EXPOUNDING THE FIRST INCOMPLETENESS THEOREM

PETER SMITH

Contents

Preface

When initially planning and then actually writing my *Introduction to Gödel's Theorems* (CUP, first published 2007), I consulted other books rather little, preferring to reconstruct strategies and proofs from memory as far as I could. I thought that this would be a good discipline, and that rethinking things through would help me to explain things as clearly as possible. But now that I am preparing a second edition, I want to pause to review how others have handled the First Incompleteness Theorem, both in the early papers from Gödel on, and then in the later textbook tradition. How do is the Theorem stated? How is it proved?

Here then are some extensive notes on the expository tradition. They don't at all aim to be comprehensive, though I'd like to know about significant omissions. The notes have been written, as much as anything, as a rather detailed aide-memoire for myself. I have done some joining up of the dots to make them tolerably readable, but I haven't put in the time to spell out everything out in the way a beginning student might want. Still, you shouldn't need much background to follow the twists and turns. Make what use of these notes that you will![1]

The notes come in three parts. Part 1 looks at early papers by the Founding Fathers. Part 2 looks at three pivotal works, Mostowki's *Sentences Undecidable in Formalized Arithmetic* and Kleene's *Introduction to Metamathematics* (both from 1952), and then Tarski, Mostowski and Robinson's *Undecidable Theories* (1953). Part III continues the story on through some sixty years of textbooks.

Part 1. From Gödel 1931 to Kleene 1943

1. Notation and terminology

Notation varies between authors, and (potentially more confusingly) so does some absolutely basic terminology. I'm cheerfully going to impose a lot more uniformity, and not always indicate what an author's own preferred usage is.

Notation Following a not-uncommon convention, I use different typefaces – *italics* vs sans serif – to distinguish informal mathematics from expressions of some formal language. Thus 'Prf' indicates a certain number theoretic relation, while '$\mathsf{Prf}(\mathsf{x},\mathsf{y})$' indicates the formal wff that expresses it in some given formal language.

As usual, we use '$\overline{\mathsf{m}}$' for the relevant formal language's canonical numeral for m: normally, that is m occurrences of the expression 'S' or whatever else is used for the successor function, followed by '$\mathsf{0}$'.

Given a system of Gödel numbering, '$\ulcorner\varphi\urcorner$' denotes the Gödel number of φ, and '$\overline{\ulcorner\varphi\urcorner}$' will then be the formal numeral for that number.

Henceforth, $Prf(m,n)$ stands for: m Gödel-numbers a proof of the wff with Gödel-number n (i.e. a proof in the relevant theory T, with some assumed system of numbering in play).

Terminology: expressing vs capturing We'll say

(1) A two-place formal predicate $\varphi(\mathsf{x},\mathsf{y})$ *expresses* the numerical relation R in the interpreted formal language L just in case '$\varphi(\overline{\mathsf{m}},\overline{\mathsf{n}})$' is true iff Rmn.

(2) A two-place formal predicate $\varphi(\mathsf{x},\mathsf{y})$ *captures* the numerical relation R in the formal theory T just in case

[1] Caveat lector! Remember how very easy it to LaTeX your work: just because what you write then looks very pretty doesn't mean that it is any more authoritative ...

(a) if Rmn, then $T \vdash \varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$,
(b) if not-Rmn, then $T \vdash \neg\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$.

So one notion concerns the expressive strength of a *language*, the other notion concerns the deductive strength of a *theory*. Dangerously, 'expresses' has – by different authors – been used for both: likewise for 'defines'. Latterly, 'represents' has become fairly standard for the second notion: but I still prefer to use 'capture' as it is helpfully mnemonic for case-by-case prove. Obviously, if T is sound, if a wff captures R it expresses it (but the converse doesn't hold).

Note that 'expressing', in our officially defined sense, is a weak condition; for if $\varphi(\mathsf{x}, \mathsf{y})$ expresses R in L so does $\varphi(\mathsf{x}, \mathsf{y}) \wedge \theta$ for any free-riding L-truth θ. Likewise 'capturing', in our officially defined sense, is a weak condition; for if $\varphi(\mathsf{x}, \mathsf{y})$ captures R in T so does $\varphi(\mathsf{x}, \mathsf{y}) \wedge \theta$ for any free-riding T-theorem θ. Many early writers when they talk of expressing/capturing have in mind wffs for doing the job which are constructed in some canonical way which won't produce arbitrary free-riding add-ons: but they aren't always clear about this.

The generalization of our definitions to cover monadic properties and relations of different adicies is obvious and needn't delay us. The natural initial expansion of the terminology to cover expressing/capturing functions is this:

(3) A two-place formal predicate $\varphi(\mathsf{x}, \mathsf{y})$ *expresses* the one-place numerical function f in the interpreted formal language L just in case '$\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$' is true iff $f(m) = n$.

(4) A two-place formal predicate $\varphi(\mathsf{x}, \mathsf{y})$ *captures* the one-place numerical function f in the formal theory T just in case
   (a) if $f(m) = n$, then $T \vdash \varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$,
   (b) if $f(m) \neq n$, then $T \vdash \neg\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$.

But, as we'll see later, we might want stronger conditions for 'capturing' a function using a relational expression, which adds the requirement that the relation is provably a functional one. We'll return to this point. But what if T's formal language allows for built-in or defined functional expressions (beyond perhaps successor, addition and multiplication)? Then the natural thing to say, of course, is that

(5) A one-place formal functional expression $\psi(\mathsf{x})$ *expresses* the one-place numerical function f in the interpreted formal language L just in case '$\psi(\overline{\mathsf{m}}) = \overline{\mathsf{n}}$' is true iff $f(m) = n$.

(6) A one-place formal functional expression $\psi(\mathsf{x})$ *captures* the one-place numerical function f in the formal theory T just in case
   (a) if $f(m) = n$, then $T \vdash \psi(\overline{\mathsf{m}}) = \overline{\mathsf{n}}$,
   (b) if $f(m) \neq n$, then $T \vdash \psi(\overline{\mathsf{m}}) \neq \overline{\mathsf{n}}$.

Note though that on the most modest assumptions, clause (6b) is redundant. For example, if $f(m) \neq n$ then for some k, $f(m) = k$ and $k \neq n$. But then by (6a) $T \vdash \psi(\overline{\mathsf{m}}) = \overline{\mathsf{k}}$. Assuming T has normal identity laws and knows enough about numerals to give $T \vdash \overline{\mathsf{k}} \neq \overline{\mathsf{n}}$, then (6b) follows.

2. Gödel 1931: in the beginning

(A) Let's start with Gödel himself and ask ourselves again: What version(s) of the First Theorem did he actually state in his epoch-making 1931 paper? And what proof(s) did he give?

The initial thing to headline is: *there are in fact two significantly different results stated in the paper: an informally presented, motivational,* 'semantic' *incompleteness theorem and the official, much more worked-through,* 'syntactic' *incompleteness theorem.* (Gödel himself was clear about the distinction, but the paper could be more explicit, and there

are indications that some early readers fumbled this. I think the first to spell out the distinction carefully in print is Mostowski in an article in Polish in 1945, part of which is directly translated as the Introduction to his 1952 book.)

In 1931 §1, Gödel announces that 'it can be shown' that we can find a formula of *Principia Mathematica* which, relative to a given coding, *expresses* the property of being (the code number of) a provable formula. Then he uses a now familiar kind of diagonalization construction to construct a Gödel sentence G, which (as he puts it) 'says about itself that it is not provable in *PM*': more about this below. He then gives the obvious simple argument for its undecidability on the *semantic* assumption that *Principia* is sound (if G were provable, it would be false: so can't be provable in a sound theory; hence it is unprovable and therefore true, hence it's false negation is unprovable too). He adds,

> The method of proof just explained can clearly be applied to any formal system that, first, when interpreted as representing a system of notions and propositions, has at its disposal sufficient means of expression to define [i.e., in our terms, express] the notions occurring in the argument above (in particular, the notion 'provable formula') and in which, second, every provable formula is true in the interpretation considered (p. 151).

Call that generalized result the 'semantic' incompleteness theorem.

The two following sections of Gödel's paper then give a different argument, this time strengthening the claim that we are dealing with a theory which can *express* the property of being provable to the claim that we are dealing with a theory which can *capture* that property. Strengthening one requirement like this allows us to weaken the other requirement from the semantic assumption of soundness to the *syntactic* assumption of ω-consistency. (Recall: a theory T is ω-inconsistent if, for some φ, proves $\varphi(\overline{\mathsf{n}})$ for each n yet also proves $\exists \mathsf{x} \neg \varphi(\mathsf{x})$.)

Taking that more slowly, the argument in the long, action-packed, §§2–3 goes through the following stages:

(1) Gödel first defines the formal system P which he is going to discuss (instead of the mess which is ramified *Principia*): this is essentially Peano's second-order axioms plus a simple type theory (pp. 151–156).

(2) He then explains his adopted system of Gödel-numbering (p. 157).

(3) Next, he explains the notion of a primitive recursive function – that's plain 'recursive' for Gödel of course (pp. 157–163).

(4) There follows the 45-stage demonstration that the relation $Prf(m, n)$ that holds when m codes for a P-proof of the wff with Gödel code n is indeed primitive recursive.

(5) Gödel now wants a proof that *Prf* can be captured in theory P. Rather than look specifically at how we might capture *Prf* given the way that relation is built up in the 45-stage development, he goes for Theorem V which says, quite generally, that *every* recursive property or relation can be captured in P. However, Gödel only sketches a proof by induction on the complexity of the definition of the (characteristic function of the) property or relation in terms of definitions by composition and recursion grounding out in the trivial initial functions. The crucial step is just asserted – 'the processes of definition ... (substitution and recursion) can both be formally reproduced in the system P'. Given the well-known strength of second-order arithmetic (let alone P), Gödel could no doubt reasonably take this to be uncontentious, but it certainly isn't spelt out (pp. 171–173).

(6) Gödel doesn't now just prove that P is incomplete: he proves Theorem VI: *The result of adding to P any recursive class κ of additional axioms is incomplete, assuming it is ω-consistent.* This is shown by construction of a Gödel sentence (using the wff

he constructs for capturing *Prf* – so we need the result that *Prf* can be captured in P and a fortiori in $P + \kappa$). The argument then uses the syntactic assumption of ω-consistency (pp.173–177). I'll again say more about the details of the construction below.

(7) Gödel next remarks that this incompleteness result also evidently generalizes. Thus:

> In the proof of Theorem VI no properties of the system P were used besides the following:
>
> (a) The class of axioms and the rules of inference (that is, the relation 'immediate consequence') are [primitive] recursively definable (as soon as we replace the primitive signs in some way by the natural numbers).
>
> (b) Every [primitive] recursive relation is definable [i.e. is 'capturable'] in the system P.
>
> Therefore, in every formal system that satisfies the assumptions 1 and 2 and is ω-consistent, there are undecidable propositions of the form $[\forall x F(x)]$, where F is a [primitive] recursively defined property of natural numbers, and likewise in every extension of such a system by a recursively definable ω-consistent class of axioms (p. 181).

(8) But we aren't done yet. Gödel now shows that we don't need the powerful expressive resources of P to express recursive properties and relations. The first-order language of addition and multiplication suffices (Theorem VII puts it thus: 'Every recursive relation is arithmetical'). *This* is, of course, the place where the β-function trick comes into play, though it's not called that in 1931. It follows that, for any way of expressing a recursive relation in P, there is an equivalent arithmetical way. And since the reasoning for this is elementary, 'this equivalence is provable in P' (but *that* isn't actually proved). This yields Theorem VIII: '*In any of the formal systems of arithmetic mentioned in Theorem VI, there are undecidable arithmetical propositions*'.

Note that Gödel doesn't in fact label the key generalization of the 'syntactic' theorem noted in (7) above as itself separate theorem. But if anything, *it is this general version (7) combined with the claim (8) that the undecidable propositions will always include arithmetical ones, that is the core of what later tradition has called the First Incompleteness Theorem.*

It is worth remarking again that the official syntactic version of the theorem *doesn't* get a full proof in 1931. There *are*, however, all the necessary ingredients for a full proof of the *semantic* version of the incompleteness theorem for P: for Gödel does in effect tell us how to use the β-function trick to construct an arithmetical wff that expresses the primitive recursive relation *Prf* (in a reasonably natural and direct way, without free-riders) and hence how to construct an arithmetical Gödel sentence which 'says of itself' that it is unprovable in P, and hence is undecidable in P assuming its soundness. And this result evidently generalizes to any theory whose axioms and rules of inference are primitive recursively definable and which can express every primitive recursive relation. But to complete a full proof of the *syntactic* version for P we'd need to complete the proof that, for every recursive relation, there is an arithmetical wff which captures it, which the 1931 paper doesn't give.

Here's another lacuna which later work will fill in: the generalized theorems apply to theories which can express/capture any primitive recursive relation. But how powerful must such a theory be (if not the full-blown theory P or some extension thereof)? Gödel's 1931 paper doesn't investigate this.

(B) To return to the sketched construction in Gödel's §1 of a sentence which 'says about itself that it is not provable in *PM*'.

Here's the argument, sticking close to Gödel's notation:

(1) The 'class signs' of *PM*, i.e. wffs with one free numerical variable, can be listed off, with φ_n the n-th one.

(2) Let $\varphi(\overline{\mathsf{n}})$ indicate the result of substituting the numeral for n for the free numerical variable, if there is one, in the wff φ.

(3) Now consider the property n has if $\varphi_n(\overline{\mathsf{n}})$ is not provable in *PM*.

(4) This numerical property can be expressed in *PM* by some 'class sign', which will one of the φ_n, say φ_q.

(5) So we have $\varphi_q(\overline{\mathsf{n}})$ is true iff $\varphi_n(\overline{\mathsf{n}})$ isn't provable.

(6) Whence, in particular, $\varphi_q(\overline{\mathsf{q}})$ is true iff $\varphi_q(\overline{\mathsf{q}})$ isn't provable..

Note this argument can be explained at the present level of abstraction without reference to Gödel numbering (we just used a numerical ordering on the class signs).

So everything now depends on the expressibility assumption (4). This should look quite plausible given the advertised strength of *PM*, but the rest of the 1931 shows how to prove such an assumption.

(C) What about the construction of the Gödel sentence and the syntactic proof of its undecidability in Gödel's §2 (pp. 174–175)?

We'll take the base case where we are proving the incompleteness of P rather than of $P + \kappa$ for some additional axioms κ (the added complication is not important here). Then the argument, with the wraps off, goes like this:

(1) Consider the relation Q [Gödel's notation] defined as follows: $Q(m, n)$ holds iff m *doesn't* number the P-proof of the result of substituting the numeral for n for the free occurrences of the variable y in the wff with code-number n.

(2) The relation *Prf* is primitive recursive, as is the function *sub* where $sub(m, n)$ returns the Gödel number of the result of substituting the numeral for n in any free occurrences of y in the wff with Gödel number m. But $Q(m, n)$ can then be defined as not-$Prf(m, sub(n, n))$, which will therefore also be primitive recursive.

(3) Since Q is primitive recursive, it can be captured by a wff $\mathsf{Q}(\mathsf{x}, \mathsf{y})$ [that can be $\neg\mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\mathsf{y}, \mathsf{y}))$, where $\mathsf{sub}(\mathsf{x}, \mathsf{y})$ is a functional expression capturing *sub*.]

(4) Form the generalization $\forall\mathsf{x}\mathsf{Q}(\mathsf{x}, \mathsf{y})$, and suppose the Gödel number of this is p [again Gödel's notation].

(5) Now consider the sentence $\forall\mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$, which is the result of substituting the numeral for p for the free occurrences of the variable y in the wff with code-number p. [Note: assuming P is sound, $\forall\mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{n}})$ will be true iff no number m numbers the P-proof of the result of substituting the numeral for n for the free occurrences of the variable y in the wff with code-number n: so this expresses the analogue of the property n has if $[R_n; n]$ is not provable which we met in §1. Hence $\forall\mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$ is true iff no number numbers of a proof of $\forall\mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$. So $\forall\mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$ is true iff it is unprovable.]

(6) And off we go, in the now familiar way. Suppose for reductio $P \vdash \forall\mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$. Then some number m must number a P-proof of the result of substituting the numeral for p for the free occurrences of the variable y in the wff with code-number p. Whence (by definition) not-$Q(m, p)$. But Q captures Q in P, so that implies $P \vdash \neg\mathsf{Q}(\overline{\mathsf{m}}, \overline{\mathsf{p}})$. Which contradicts P's consistency (and a fortiori its ω-inconsistency).

(7) We've just shown that $\forall \mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$ is not P-provable. That is to say, for every number m, m *doesn't* number the P-proof of the result of substituting the numeral for p for the free occurrences of the variable y in the wff with code-number p. That is to say, for every m, $Q(m, p)$. But Q captures Q, so that implies that for every m $P \vdash \mathsf{Q}(\overline{\mathsf{m}}, \overline{\mathsf{p}})$. Now suppose $P \vdash \neg\forall \mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$. Then that would contradict P's ω-inconsistency.

However, as I say, that is Gödel 'with the wraps off'. His own presentation, at first sight, seems much less approachable. That's because he opts to talk in coding mode, rather than about expressions directly (thus when he talks e.g. of FORMULAS, Gödel is actually talking of a class of numbers, and being PROVABLE is a property of numbers, etc.). So, in particular,

(i) Instead of talking about the *wff* $\forall \mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$ he talks of the *number* $17Gen\ r$. That's because our wff is also the result of taking the wff $\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$ (to which Gödel gives the number r), and then universally generalizing on x (the variable with Gödel number 17). Since the function x Gen y is defined as yielding the number of the result of generalizing the wff number y with respect to the variable numbered x, 17 Gen r is the number for our Gödel sentence.

(ii) Gödel defines the numerical property *Bew* (for 'BEWEISBARE FORMEL', i.e. [numbers a] provable formula) thus: $Bew(n)$ iff $\exists x Prf(x, n)$. And then, instead of saying e.g. that $\forall \mathsf{x}\mathsf{Q}(\mathsf{x}, \overline{\mathsf{p}})$ is provable he can write, and prefers to write, the coded claim $Bew(17Gen\ r)$, and so forth.

Still, if we take the coding wraps off, Gödel's construction and proof is the familiar one I sketched above.

3. TARKSI 1933: TRUTH (BUT NOT PROOF)

(A) Along with John von Neumann, two of those who quickly saw the importance of Gödel's result were Rudolf Carnap (about whom more later) and Alfred Tarski.

Tarski's monograph on 'The Concept of Truth in the Languages of the Deductive Sciences' (Polish version 1933, English translation 1936) was composed in 1931: as it was going to press, he added the proof of the inexpressibility of truth in languages including enough arithmetic, based on the newly available Gödelian ideas. So Tarksi is the first logician to publish a relevant major work in the wake of Gödel's result.

Here's his familiar argument about truth in rather more modern dress.

(1) We are considering a language L which can express enough arithmetic (it's a language for type theory in Tarski's example). And we suppose for reductio $\mathsf{Tr}(\mathsf{x})$ is a one-place predicate expressing truth in L in the sense that $\mathsf{Tr}(\overline{\ulcorner \varphi \urcorner}) \leftrightarrow \varphi$ is always true in L. (We could run the following argument mutatis mutandis for some alternative system of systematic structural description instead of Gödel numbering, but we'll stick to what we know.)

(2) Enumerate L's open wffs with the free variable x as $\varphi_n(\mathsf{x})$, and consider now the arithmetical property D defined by Dn iff $\neg\mathsf{Tr}(\overline{\ulcorner \varphi_n(\overline{\mathsf{n}}) \urcorner})$ is true in L. This is expressible in L [in fact, though Tarski doesn't spell this out, by $\neg\mathsf{Tr}(\mathsf{sub}(\overline{\mathsf{n}}, \overline{\mathsf{n}}))$, for suitable sub expressing the function $sub(m, n)$ that takes the m-th wff φ_m and returns the Gödel number of the code for the result of substituting the numeral for n for the free variable in φ_m].

(3) The wff expressing D in L must be φ_k for some k. So this is true in L for any n:

$$\varphi_k(\overline{\mathsf{n}}) \leftrightarrow \neg\mathsf{Tr}(\overline{\ulcorner \varphi_n(\overline{\mathsf{n}}) \urcorner})$$

whence, of course,

$$\varphi_k(\overline{\mathsf{k}}) \leftrightarrow \neg\mathsf{Tr}(\overline{\ulcorner \varphi_k(\overline{\mathsf{k}}) \urcorner}).$$

But by the assumption that Tr expresses truth,

$$\varphi_k(\overline{\mathsf{k}}) \leftrightarrow \mathsf{Tr}(\overline{\ulcorner \varphi_k(\overline{\mathsf{k}}) \urcorner})$$

and contradiction follows.

(B) Three points about this. First, the real parallel here is with Gödel's initial *semantic* argument for incompleteness. The argument turns on assumptions about what can be *expressed* in L.

Second Dawson (1984) suggests that Gödel's correspondence with Bernays in 1931 'furnishes independent evidence of Gödel's awareness of the formal undefinability of the notion of truth – a fact nowhere mentioned in [Gödel's published paper].'

Third, it is worth remarking on what Tarski *doesn't* go on to say. Suppose we have a (primitively recursively) axiomatized formal theory T couched in some language L which can express enough arithmetic. Then truth in L isn't provability in T because while T-provability *is* expressible in L as Gödel shows in 1931, truth-in-L *isn't*. So assuming that T is sound and everything provable in it is true, this means that there must be truths of T which it can't prove.

And we might well take this to be in a sense the Master Argument for incompleteness, revealing the roots of the phenomenon. Gödel himself wrote (in response to a query)

> I think the theorem of mine that von Neumann refers to is ... that a complete epistemological description of a language A cannot be given in the same language A, because the concept of truth of sentences in A cannot be defined in A. It is this theorem which is the true reason for the existence of undecidable propositions in the formal systems containing arithmetic. I did not, however, formulate it explicitly in my paper of 1931 but only in my Princeton lectures of 1934. The same theorem was proved by Tarski in his paper on the concept of truth.

Gödel's letter is quoted by Feferman (1984), who also has a very interesting discussion of why Gödel chose not to highlight this line of argument for incompleteness in his original paper. But it is slightly odd that *Tarski* didn't highlight the argument either, to point up his moral that truth is one thing (and despite positivist worries, can be given a sharp formal characterization) and proof is something other.

4. Gödel 1934: the Princeton lectures

Gödel lectured on his incompleteness results at the Institute for Advanced Study in 1934. Notes were taken by Kleene and Rosser, approved with corrections by Gödel, and quite widely circulated: a further corrected version became available in Davis (1965). There are nine sections if we set aside the postscriptum added thirty years after the event for the reprint. Here's a quick review of what happens:

(1) A short introduction characterizes the notion of a 'formal mathematical system', emphasizing that we require syntactic properties like being a well-formed derivation to be decidable.

(2) The next sections first define the primitive recursive functions, ...

(3) ... then present a particular formal system for an axiomatized second-order arithmetic (so unlike 1931's system P, the system here uses only the first two types up the hierarchy). A quite significant addition is that in the 1934 system there's a primitive description operator, or more exactly its a 'least number' operator ϵ, where $\epsilon x \varphi(x)$ denotes the least number that satisfies φ if there is such a number, or denotes zero otherwise.

(4) Gödel now defines a system of Gödel numbering, and again shows that the *Prf* relation for the new system is primitive recursive.

(5) This pivotal section has two parts. The first sketches a proof that 'every recursive function, class, and relation is represented [i.e. captured] by some formula of our formal system'. This is a little more detailed that the discussion in 1931, and things go a little easier as Gödel can make use of the ϵ-operator. Thus, consider e.g. the definition by recursion of the factorial function:

$fact(0) = 1$
$fact(Sn) = Sn \cdot fact(n)$

This can now be wrapped up into a single definition:

$fact(n) = \epsilon x \exists f\{f(0) = 1 \wedge f(Sn) = Sn \cdot f(n) \wedge f(n) = x\}$.

The trick obviously generalizes, to give us a way of using second-order quantification and a description operator to turn recursive definitions into explicit definitions. We can evidently use this trick to come up with a wff of the system which expresses any given primitive recursive function (which has to be definable by a sequence of compositions and recursions). So that shows how why any primitive recursive function can be *expressed* in the described formal system. However, Gödel really wants a proof that this trick also gives us a wff that *captures* the relevant p.r. function. But here Gödel again just says the proof 'is too long to be given here' (p. 359).

The second part of §5 proves the syntactic first incompleteness theorem for the given formal system, pretty much as in 1931: see below for more on this. Gödel then sketches a proof of the Second Theorem.

(6) Gödel now generalizes by analysing the conditions for the proofs so far to go through.

(7) The next section briefly notes Carnap's generalization of the construction of the Gödel sentence to the claim that for any wff $\varphi(\mathsf{x})$ (not just the wff which expresses the property of being not-PROVABLE) we can find a sentence γ which is true iff $\varphi(\overline{\ulcorner\gamma\urcorner})$. Gödel then derives the inexpressibility of truth by the same line of argument we noted in Tarski's paper – so this is the passage referred to in that letter to von Neumann quoted above. Gödel does explicitly draw the moral:

> So we see that the class α of numbers of true formulas cannot be expressed by a propositional function of our system whereas the class β of provable formulas can. Hence $\alpha \neq \beta$ and if we assume $\beta \subseteq \alpha$ (i.e., every provable formula is true) we have $\beta \subset \alpha$, i.e., there is a proposition A which is true but not provable. A then is not true and therefore not provable either, i.e., A is undecidable. (p. 363)

(8) We now get a clearer presentation of the β-function trick for expressing recursive functions using arithmetical predicates (and it is now shown that this implies a theorem about the undecidability of a certain statement about the solution of a Diophantine equation).

(9) Finally there is a quick look at the idea of Herbrand-Gödel computability.

In summary, though, there is perhaps surprisingly little development of the incompleteness theorems themselves in the three years between the original paper and the Princeton lectures. That's perhaps why most of Kleene's introductory note to Gödel's lectures in the *Collected Works* concerns post-1934 developments.

(B) In his §5, Gödel (p. 360, fn. 22) credits Herbrand for the lines of a somewhat shorter proof of incompleteness than the 1931 version.

(1) Start with the relation *Prf* which is such that $Prf(m, n)$ holds when m numbers a proof in T (the system under consideration) of the formula with number n. This relation is primitive recursive, and hence can be captured in T by a wff $\mathsf{Prf}(\mathsf{x}, \mathsf{y})$.

(2) Consider next the two-place function $sub(m, n)$ which returns the Gödel number of the result of substituting the numeral for n for the free occurrences of the variable y in the wff with code-number m. This function is primitive recursive, and so can be captured in T by a functional expression $\mathsf{sub}(\mathsf{x}, \mathsf{y})$.

(3) Now take the formula $\mathsf{U}(\mathsf{y}) =_{\text{def}} \forall \mathsf{x} \neg \mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\mathsf{y}, \mathsf{y}))$, and let p be the Gödel number of this.

(4) Let $\mathsf{U}(\overline{\mathsf{p}})$ have Gödel number g (n.b. that's the code number of the original, unabbreviated wff).

(5) Note that $\mathsf{U}(\overline{\mathsf{p}})$ is the result of substituting the numeral p for the free occurrences of the variable y in the wff with code-number p. So $sub(p, p) = g$. Whence, since sub captures *sub*, $T \vdash \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}) = \overline{\mathsf{g}}$.

(6) Suppose $T \vdash \mathsf{U}(\overline{\mathsf{p}})$, i.e. $T \vdash \forall \mathsf{x} \neg \mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$. Then for some m, m numbers a proof of the wff with Gödel number g, i.e. $Prf(m, g)$. Hence since Prf captures *Prf*, $T \vdash \mathsf{Prf}(\overline{\mathsf{m}}, \overline{\mathsf{g}})$. Whence $T \vdash \mathsf{Prf}(\overline{\mathsf{m}}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$, making T inconsistent.

(7) We've just shown that $\mathsf{U}(\overline{\mathsf{p}})$ is not T-provable. That is to say, for every number m, m *doesn't* number the T-proof of the result of substituting the numeral for p for the free occurrences of the variable y in the wff with code-number p. Hence, for every m, not-$Prf(m, sub(p, p))$. But Prf captures *Prf*, so that implies that for every m, $T \vdash \neg \mathsf{Prf}(\overline{\mathsf{m}}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$. Now suppose $T \vdash \neg \mathsf{U}(\overline{\mathsf{p}})$, i.e. $T \vdash \exists \mathsf{x} \mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$. Then that would contradict T's ω-inconsistency.

Now on the face of it, as I've presented things, this isn't really much slicker or shorter than the original proof as I presented *that*. But in fact, a comparison of the originals of the 1931 and 1934 papers will indeed show that the presentation of the argument in the latter *is* quite a lot nicer. However, this hasn't really anything to do with the slight difference in the way the pieces of the arguments fit together. The difference in the accessibility of the two proofs is that – as I said – the original prefers to talk of coding-numbers rather than formulas, while the later lectures talk more directly about the syntactic gadgets that are, after all, the theorem's topic. It's *that*, surely, which makes things now seem to go more easily.

(C) A footnote. Though the 1934 presentation of the incompleteness proof is more accessible than the 1931 version, Gödel still misses a major presentational trick.

Abbreviate still further, and put G for the Gödel sentence $\forall \mathsf{x} \neg \mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$. Note that since g and $\ulcorner \mathsf{G} \urcorner$ are the same number, we have $T \vdash \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}) = \overline{\ulcorner \mathsf{G} \urcorner}$. So, given that (non-trivial) claim about T's capturing powers, we can (trivially) infer

$$T \vdash \mathsf{G} \leftrightarrow \forall \mathsf{x} \neg \mathsf{Prf}(\mathsf{x}, \overline{\ulcorner \mathsf{G} \urcorner}).$$

Now put $\mathsf{Prov}(\mathsf{y}) =_{\text{def}} \exists \mathsf{x}\, \mathsf{Prf}(\mathsf{x}, \mathsf{y})$ (so some number satisfies $\mathsf{Prov}(\mathsf{y})$ iff it numbers a theorem). And we have, even more neatly,

$$T \vdash \mathsf{G} \leftrightarrow \neg \mathsf{Prov}(\overline{\ulcorner \mathsf{G} \urcorner}).$$

That is elegant and revealing: so it is a pity (and a little odd?) that Gödel didn't make the point.

Moreover, that there is a sentence G which is – in the illustrated sense – a 'fixed point' for the predicate $\neg \mathsf{Prov}(\mathsf{y})$ is a consequence of the more general, and now very familiar, Diagonalization Lemma, which says that in the setting of the right kind of theory T,

then for any one-place predicate φ there is a 'fixed point' γ such that $T \vdash \gamma \leftrightarrow \varphi(\overline{\ulcorner\gamma\urcorner})$. But Gödel doesn't remark on this here.

5. Carnap 1934: does he prove the diagonal lemma?

'It was Carnap who introduced Kurt Gödel to logic, in the serious sense,' as Goldfarb (2005, p. 185) puts it. And Carnap's 1934 *Logische Syntax der Sprache* is in turn the first extended response to the impact of Gödelian incompleteness on the logicist project.

Now, modern treatments of the First Theorem go via the general Diagonalization Lemma which we just noted, and that lemma is often credited to Carnap 1934. But is this attribution right?

(A) Look again at Gödel's construction in 1931, that we explained on p. 6 above (and we quietly generalize to any suitable theory T, relativizing *Prf* etc. to that theory).

Consider again the relation $Q(m, n)$, i.e. not-$Prf(m, sub(n, n))$. So this can be expressed, more explicitly, by $\neg\mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\mathsf{y}, \mathsf{y}))$. Gödel then constructs a sentence that 'says' it is unprovable by first quantifying on the first variable to get $\forall\mathsf{x}\neg\mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\mathsf{y}, \mathsf{y}))$, taking the Gödel number of that, p, and substituting its numeral for the second variable to get $\forall\mathsf{x}\neg\mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$.

Abbreviate $\forall\mathsf{x}\neg\mathsf{Prf}(\mathsf{x}, \mathsf{y})$ by $\varphi(\mathsf{y})$. Then

(i) our Gödel sentence, i.e. $\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$, is true just so long as the number denoted by $\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}})$ satisfies $\varphi(\mathsf{y})$.

(ii) But the number denoted by $\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}})$ is, by definition, the number of the result of putting the numeral for p for free occurrences of y in the wff with Gödel number p, i.e. the number of $\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$ itself.

(iii) So $\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$ is true just so long as $\ulcorner\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))\urcorner$ satisfies $\varphi(\mathsf{y})$.

(iv) In other words, $\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$ is true just so long as $\varphi(\overline{\ulcorner\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))\urcorner})$ is true.

(v) Or to put that more vividly, abbreviate $\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$ as γ. Then γ is true iff $\varphi(\overline{\ulcorner\gamma\urcorner})$ is true.

(vi) But $\varphi(\mathsf{y})$ evidently expresses the numerical property NOT PROVABLE. So γ is true iff γ is not provable in the relevant T.

(B) So far, that's pretty much Gödel himself in a new notational dress. *But now we note that steps (i) to (v) don't depend at all on the details of the wff* φ. So in fact we've proved something quite general: whatever open wff φ we take, there will be a wff γ [in fact, the wff $\varphi(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$ where p is the Gödel number of $\varphi(\mathsf{sub}(\mathsf{x}, \mathsf{x}))$] such that γ is true iff $\varphi(\overline{\ulcorner\gamma\urcorner})$ is true. Call this general semantic result the *Diagonal Equivalence*. And it is this equivalence that Gödel refers to in his 1934.

Now, in §35 of his 1934, Carnap neatly proves the general Diagonal Equivalence pretty much as we have just done, and in §36 he uses this, together with the assumption that NOT PROVABLE is expressible by an open wff in his Language II (better *Theory* II) to show that Language II is incomplete.

But note that constructing the semantic Diagonal Equivalence is not to establish the *Diagonalization Lemma* as normally understood in the modern sense, which is a syntactic thesis, not about truth-value equivalence but about provability. To repeat: the Diagonalization Lemma is the claim that, in the setting of the right kind of theory T, then for any one-place predicate φ there is a γ such that true $T \vdash \gamma \leftrightarrow \varphi(\overline{\ulcorner\gamma\urcorner})$. And Carnap doesn't actually state or prove that in §35.

When we turn to §36, we see that his argument for incompleteness is the simple semantic argument depending on the assumed soundness of his Language II. So at this

point Carnap is giving a version of the semantic incompleteness argument sketched in the opening section of Gödel 1931 (the one that appeals to a soundness assumption), and not a version of Gödel's official syntactic incompleteness argument which appeals to ω-consistency. Indeed, Carnap doesn't even mention ω-consistency in the context of his §36 incompleteness proof. He doesn't need to.

To put it bluntly: in §§35–36, Carnap doesn't use the theorem that his Language II proves $\gamma \leftrightarrow \varphi_p(\overline{\ulcorner \gamma \urcorner})$ for the case where φ_p expresses NOT PROVABLE; i.e. he doesn't appeal to an application of the Diagonalization Lemma in the modern sense. He doesn't need it (yet), and he doesn't prove it (here).

(C) What about later in the book? Carnap's notation and terminology together don't make for an easy read. But as far as I can see, when he returns to Gödelian matters later, he still is using the semantic Diagonal Equivalence and not the syntactic Diagonalization Lemma. If the latter was going to appear anywhere, you'd expect to find it in §60 when Carnap returns to the incompleteness of arithmetics: but it isn't there. (An indication: Carnap there talks of provability being 'definable' in arithmetics, and it is indeed *expressible* – but we know it isn't *capturable* by a trivial argument from the Diagonalization Lemma proper.[2] So Carnap hereabouts is still dealing with semantic expressibility, not the syntactic notion of capturing needed for the Lemma.)

So, in summary: Yes, Carnap notices we can generalize Gödel's construction and get the general semantic Diagonal Equivalence. But this isn't the modern Diagonalization Lemma. Which isn't to say that we can't get the Lemma very easily with the ingredients now to hand: but still, Carnap didn't actually take the step.

(D) A footnote. Carnap's 'rescue' of logicism depends on adopting a logic that allows the infinitary ω-rule [informally, from $\varphi(0), \varphi(1), \varphi(2), \varphi(3), \ldots$ you can infer $\forall n \varphi(n)$]. An inference that invokes this rule is still, he claims, analytic. A question arises, though: how many applications of the ω-rule are we allowed in a proof? ω-many? More? What differences do restrictions placed on the use of the infinitary rule make to incompleteness issues?

This question is nicely tackled in a paper by J. Barkley Rosser (1937) which shows, for example, that the consistency of the system P_n (Peano Arithmetic plus up to n uses of the ω-rule in an argument) can't be proved in that system, though it can be proved in P_{n+1}, and also there are still undecidable sentences in P_ω.

6. Kleene 1936: general recursive functions and a new proof

Gödel's 1931 paper and 1934 lectures concern the incompleteness of *primitive-recursively axiomatized* theories. For such theories the crucial *Prf* relation is primitive recursive, and hence (assuming a rich enough language) expressible and (assuming a rich enough set of axioms and rules of inference) capturable, which enables the semantic and syntactic versions of the First Theorem go through.

One thing that happens between 1931 and 1936 is that the theory of recursive functions, Turing computability, and the Lambda calculus takes off, and various definitions of effective computability are shown to be equivalent. Now, as is familiar, what we need by way of new function-building operators to extend the class of primitive recursive functions to the class of general recursive functions is the minimization (least number) operator which returns the first output of an open-ended search. But open-ended searches can evidently be expressed by the existential quantifier, and (plausibly) captured by the

[2] Suppose $\mathsf{Pr}(\mathsf{x})$ captures provability, By the Diagonalization Lemma, for some γ, $T \vdash \gamma \leftrightarrow \neg \mathsf{Pr}(\overline{\ulcorner \gamma \urcorner})$. By the definition of capturing, if $T \vdash \gamma$ then $T \vdash \mathsf{Pr}(\overline{\ulcorner \gamma \urcorner})$, and if not-$(T \vdash \gamma)$ then $T \vdash \neg \mathsf{Pr}(\overline{\ulcorner \gamma \urcorner})$. It quickly follows that T is inconsistent.

use of the quantifier too. So we should expect that normally if a language/theory can express/capture the primitive recursive functions it can express/capture general recursive functions too. Hence it should be easy to extend the First Theorem (in either the semantic or the syntactic version) to apply to *recursively* axiomatized theories as will as to primitive-recursively axiomatized theories.

But this extension is actually not very exciting. For a start, a recursively axiomatized theory can be primitive-recursively re-axiomatized using Craig's Trick, so the extended incompleteness theorems won't cover new theories in the extensional sense of axiomatizable-sets-of-theorems. Ok, the re-axiomatizations by Craig's Trick are totally unnatural; but then what would a natural presentation of a theory look like which wasn't already primitive-recursively axiomatized? Assuming a normal-looking logic, its class of axioms would have to be such that we could on occasion only check whether a wff is an axiom by an unbounded search. Which is already an entirely *un*natural way of setting up a theory!

For our purposes, the real interest in the general theory of effective computability is that it gives us new ways of proving the old First Theorem, not that it enables us to extend the application of the Theorem. One such new way is given by Stephen Kleene in his 1936. Here's a version of the argument from his §2:

(1) Any total one-place general recursive function $\varphi_e(n)$ can be identified with the function $U(\mu y[T(e, n, y) = 0])$ for a certain index e. Here, U and T are fixed primitive recursive functions. This is Kleene's Normal Form Theorem.

(2) Trivially, for given e, the defined function is indeed total iff $\forall n \exists y\, T(e, n, y) = 0$.

(3) Theorem: *the set of indices e such that $\forall n \exists y\, T(e, n, y) = 0$ is not recursively enumerable.* For suppose otherwise. Then there is a recursive function θ which enumerates the set. Then consider the function $\delta(n) = U(\mu y[T(\theta(n), n, y) = 0]) + 1$. Then this is recursive, total, but differs from any total φ_e i.e. any $\varphi_{\theta(j)}$, i.e. any $U(\mu y[T(\theta(j), n, y) = 0])$ when $n = j$. But that's impossible by the Normal Form Theorem.

(4) Now consider a sound recursively axiomatized theory Θ which can express the primitive recursive T, and imagine a standard scheme for Gödel numbering formulae. Then we can write down a formula A_e which is true just when $\forall n \exists y\, T(e, n, y) = 0$. Start effectively enumerating the proofs of Θ, and we can extract an effective enumeration of the provable A_e. But by our theorem that can't be an effective enumeration of the *true* A_e (or else that would give us an effective enumeration of the e such that $\forall n \exists y\, T(e, n, y) = 0$, since we can recover the index e from the formula A_e).

(5) So there is a formula A_e which is true but not provable in Θ. Given Θ's soundness, its negation isn't provable either. So Θ is incomplete.

That's pretty. But note (i) the proof as given by Kleene here is a *semantic* incompleteness proof, and (ii) that the undecided sentence is $\forall\exists$ so is Π_2, In §11, we'll see how Kleene later extracts a syntactic incompleteness theorem with a Π_1 undecidable sentence from a variant Normal Form Theorem.

7. Rosser 1936: strengthening the first theorem

(A) J. B. Rosser's short 'Extensions of some theorem's of Gödel and Church' (1936) is famous for the result we'll come to in (B) below, but in fact proves a lot more.

To start with, Rosser notes that *if a set can be enumerated by a total recursive function then it can enumerated by a primitive recursive function (allowing repetitions as usual).* This is intuitive. Suppose the effectively computable $\varphi(n)$ enumerates the non-empty Σ, with $s \in \Sigma$. Then consider the following derived computation, defining a function ψ.

Given input n, do n steps of a 'dovetailed' zig-zag computation through the $\varphi(j)$ – i.e. do one step of computing $\varphi(0)$; one of $\varphi(1)$ and another of $\varphi(0)$; one of $\varphi(2)$ and further steps of $\varphi(1)$, $\varphi(0)$; one of $\varphi(3)$ and further steps of $\varphi(2)$, $\varphi(1)$, $\varphi(0)$; and so on for n steps in total. If at the final step, step n, the computation of some $\varphi(m)$ finishes giving output k, put $\psi(n) = k$, otherwise put $\psi(n) = o$. Evidently $\psi(m)$ also enumerates Σ, but it is primitive recursive, as a computation of $\psi(n)$ involves just n steps, i.e. involves no unbounded searches. (Rosser officially appeals to Kleene's Normal Form theorem for a proper proof.)

Then Rosser notes the following. Suppose a theory T captures primitive recursive functions, and also the theorems of T are recursively enumerable (e.g. because the axioms are recursively enumerable). Then there is a primitive recursive function $\psi(n)$ which enumerates the Gödel numbers of the T theorems, and can be captured by a two-place predicate $\mathsf{P}(\mathsf{x},\mathsf{y})$. But then we can just rerun the Gödel construction and the incompleteness argument using $\mathsf{P}(\mathsf{x},\mathsf{y})$ rather than $\mathsf{Prf}(\mathsf{x},\mathsf{y})$. So, *the first theorem with its original line of proof will apply to a theory T which includes enough arithmetic so long as its axioms and legitimate applications of rules are recursively enumerable.*

Still, as I said in the preamble to talking about Kleene's paper, that's perhaps not *especially* interesting, especially in the light of Craig's later theorem that any T whose theorems are recursively enumerable can be re-axiomatized as a primitive-recursively axiomatized theory T', and since the original Gödel theorem will apply to the re-axiomatization T', that already shows that the equivalent T will be incomplete too.

(B) So now let's turn to what Rosser's paper is most remembered for: by constructing a fancier Gödel-Rosser sentence R, we can get a syntactic proof of incompleteness which depends only on assuming simple consistency as opposed to ω-consistency.

The construction is familiar. Instead of $\mathsf{Prf}(\mathsf{x},\mathsf{y})$ use the more complex, 'consistency-minded' predicate $\mathsf{RPrf}(\mathsf{x},\mathsf{y}) =_{def} \mathsf{Prf}(\mathsf{x},\mathsf{y}) \wedge \neg\exists\mathsf{w}[\mathsf{w} \leq \mathsf{x} \wedge \mathsf{Prf}(\mathsf{w},\mathsf{neg}(\mathsf{y}))]$, where $\mathsf{neg}(\mathsf{x})$ captures a suitable function *neg* which, fed the Gödel number of a wff returns the Gödel number of its negation (this expresses the property *RPrf* where $RPrf(m,n)$ iff m numbers a proof of the wff with number n and there is no smaller-numbered proof of the negation of that wff. Now construct a Gödel-Rosser sentence R from RPrf as the original Gödel sentence was constructed from Prf. That is to say, form $\forall\mathsf{x}\neg\mathsf{RPrf}(\mathsf{x},\mathsf{sub}(\mathsf{y},\mathsf{y}))$; take the Gödel number of that, r, and put $\mathsf{R} =_{def} \forall\mathsf{x}\neg\mathsf{RPrf}(\mathsf{x},\mathsf{sub}(\overline{\mathsf{r}},\overline{\mathsf{r}}))$, which 'says' *if I am provable, then there is a already a proof of my negation.*

It can then be shown, on the usual assumptions about what the relevant theory T can capture, that T proves neither R nor $\neg\mathsf{R}$ so long as it is plain consistent.

(C) Rosser's actual proof of this in 1936 is not fully spelt out, however. Here's a reconstruction.

First, we note that if T is consistent, then φ is provable if and only if it is Rosser-provable (i.e. φ is provable by a T-proof such that there there is no T-proof of $\neg\varphi$ with a smaller Gödel number). Then Rosser announces two lemmas which (still assuming T is consistent) come to this:

(i) if φ is provable, then $\mathsf{RProv}(\overline{\ulcorner\varphi\urcorner})$ is provable.

(ii) if $\neg\varphi$ is provable, then $\neg\mathsf{RProv}(\overline{\ulcorner\varphi\urcorner})$ is provable.

where $\mathsf{RProv}(\mathsf{y}) =_{\text{def}} \exists\mathsf{x}\,\mathsf{RPrf}(\mathsf{x},\mathsf{y})$. With these two lemmas to hand, the argument then proceeds easily:

(iii) Suppose $T \vdash \mathsf{R}$. Then, by (i) $T \vdash \mathsf{RProv}(\overline{\ulcorner\mathsf{R}\urcorner})$. But $\mathsf{RProv}(\overline{\ulcorner\mathsf{R}\urcorner})$ is trivially T-provably equivalent to $\neg\mathsf{R}$, so we have $T \vdash \neg\mathsf{R}$, contradicting the assumption of T's consistency.

(iv) Now suppose $T \vdash \neg\mathsf{R}$. Then by (ii) $T \vdash \neg\mathsf{RProv}(\overline{\ulcorner \mathsf{R} \urcorner})$ whence, by the trivial T-provable equivalence, $T \vdash \mathsf{R}$, contradicting the assumption of T's consistency.

Hence R is undecidable in T.

So do we prove the crucial lemmas (i) and (ii)?

(v) Suppose $T \vdash \varphi$. Then some number m numbers the proof of φ, i.e. we have $Prf(m, \ulcorner\varphi\urcorner)$. Further, because of T's consistency, no smaller number $w \leq m$ numbers a proof of its negation, so for all $w \leq m$ we have not-$Prf(w, neg(\ulcorner\varphi\urcorner)]$. But T can capture the relevant relations/functions, which means that we have T can (a) prove $\mathsf{Prf}(\overline{\mathsf{m}}, \overline{\ulcorner\varphi\urcorner})$ and (b) for each $w \leq m$ it proves $\neg\mathsf{Prf}(\overline{\mathsf{w}}, \mathsf{neg}(\overline{\ulcorner\varphi\urcorner}))$; and because any sensible T knows about bounded quantifiers, (b) implies that (c) T proves $\neg\exists\mathsf{w}[\mathsf{w} \leq \overline{\mathsf{m}} \wedge \mathsf{Prf}(\mathsf{w}, \mathsf{neg}(\mathsf{y}))]$. Putting (a) and (c) together, and existentially quantifying on $\overline{\mathsf{m}}$, it immediately follows that $T \vdash \mathsf{RProv}(\overline{\ulcorner\varphi\urcorner})$.

(vi) Suppose $T \vdash \neg\varphi$. Then some number m numbers the proof of $\neg\varphi$, i.e. we have $Prf(m, neg(\ulcorner\varphi\urcorner))$, whence $T \vdash \mathsf{Prf}(\overline{\mathsf{m}}, \overline{\ulcorner\varphi\urcorner})$. Granting that T knows about bounded quantifiers, it follows that

$$T \vdash \forall\mathsf{x}(\overline{\mathsf{m}} \leq \mathsf{x} \rightarrow \exists\mathsf{w}[\mathsf{w} \leq \overline{\mathsf{m}} \wedge \mathsf{Prf}(\overline{\mathsf{m}}, \mathsf{neg}(\overline{\ulcorner\varphi\urcorner}))]).$$

Now, since we are assuming T's consistency and that $T \vdash \neg\varphi$. it follows that $T \nvdash \varphi$, so for all m, not-$Prf(m, \ulcorner\varphi\urcorner)$, whence for all m, $T \vdash \neg\mathsf{Prf}(\overline{\mathsf{m}}, \overline{\ulcorner\varphi\urcorner})$. So, a fortiori, we'll have

$$T \vdash \forall\mathsf{x}(\mathsf{x} \leq \overline{\mathsf{m}} \rightarrow \neg\mathsf{Prf}(\mathsf{x}, \overline{\ulcorner\varphi\urcorner})).$$

Whence,

$$T \vdash \forall\mathsf{x}(\neg\mathsf{Prf}(\mathsf{x}, \overline{\ulcorner\varphi\urcorner})) \vee \exists\mathsf{w}[\mathsf{w} \leq \overline{\mathsf{m}} \wedge \mathsf{Prf}(\overline{\mathsf{m}}, \mathsf{neg}(\overline{\ulcorner\varphi\urcorner}))]).$$

But that, abbreviated, gives us $T \vdash \neg\mathsf{RProv}(\overline{\ulcorner\varphi\urcorner})$, and we are done.

8. Turing 1936, 1938: incompleteness assumed

One might perhaps have expected Turing to say more about the *proof* of incompleteness than he does. His great paper 'On computable numbers, with an application to the *Entscheidungsproblem*' was published in 1936. Turing emphasizes that what he will prove (concerning the undecidability of first-order logic) 'is quite different from the well-known results of Gödel'. Interestingly, he *doesn't* go on to remark that his methods in the paper yield a different proof of Gödelian incompleteness. I don't know (yet) who first noted that the proof of the undecidability of the halting problem can be parlayed into an incompleteness proof.

Turing's 1938 dissertation (published as his 1939) starts with the words 'The well-known theorem of Gödel [Turing cites both the 1931 paper and the 1934 Princeton lectures] shows that every system of logic is in a certain sense incomplete ...'. The dissertation examines what happens if you march along a sequence of theories adding at each stage the unprovable Gödel sentence (or equivalently, the unprovable consistency sentence) of what you've got so far. So this couldn't be more intimately related to Gödelian matters. However, just for the record, Turing again doesn't actually re-examine the proof of the First Theorem itself.

9. Rosser 1939: the story so far

Rosser's short paper 'An informal exposition of proofs of Gödel's Theorems and Church's Theorem' aims to do what the title suggests, i.e. give an accessible account of the incompleteness and undecidability theorems. His bibliography is very short. On incompleteness, he mentions Gödel 1931, Kleene 1936 and his own 1936. (He says, parenthetically, 'A less difficult exposition of Gödel's work is to be found in Carnap's *The logical Syntax of*

Language': but that seems rather misleading, given that Carnap doesn't get beyond the semantic version of the theorem.)

(1) The paper starts by reviewing some of the general conditions under which the described Gödelian arguments hold for a theory T. Essentially, the ones Rosser mentions concern the formal axiomatizability of T. Interestingly, nothing is explicitly made of the expressing/capturing distinction here.

(2) Rosser proves, as his Lemma 1, the general Diagonal Equivalence (not the Diagonalization Lemma) in this form: *Suppose the numerical property P is expressible, e.g. by the formula $\varphi(\mathsf{x})$. Then there is a formula with number n which is true just when n has the property P.*

The proof goes exactly as you'd now expect. Define $sub(m, n)$ as usual to be the number of the formula got by taking the formula with the number m and replacing all occurrences of x in it by the formula of T which denotes the n. T can express sub using a wff sub (Rosser remarks on how Gödel proves this expressibility claim, but adds that the proof is 'very complicated and technical, and will not even be sketched here'). Then off we go as before.

(3) Rosser now announces that Gödel shows that for a large class of theories T (using our notation):

(a) The property of numbering a T-theorem [Gödel's property *Bew*, PROVABLE] is expressible in T by a wff $\mathsf{Prov}(\mathsf{x})$.

(b) If T is ω-consistent, and $T \vdash \mathsf{Prov}(\overline{\ulcorner \varphi \urcorner})$, then $T \vdash \varphi$.

(c) If $T \vdash \varphi$, then $T \vdash \mathsf{Prov}(\overline{\ulcorner \varphi \urcorner})$.

For (b), note that if $T \nvdash \varphi$, then for each m, not-$Prf(m, n)$, and hence – by T's assumed capturing powers – for, each m, $T \vdash \neg\mathsf{Prf}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$, making T ω-inconsistent after all given we are assuming $T \vdash \exists\mathsf{x}\,\mathsf{Prf}(\mathsf{x}, \overline{\ulcorner \varphi \urcorner})$. For (c), assume for some m, $Prf(m, n)$: then $T \vdash \mathsf{Prf}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$, and existentially quantifying gives us$T \vdash \mathsf{Prov}(\overline{\ulcorner \varphi \urcorner})$.

(4) Now construct a standard Gödel sentence $\mathsf{G} =_{\text{def}} \neg\mathsf{Prov}(\overline{\mathsf{g}})$, where $g = sub(p, p)$ and p is the Gödel number of $\neg\mathsf{Prov}(\mathsf{sub}(\mathsf{x}, \mathsf{x}))$. But then $g = \ulcorner \mathsf{G} \urcorner$. So $\mathsf{G} = \neg\mathsf{Prov}(\overline{\ulcorner \mathsf{G} \urcorner})$. Then

(i) If T is consistent, then $T \nvdash \mathsf{G}$. [For if $T \vdash \mathsf{G}$, then by (c) $T \vdash \mathsf{Prov}(\overline{\ulcorner \mathsf{G} \urcorner})$, making T inconsistent.]

(ii) If T is ω-consistent, then $T \nvdash \neg\mathsf{G}$. [For if $T \vdash \neg\mathsf{G}$, then by definition $T \vdash \mathsf{Prov}(\overline{\ulcorner \mathsf{G} \urcorner})$, whence by (b) $T \vdash \mathsf{G}$, making T inconsistent.]

(5) Steps (3) and (4) make for a neat outline presentation of the First Theorem. Though Rosser implies that the proof depends on Lemma 1. Strictly speaking it doesn't. It depends on the *capturing* powers of T not the *expressive* powers of T.

(6) Next Rosser introduces his Rosser-provability property, and outlines his 1936 proof with an tweak. Previously, to show that $T \nvdash \neg\mathsf{R}$, he invokes the lemma

(i) if $T \vdash \neg\varphi$, then $T \vdash \neg\mathsf{RProv}(\overline{\ulcorner \varphi \urcorner})$.

while this time he appeals to the lemma

(i*) if $T \vdash \mathsf{RProv}(\overline{\ulcorner \varphi \urcorner})$, then $T \nvdash \neg\varphi$.

which, assuming T's consistency, follows from (i), and trivially does the needed job. Otherwise there is nothing new here. Nor is there anything new in the presentation of Kleene's proof which follows. Note, however, that Rosser doesn't comment on the fact that the Kleene proof depends on a soundness assumption whereas the previous two proofs don't. (Together with the segue from the semantic Lemma 1 into the syntactic Gödelian proofs, that's an unfortunate presentational lapse.)

10. Hilbert and Bernays 1939: the First Theorem revisited

The second volume of Hilbert and Bernays, *Grundlagen der Mathematik*, appeared in 1939: §5.1b proves the First Theorem. In the next subsection, famously, Hilbert and Bernays start on giving the first full proof of a version of the Second Theorem, and it's for *that* that the *Grundlagen* treatment of incompleteness is most remembered. But here, I want to note what they say a propos of the First Theorem.[3]

(1) H&B specify they are concerned with a 'formalism F' which captures the (primitive) recursive functions and where the relevant *Prf* relation and function *sub* are (primitive) recursive.

(2) They then first construct the free-variable wff $\neg\mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))$ (where p is the Gödel number of $\neg\mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\mathsf{y}, \mathsf{y}))$, and the argument for the unprovability of that wff goes as before along Gödel's lines, assuming F's capturing powers.

(3) Note, however, that we need Gödel's universally quantified version of this if we are to talk of negating it and showing that the result is also unprovable assuming ω-consistency. H&B again give what is in effect Gödel's proof.

(4) They then draw the following corollary. There is a (primitive) recursive function f, which can be expressed in F by the function sign f, such that (i) for all n, $F \vdash \mathsf{f}(\overline{\mathsf{n}}) = \mathsf{0}$, but (ii) $F \nvdash \forall\mathsf{x}\,\mathsf{f}(\mathsf{x}) = \mathsf{0}$, even though (iii) $\forall\mathsf{x}\,\mathsf{f}(\mathsf{x}) = \mathsf{0}$ is true. [Just take f to be the characteristic function (with 0 for 'true') of the primitive recursive property not-$Prf(x, q)$ (where $q = sub(p, p)$. This vivid way of presenting the result seems to be new to H&B.]

(5) They then turn to Rosser's proof, and do this is more fully worked-out way than Rosser himself gives.

All this is cleanly done in about as clear a version as we have yet. (Just one query: H&B assert that the provability of $\mathsf{Prf}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ implies that the relation 'holds': are they entitled to this? The capturing has to be done 'in the right way' or else in unsound F we can have a wff that captures without expressing, e.g. by carrying along a free-loading *false* theorem.)

So what systems do satisfy the constraints in (1), and in particular can capture all (primitive) recursive functions? In Volume I §8, H&B prove that all recursive functions can be captured in the arithmetic Z. I'll return to the details of their proof when I have it to hand.

11. Kleene 1943: proving the First Theorem again

Kleene's 1943 paper 'Recursive predicates and quantifiers' is another of the seminal documents in the development of computability theory. Part III of the paper is called 'Incompleteness theorems in the foundations of number theory'.

(1) Kleene proves a variant Normal Form Theorem. It yields the following: there is a three-place primitive recursive relation T such that for any two-place general recursive relation R, there is an index e such that $\exists x R(n, x)$ iff $\exists x T(e, n, x)$.

(2) Consider the property $D(n) =_{\text{def}} \forall x \neg T(n, n, x)$. Then there can be no two-place recursive predicate R such that $\exists x R(n, x)$ iff $D(n)$. [For otherwise, by (1), for some e, $\exists x T(e, n, x)$ iff $\exists x R(n, x)$ iff $\forall x \neg T(n, n, x)$. Put $n = e$ and contradiction follows.] It follows that $D(n)$ is not recursive. For if it were, then $R(n, x) =_{\text{def}} D(n) \wedge x = x$

[3] I don't read German, and an English translation of this part of the book is not yet available. So I'm relying on the French translation. And I don't pretend that my understanding of that is 100% reliable. But I hope that my concerns are broad-brush enough for what I say not to be sensitively dependent on nuances I might miss.

would be two place and recursive, but $\exists R(n, x)$ iff $D(n)$ contrary to the previous result.]

(3) Kleene defines 'a complete formal deductive theory [for the predicate $P(a)$]' to be one where 'those and only those of the formulas expressing the true instances of the predicate should be provable'. Note though that for him 'predicate' means *property*. So this is the requirement that there is a formal predicate φ such that the relevant theory Θ proves $\varphi(\overline{\mathsf{n}})$ if and only if n has the property P. Evidently a complete formal theory in this sense has to be consistent.

(4) Next, Kleene states Thesis II: *'For any given formal system and given predicate $P(a)$, the predicate that $P(a)$ is provable is expressible in the form $\exists x R(a, x)$ where R is general recursive.*' This is presumably the claim that for appropriately formal Θ, and formal predicate φ, there will be a recursive provability relation $R(n, m)$ i.e. m numbers a proof of $\varphi(\overline{\mathsf{n}})$, so that $\Theta \vdash \varphi(\overline{\mathsf{n}})$ iff $\exists x R(n, x)$.

(5) Suppose Θ is, in the defined sense, a complete formal deductive theory for the property $P(n)$. Then there is a formal wff φ and a recursive relation R such that $P(n)$ iff $\Theta \vdash \varphi(\overline{\mathsf{n}})$ iff $\exists x R(n, x)$. So, cutting out the middle step, if P has a complete formal deductive theory, then for some recursive relation R, $P(n)$ is equivalent to $\exists x R(n, x)$.

(6) From the first part of (2) and (5) it follows that there is no complete formal deductive system for the property $D(n)$.

So far, so good. Kleene, however, comments that (6) '*is the famous theorem of Gödel on formally undecidable propositions in generalized form*'. But this is a bit hasty. After all there can be weak but *complete* theories of arithmetic such as the-variable-free 'Baby Arithmetic' of addition and multiplication, which trivially is not complete for D. So, just by itself, not being complete for D doesn't entail being an incomplete theory Θ (on the usual understanding of failing to proof or refute every sentence of Θ's language). So for under what conditions does incompleteness-for-D go with incompleteness?

Kleene's following remarks start to spell things out a bit more:

> In the present form of the theorem, we have a preassigned predicate $\forall x \neg T(a, a, x)$ and a method which, to any formal system whatsoever for this predicate, gives a number f for which the following is the situation.
>
> Suppose that the system meets the condition that the formula expressing the proposition $\forall x \neg T(f, f, x)$ is provable only if that proposition is true. Then the proposition is true but the formula expressing it unprovable. This statement of results uses the interpretation of the formula, but if the system has certain ordinary deductive properties for the universal quantifier and recursive predicates, our condition on the system is guaranteed by the metamathematical one of consistency. If the system contains also a formula expressing the negation of $\forall x \neg T(f, f, x)$, and if the system meets the further condition that this formula is provable only if true, then this formula cannot be provable, and we have a formally undecidable proposition. The further condition, if the system has ordinary deductive properties, is guaranteed by the metamathematical one of ω-consistency.

But this is still less than ideal. What's the 'method' for finding the f? Indeed, what's meant by a formal system for the property $\forall x \neg T(a, a, x)$ (we know there can't be a complete one)? Kleene isn't explicit.

I take it, however, that the argument he has in mind is this:

(i) Suppose Θ is a recursively axiomatized ω-consistent theory which captures the relation T by the wff T. Let $R(n, m)$ be the recursive relation that holds when m numbers a Θ-proof of $\forall \mathsf{x}\, \mathsf{T}(\overline{\mathsf{n}}, \overline{\mathsf{n}}, \mathsf{x})$. By the variant Normal Theorem, there is an index f such that $\exists x R(n, x)$ iff $\exists x T(f, n, x)$. We'll now consider the wff $\forall \mathsf{x} \neg \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \mathsf{x})$.

(ii) Suppose $\Theta \vdash \forall \mathsf{x} \neg \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \mathsf{x})$. Then $\exists x R(f, x)$ so $\exists x T(f, f, x)$. So for some m, $T(f, f, m)$. Since Θ captures T, $\Theta \vdash \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \overline{\mathsf{m}})$ making Θ inconsistent. So if Θ is consistent, $\Theta \nvdash \forall \mathsf{x} \neg \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \mathsf{x})$.

(iii) We've just proved $\neg \exists x R(f, x)$. Whence $\neg \exists x T(f, f, x)$. So for all m, $\neg T(f, f, m)$. Since Θ captures T, for all m, $\Theta \vdash \neg \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \overline{\mathsf{m}})$. If $\Theta \vdash \neg \forall \mathsf{x} \neg \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \mathsf{x})$, then Θ is ω-inconsistent. So if Θ is ω-consistent, $\Theta \nvdash \neg \forall \mathsf{x} \neg \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \mathsf{x})$.

So $\forall \mathsf{x} \neg \mathsf{T}(\overline{\mathsf{f}}, \overline{\mathsf{f}}, \mathsf{x})$ is undecidable in Θ.

Note that compared with Kleene's 1936 argument, this is a *syntactic* proof of incompleteness (applying to axiomatized theories which capture enough and are ω-consistent), and the undecidable sentence is Π_1 rather than Π_2. So this is a double improvement on his 1936. Though with all the details now filled in, the new incompleteness proof isn't notably slicker than Gödel's proof (and not clearly a 'generalised form' either). Still, it provides an elegant twist on the original.

Finally, Kleene remarks that a variant argument using a Rosserized version of relation Rwill yield Rosser's improved theorem. But we needn't delay over this.

Part 2. Three classics of 1952/53

To be continued ...

References

Carnap, R., 1934. *Logische Syntax der Sprache.* Vienna: Springer. Translated into English as Carnap 1937.

Carnap, R., 1937. *The Logical Syntax of Language.* London: Paul, Trench.

Copeland, B. J. (ed.), 2004. *The Essential Turing.* Oxford: Clarendon Press.

Davis, M., 1965. *The Undecidable: basic papers on undecidable propositions, unsolvable problems, and computable functions.* Hewlett, NY: Raven Press.

Dawson, J. W., 1984. The reception of Godel's incompleteness theorems. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1984: 253–271.

Feferman, S., 1984. Kurt Gödel: conviction and caution. *Philosophia Naturalis*, 21: 546–562. In Feferman 1998, pp. 150–164.

Feferman, S., 1998. *In the Light of Logic.* New York: Oxford University Press.

Gödel, K., 1931. On formally undecidable propositions of *Principia Mathematica* and related systems I. In Gödel 1986, pp. 144–195.

Gödel, K., 1986. *Collected Works, Vol. 1: Publications 1929–1936.* New York and Oxford: Oxford University Press.

Goldfarb, W., 2005. On Gödel's way in: the influence of Rudolf Carnap. *Bulletin of Symbolic Logic*, 11: 185–193.

Hilbert, D. and Bernays, P., 1939. *Grundlagen der Mathematik, Vol II.* Berlin: Springer.

Kleene, S. C., 1936. General recursive functions of natural numbers. *Mathematische Annalen*, 112: 727–742. Reprinted in Davis 1965.

Kleene, S. C., 1943. Recursive predicates and quantifiers. *Transactions of the American Mathematical Society*, 53: 41–73. Reprinted in Davis 1965.

Kleene, S. C., 1952. *Introduction to Metamathematics.* Amsterdam: North-Holland Publishing Co.

Mostowski, A., 1952. *Sentences Undecidable in Formalized Arithmetic: An Exposition of the Theory of Kurt Gödel*. Amsterdam: North-Holland Publishing Co.

Rosser, J. B., 1936. Extensions of some theorems of Gödel and Church. *Journal of Symbolic Logic*, 1: 230–235. Reprinted in Davis 1965.

Rosser, J. B., 1937. Gödel theorems for non-constructive logics. *Journal of Symbolic Logic*, 2: 129–137.

Rosser, J. B., 1939. An informal exposition of proofs of Gödel's Theorems and Church's Theorem. *Journal of Symbolic Logic*, 4: 53–60. Reprinted in Davis 1965.

Tarski, A., 1933. *Pojęcie prawdy w językach nauk dedukcyjnych*. Warsaw. Translated into English in Tarksi 1956, pp. 152–278.

Tarski, A., 1936. The concept of truth in formalized languages. In Tarski 1956.

Tarski, A., 1956. *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press.

Turing, A., 1936. On computable numbers, with an application to the *Entscheidungsproblem*. *Proceedings of the London Mathematical Society*, 42: 230–265. In Copeland 2004, pp. 58–90.

Turing, A., 1939. Systems of logic based on ordinals. *Journal of the London Mathematical Society*, 45: 161–228.