' – an absolutely crucial type of argument which will feature repeatedly in this book.

2.1 Kinds of function

(a) Functions, and in particular functions from natural numbers to natural numbers, will feature pivotally in everything that follows.

Note though that our concern will be with *total* functions. A total one-place function maps each and every element of its *domain* to some unique corresponding value in its *codomain*. Similarly for many-place functions: for example, the total two-place addition function maps any two numbers to their unique sum.

For certain wider mathematical purposes, especially in the broader theory of computation, the more general idea of a *partial* function can take centre stage. This is a mapping f which does not necessarily have an output for each argument in its domain (for a simple example, consider the function mapping a natural number to its natural number square root, if it has one). However, we won't need to say much about partial functions in this book, and hence – by default – plain 'function' will henceforth always mean 'total function'.

(b) The conventional notation to indicate that the one-place total function f maps elements of the domain Δ to values in the codomain Γ is, of course, $f\colon \Delta \to \Gamma$. Let f be such a function. Then we say

1. The *range* of f is $\{f(x) \mid x \in \Delta\}$, i.e. the set of elements in Γ that are values of f for arguments in Δ. Note, the range of a function need not be the whole codomain.

2. f is *surjective* iff the range of f indeed *is* the whole codomain Γ – i.e. just if for every $y \in \Gamma$ there is some $x \in \Delta$ such that $f(x) = y$. (If you prefer that in plainer English, you can say that such a function is *onto*, since it maps Δ onto the whole of Γ.)

3. f is *injective* iff f maps different elements of Δ to different elements of Γ – i.e. just if, whenever $x \neq y$, $f(x) \neq f(y)$. (In plainer English, you can say that such a function is *one-to-one*.)

4. f is *bijective* iff it is both surjective and injective. (Or if you prefer, f is then a *one-one correspondence* between Δ and Γ.)[1]

These definitions generalize in natural ways to many-place functions that map two or more objects to values: but we needn't pause over this.

(c) Our special concern, we said, is going to be with numerical functions. It is conventional to use 'ℕ' for the set of natural numbers (which includes zero, remember). So $f\colon \mathbb{N} \to \mathbb{N}$ is a one-place total function, defined for all natural numbers, with number values. While $c\colon \mathbb{N} \to \{0,1\}$, for example, is a one-place numerical function whose values are restricted to 0 and 1.

'$\mathbb{N}^2$' standardly denotes the set of ordered pairs of numbers, so $f\colon \mathbb{N} \to \mathbb{N}^2$ is a one-place function that maps numbers to ordered pairs of numbers. Note, $g\colon \mathbb{N}^2 \to \mathbb{N}$ is another *one*-place function which this time maps an ordered pair (which is one thing!) to a number. So, to be really pernickety, if we want to indicate a function like addition which maps *two* numbers to a number, we really need a notation such as $h\colon \mathbb{N}, \mathbb{N} \to \mathbb{N}$.

2.2 Characteristic functions

As well as talking about numerical *functions*, we will also be talking a lot about numerical *properties* and *relations*. But discussion of these can be tied back to discussion of functions using the following idea:

> The *characteristic function* of the numerical property P is the one-place function $c_P\colon \mathbb{N} \to \{0,1\}$ such that if n is P, then $c_P(n) = 0$, and if n isn't P, then $c_P(n) = 1$. (So if P is the property of being even, then c_P maps even numbers to 0 and odd numbers to 1.)
>
> The characteristic function of the two-place numerical relation R is the two-place function $c_R\colon \mathbb{N}, \mathbb{N} \to \{0,1\}$ such that if m is R to n, then $c_R(m,n) = 0$, and if m isn't R to n, then $c_R(m,n) = 1$.

The notion evidently generalizes to many-place relations in the obvious way.

The choice of values for the characteristic function is, of course, pretty arbitrary; any pair of distinct objects would do as the set of values. Our choice is supposed to be reminiscent of the familiar use of 0 and 1, one way round or the other, to stand in for *true* and *false*. And our selection of 0 rather than 1 for *true* – not the usual choice, but it was Gödel's – is merely for later neatness.

Now, the numerical property P partitions the numbers into two sets, the set of numbers that have the property and the set of numbers that don't. Its corresponding characteristic function c_P also partitions the numbers into two sets, the set of numbers the function maps to the value 0, and the set of numbers the function maps to the value 1. And these are of course exactly the *same*

[1] If these notions really *are* new to you, it will help to look at the on-line exercises.

partition both times. So in a good sense, *P and its characteristic function c_P encapsulate just the same information about a partition.* That's why we can typically move between talk of a property and talk of its characteristic function without loss of relevant information. Similarly, of course, for relations.

We will be making use of characteristic functions a lot, starting in the next chapter. But the rest of this chapter discusses something else, namely . . .

2.3 Enumerable sets

Suppose that Σ is some set of items: its members might be natural numbers, computer programs, infinite binary strings, complex numbers or whatever. Then, as a suggestive first shot, we can say

> The set Σ is *enumerable* iff its members can – at least in principle – be listed off in some numerical order (a zero-th, first, second, . . .) with every member appearing on the list; repetitions are allowed, and the list may be infinite.

It is tidiest to think of the empty set as the limiting case of an enumerable set; after all, it is enumerated by the empty list.

One issue with this rough definition is that, if we are literally to 'list off' elements of Σ, then we need to be dealing with elements which are either things that themselves can be written down (like finite strings of symbols), or which at least have standard representations that can be written down (in the way that natural numbers have numerals which denote them). That condition will be satisfied in most of the cases that interest us in this book; but we need the idea of enumerability to apply more widely.

A more immediate problem is that it is of course careless to talk about 'listing off' *infinite* sets as if we can complete the job. What we really mean is that any member of Σ will eventually appear on this list, if we go on long enough.

Let's give a more rigorous definition, then, that doesn't presuppose that we have a way of writing down the members of Σ, and doesn't imagine us actually making a list. So officially we will now say

> The set Σ is enumerable iff either Σ is empty or else there is a surjective function $f\colon \mathbb{N} \to \Sigma$ (so Σ is the range of f: we can say that such a function enumerates Σ).

This is equivalent to our original informal definition (at least in the cases we are most interested in, when it makes sense to talk of listing the members of Σ).

Proof Both definitions trivially cover the case where Σ is empty. So concentrate on the non-empty cases.

Pretend we can list off all the members of Σ in some order, repetitions allowed. Count off the members of the list from zero, and define the function f as follows: