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 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 (at rather varying levels of detail). 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*, Kleene's classic *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 (for all m, n of course).

[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 . . .

(2) A two-place formal predicate $\varphi(\mathsf{x},\mathsf{y})$ *captures* the numerical relation R in the formal theory T just in case
   (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 $\underline{\text{ca}}$se-by-case $\underline{\text{p}}$rove. 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 $\varphi(\mathsf{x},\mathsf{y}) \land \theta$ 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 $\varphi(\mathsf{x},\mathsf{y}) \land \theta$ 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, for all m, n,
   (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, relating any given number to a unique value. This can be done in two ways. We can add the requirement

   (c) for each m, $T \vdash \exists!\mathsf{y}\varphi(\overline{\mathsf{m}},\mathsf{y})$,

where $\exists!$ is the usual defined uniqueness quantifier. Or we can add the stronger requirement

   (c′) $T \vdash \forall\mathsf{x}\exists!\mathsf{y}\varphi(\mathsf{x},\mathsf{y})$.

If (c) holds we'll say T *captures f as a function*; if (c′) holds we'll say that T *fully captures f as a function.* The conditions on capturing-as-a-function (on modest assumptions) come to this unified condition:

   (u) if $f(m) = n$, $T \vdash \forall\mathsf{y}(\varphi(\overline{\mathsf{m}},\mathsf{y}) \leftrightarrow \mathsf{y} = \overline{\mathsf{n}})$.

It is now well known that, on modest assumptions about T, if T captures f using the predicate φ, then there will be predicates (perhaps different from φ) which capture/fully capture f as a function. So the differences between the three grades of capturing a function with a relational predicate are perhaps not of very deep significance: but they can matter for the details of technical proofs.

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.

As you would expect, if the functional expression $\psi(\mathsf{x})$ captures f, then the relational expression $\psi(\mathsf{x}) = \mathsf{y}$ also fully captures f as a function.

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 (primitive) 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: we have to wait to Tarski, Mostowski and Robinson 1953 for an analysis.

(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, though changing 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 the relation Q is primitive recursive, it can be captured by a relational 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, as that will be definable in P.]

(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* $17\,Gen\ r$. 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 $Bew(17\,Gen\ r)$, 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 an interpreted 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, for any φ, the L-sentence $\mathsf{Tr}(\overline{\ulcorner\varphi\urcorner}) \leftrightarrow \varphi$ is true. (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 $\neg\mathsf{Tr}(\overline{\ulcorner\varphi_n(\overline{\mathsf{n}})\urcorner})$ a true L-sentence. This is expressible in L. [In fact, though Tarski doesn't spell this out, $\neg\mathsf{Tr}(\mathsf{sub}(\overline{\mathsf{n}}, \overline{\mathsf{n}}))$ would do the job, for a suitable functional expression 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' (with functions captures by functional expressions, by the way). 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:

$$\mathit{fact}(0) = 1$$
$$\mathit{fact}(Sn) = Sn \cdot \mathit{fact}(n)$$

This can now be wrapped up into a single definition:

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

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 $\varphi(\ulcorner\gamma\urcorner)$. 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 $\mathit{Prf}(m, n)$ 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 $\mathit{sub}(m, n)$ 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 defined functional expression $\mathsf{sub}(\mathsf{x}, \mathsf{y})$.

(3) Now take the formula $\mathsf{U}(\mathsf{y}) =_{\mathrm{def}} \forall \mathsf{x} \neg \mathsf{Prf}(\mathsf{x}, \mathsf{sub}(\mathsf{y}, \mathsf{y}))$, 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 $\mathit{sub}(p, p) = g$. 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. $\mathit{Prf}(m, g)$. 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-$\mathit{Prf}(m, \mathit{sub}(p, p))$. 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 $\mathsf{Prov}(\mathsf{y}) =_{\mathrm{def}} \exists \mathsf{x}\, \mathsf{Prf}(\mathsf{x}, \mathsf{y})$ (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-$\mathit{Prf}(m, \mathit{sub}(n, n))$. 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

[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.

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 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 really rather pretty. But note (i) the proof as given by Kleene here is a *semantic* incompleteness proof, and (ii) that the undecidable sentence is $\forall\exists$ so is Π_2, In §12, 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 $\mathsf{RPrf}(\mathsf{x}, \mathsf{y}) =_{def} \mathsf{Prf}(\mathsf{x}, \mathsf{y}) \land \neg\exists \mathsf{w}[\mathsf{w} \leq \mathsf{x} \land \mathsf{Prf}(\mathsf{w}, \mathsf{neg}(\mathsf{y}))]$, 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 $\mathsf{R} =_{def} \forall \mathsf{x} \neg \mathsf{RPrf}(\mathsf{x}, \mathsf{sub}(\bar{\mathsf{r}}, \bar{\mathsf{r}}))$, 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. (As background, note first that $sub(r, r) = \ulcorner\mathsf{R}\urcorner$, so (again on the usual assumptions about what the relevant theory T can capture), we have $T \vdash \mathsf{sub}(\bar{\mathsf{r}}, \bar{\mathsf{r}}) = \overline{\ulcorner\mathsf{R}\urcorner}$. Abbreviate $\exists \mathsf{x} \mathsf{RPrf}(\mathsf{x}, \mathsf{y})$ as $\mathsf{RProv}(\mathsf{y})$. Then $T \vdash \mathsf{R} \leftrightarrow \neg\mathsf{RProv}(\mathsf{sub}(\bar{\mathsf{r}}, \bar{\mathsf{r}}))$ whence $T \vdash \mathsf{R} \leftrightarrow \neg\mathsf{RProv}(\overline{\ulcorner\mathsf{R}\urcorner})$.)

Rosser starts by remarking 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 he 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 $\mathsf{RProv}(\mathsf{y}) =_{\mathrm{def}} \exists \mathsf{x}\, \mathsf{RPrf}(\mathsf{x}, \mathsf{y})$. 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 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 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 $\neg\exists\mathsf{w}[\mathsf{w} \leq \overline{\mathsf{m}} \land \mathsf{Prf}(\mathsf{w}, \mathsf{neg}(\mathsf{y}))]$. 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

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

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

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

Whence,

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

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 explicitly noted that the proof of the undecidability of the halting problem can be parlayed into an incompleteness proof (though Kleene's 1943 paper gives a closely related 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, $\mathit{Prf}(m, n)$: 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 $g = \mathit{sub}(p, p)$ 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 that 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 *open* 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-$\mathit{Prf}(x, q)$ (where $q = \mathit{sub}(p, p)$. This vivid way of presenting the result seems to be new to H&B.]

[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.

(5) They then turn to Rosser's proof, and do this is in a 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.

Entry on Hilbert and Bernays to be completed

11. Quine 1940: the theory of syntax self-applied

Quine's *Mathematical Logic* of 1940 concludes with a chapter 'Syntax', where he turns 'to the formalization of the metamathematical or syntactical machinery involved in discourse *about* a formalism such as presented in the foregoing chapters. Gödel's theorem regarding the incompletability of logic and arithmetic is derived along novel lines, and its scope is somewhat extended.' So how do Quine's treatment go?

Entry on Quine to be written

12. 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 $D(n) =_{\mathrm{def}} \forall x \neg T(n,n,x)$. 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 $R(n,x) =_{\mathrm{def}} D(n) \wedge x = x$ 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, the weak but *complete* variable-free 'Baby Arithmetic' of addition and multiplication 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.

There is also a sense in which the result is an improvement on Gödel's original. For in the general version of the original argument, we take a theory Θ, and form the relational predicate Prf which captures the relation m-numbers-a-Θ-proof-of-the-wff-numbered-n, then we form a Gödel sentence from Prf. So changing Θ (e.g. by adding new axioms) will involve changing Prf so changing the underlying form of the corresponding Gödel sentence. In Kleene's proof, however, the relational predicate T stays fixed as we change Θ, and we keep the same form of undecidable sentence built from T even as we add axioms, this time just changing the numeral $\bar{\mathsf{f}}$.

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

Part 2. Three classics of 1952/53

13. Mostowski 1952: 'An Exposition of the Theory of Kurt Gödel'

In the preface to his short *Sentences Undecidable in Formalized Arithmetic* (whose subtitle is given as the heading to this section), Mostowski writes

> In the present booklet an attempt is made to present as clearly and as rigorously as possible the famous theory of undecidable sentences created by Kurt Gödel in 1931.

Well, clarity and rigour can often be in tension: and here Mostowski achieves the second aim at some cost to the first (not helped by some notational choices). Still, this is the most extended treatment since Gödel's original, and must have once been very influential (read at least by early textbook writers even if not directly by their students). So it is worth looking at in some detail, especially as it is probably not very well known these days – I, for one, didn't even really register the existence of the book until after *IGT* was written.

The Introduction gives an informal presentation of Gödel's theorem. Chs. I to V set the scene for the formal presentation. Ch. VI proves incompleteness. There is an appendix on further results.

(A) The Introduction is notable for making a distinction between a 'semantical' and 'syntactic' proof. Mostowski aims later to present many lemmas en route to incompleteness in a general form that can then be combined with more specific propositions to give syntactic and semantic versions of results. (This emphasis on generalizing/unifying seems a new turn in discussions of Gödel's theorem.)

After some initial remarks, things are set out as follows:

(1) Mostowski first presents a neat form of Richard's Paradox, which with minor notational change goes like this:

  (i) Take expressions of English that define properties of positive integers. These are denumerable and so can be listed in lexical order $W_1, W_2, W_3 \ldots$.

  (ii) Consider now the property of an integer which n has iff n doesn't have the property expressed by W_n (for short, iff $\neg W_n(n)$).

  (iii) This property must be expressed by some expression W_q. So then $W_q(n)$ iff $\neg W_n(n)$.

  (iv) Putting $n = q$ leads to contradiction.

  Moral? The notion of 'being an expression of English defining a property of the positive integers' doesn't fix a determinate totality of expressions W_i.

(2) Suppose now we set out to generate a formal analogue of Richard's Paradox: then we can get Gödel's theorem.

  (i) We take a formal system S for arithmetic, and $W_1, W_2, W_3 \ldots$ now enumerates the one-place formal predicates. (No worries this time about determinacy of the process, as we are dealing with a formal language.)

  (ii) Next we trade in the notion of 'n not having the property expressed by W_n' for the unproblematic syntactic property of n's being such that $S \vdash \neg W_n(\overline{\mathsf{n}})$.

  (iii) Now introduce Gödel numbering. Then there is a function $\varphi(n)$ which gives the number of $W_n(\overline{\mathsf{n}})$, and a property Th of being the number of a formal theorem of S. Both this numerical function and this property is expressible in S (along with the arithmetization of negation) if the theory is sufficiently strong. So (putting their expressions together) there is some expression W_q which expresses the property which n has if $S \vdash \neg W_n(\overline{\mathsf{n}})$.

  (iv) 'The intuitive content of the sentence $W_q(\overline{\mathsf{q}})$' is then that $W_q(\overline{\mathsf{q}})$ is unprovable. (But that is no paradox).

  (v) And with 'expression' here construed in our official sense, the obvious semantic argument for incompleteness follows (though Mostowski oddly credits this to Tarski rather than the opening pages of Gödel).

  (vi) Mostowski then considers what happens if we replace the semantic notion of expression with (in our terms) capturing. He in effect notes that we can't capture the property of being an S-theorem in S, but we *can* capture the relation Prf using a two-place wff, and he constructs the Gödel sentence from this in the usual way and gives the syntactic proof of incompleteness.

Note: there's a slight presentational hiccup here in that Mostowski treats both expressing (*our* sense) and capturing as forms of 'expressing' in *his* sense, and presents the initial stages of (2) as if supposedly neutral between the two – except, as he himself notes, he has to backtrack given that the property Th is expressible in one of his senses but not the other.

Still, the needed distinctions *are* in the end properly made, and he goes on to say

> The different kinds of incompleteness proofs lead to different important corollaries. We obtain them when we investigate the problem of formalization of these proofs. It turns out that the semantical proof is not formalizable within S itself. As a corollary we obtain the important theorem that the notion of 'truth' for the system S is not definable within S. The syntactical proofs are on the contrary formalizable within S and studying carefully this fact we can recognize with Godel that the consistency of S is not provable by means formalizable within S. (p. 12)

Which is a nice observation: if we imagine formalizing the semantic and syntactic first theorems, we get respectively Tarski's Theorem and Gödel's Second Theorem.

(B) What happens next in the book? In Chapter I, Mostowski introduces a standard pairing function, and explains what is essentially Gödel's β-function trick for coding up finite sequences and hence for converting a definition by primitive recursion into an explicit definition in basically the now familiar way.

But one wrinkle here is the use of the minimization operator. To take a simple case, suppose $f(0) = k$, $f(n+1) = g(f(n))$. Then we can define $f(n)$ by taking the least number g such that the values for the decoding β-function for $0, 1, \ldots n$, i.e. the values

$a_0, a_1, \ldots, a_n$ are such that $a_0 = k$, and $a_{n+1} = g(a_n)$, and then applying to g the decoding function that extracts a_n.

In Chapter II, 'System S and its syntax' describes a formal system system and simultaneously its arithmetization. In fact, officially all the real work is done at the latter level: 'Instead of expressions we shall deal uniquely with the integers which correspond to them. Our exposition will therefore belong entirely to arithmetic.' (So here's a case of going for rigour at the expense of a certain easy readability, though Mostowski introduces helpfully mnemonic abbreviations into the arithmetic metatheory.)

The arithmetical system S is rather non-standard in various ways:

(1) The primitive expressions are distinguished free and bound variables; the arithmetical constant 1 and functions + and ×; the equality sign; the propositional connective →; and the least-number operator μ; plus parentheses.

(2) Mostowski takes $1 = 1 + 1$ as a falsum, and then his axioms for propositional logic are Church's.

(3) The least number operator applied to a condition φ returns the least number n such that $\varphi(n)$ if there is one, or else defaults to 1. We can thus define $\exists x\varphi(x)$ as $\varphi(\mu x\varphi(x))$, and give rules for μ instead of the quantifier. Mostowski gives rules for this which include, in effect, the least number principle.

(4) The adopted identity axioms are (i) the expected reflexivity, symmetry and transitivity principles. But instead of a general version of Leibniz's law we get principles that (i) identicals can be substituted on the left in $a + b$ and $a \times b$, (iii) that the minimizations $\mu v_i\varphi(v_i)$ and $\mu v_j\varphi(v_j)$ will be equal (assuming v_i and v_j don't otherwise appear in φ), and (iv) minimization on equivalent wffs produces identical results.

(5) The special rules for arithmetic for addition and multiplication are much as you'd expect, presented as open sentences with free variables, and omitting induction.

(6) A formal S-proof is defined in the obvious way, with modus ponens as the sole rule.

Chapter III then proves that S includes (with predictable definitions) the usual propositional and predicate calculus, that Leibniz's Law holds in a general form for expressions of S, and the usual induction principle can be derived because of the least number principle. It is also shown that various arithmetical principles hold, including natural results for a defined '<'-sign.

The semantics of S is dealt with in Chapter IV. This is essentially a Tarskian semantics, but arithmetized. Thus infinite sequences of integers a_i – with $a_1 = 1$ for all but finitely many indices i – are coded with finite integers, and otherwise definitions go as you'd expect.

(C) Chapter V discusses expressing and capturing recursive functions and relations.

(1) Henceforth, $\mathfrak{K}$ [Mostowski's notation] is a consistent set of S-wffs [he allows open sentences], including the S-axioms and closed under implication. Then Mostowski will say

   (i) The two-place relation R is $\mathfrak{K}$-definable [his notation] by the two-place predicate φ just in case

      (a) if Rmn, then $\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ is in $\mathfrak{K}$,

      (b) if not-Rmn, then $\neg\varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ is in $\mathfrak{K}$.

   (ii) The one-place function f is $\mathfrak{K}$-definable by the functional expression ψ iff, for any n, if $f(m) = n$, the sentence $\psi(\overline{\mathsf{m}}) = \overline{\mathsf{n}}$ is in $\mathfrak{K}$.

The definitions expand in obvious ways to predicates and functions of different arities. And on modest assumptions, the conditional in the second definition can be strengthened to a biconditional.

Putting things in our terminology, when $\mathfrak{K} = Tr$, i.e. the set of S-truths (as specified by the semantics), then $\mathfrak{K}$-defining is *expressing*; and when more restrictedly $\mathfrak{K} = Th$, i.e. the set of S-theorems, then $\mathfrak{K}$-defining is *capturing*. In Mostowski's terminology, when $\mathfrak{K} = Tr$, being $\mathfrak{K}$-definable is simply being *definable*, when $\mathfrak{K} = Th$ being $\mathfrak{K}$-definable is being *recursive*. He announces that recursiveness in this sense coincides with Turing computability, general recursiveness, etc.; but he doesn't make use of this. So, for the moment, we can take Mostowski's 'recursive' to be a mere shorthand tag for the class of Th-definable relations and functions, without further descriptive content.

(2) Mostowski now proves a number of results about $\mathfrak{K}$-definability in general:

(i) For any $\mathfrak{K}$, 'Recursive functions and relations are $\mathfrak{K}$-definable.' [This is just a truism given what Mostowski means by recursive. For example, given if $R(m, n)$ then $S \vdash \mathsf{R}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$, it follows that if $R(m, n)$ then $S \vdash \mathsf{R}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ is in any $\mathfrak{K}$ including the S-axioms and closed under implication. Etc.]

(ii) *Initial functions* The identity relation, the less than relation, sum, multiplication and the n-constant functions for any n are all recursive so $\mathfrak{K}$-definable. [Likewise for projection functions. All straightforward, given easy results from Ch. III.]

(iii) *Composition* Suppose the one-place property P is $\mathfrak{K}$-definable by $\varphi(\mathsf{x})$, and the one-place functions f, g are $\mathfrak{K}$-definable by $\psi(x)$, $\chi(x)$. Then the property that n has when $P(f(n))$ is $\mathfrak{K}$-definable by $\varphi(\psi(\mathsf{x}))$. And the function $g(f(n))$ is is $\mathfrak{K}$-definable by $\chi(\psi(\mathsf{x}))$. With obvious generalisations for many-place relations/functions. [Similarly straightforward, given easy results from Ch. III.]

(iv) *Defn. by recursion* If g is $\mathfrak{K}$-definable, then the function f defined by $f(0) = k$, $f(n') = g(f(n))$ is $\mathfrak{K}$-definable. Similarly for more complex cases of definition by primitive recursion. [The proof uses the facts that (a) f can be explicitly defined using minimisation and the decoding β-function, (b) if a relation R is $\mathfrak{K}$-definable, and $\forall m \exists n Rmn$, then the function $\mu x Rxn$ is $\mathfrak{K}$-definable, (c) the decoding β-function is $\mathfrak{K}$-definable.]

(v) Boolean combinations of $\mathfrak{K}$-definable properties are $\mathfrak{K}$-definable.

This gets the result that all primitive recursive functions are $\mathfrak{K}$-definable.

(3) Mostowski now shows that some key relations and functions involved in the arithmetization of syntax are recursive so $\mathfrak{K}$-definable:

(i) The property of being the number of an S-term, an S-wff, an S-sentence, etc., are recursive so $\mathfrak{K}$-definable. So are substitution relations, etc. [Look at the definitions in Ch. II, and note they only involved bounded quantifiers.]

(ii) The property of numbering an S-axiom, the relation that holds between the numbers of the premisses and conclusion of a modus ponens, and hence the property of numbering an S-proof are all recursive. And if the property of belonging to class $\mathfrak{L}$ of wffs is $\mathfrak{K}$-definable, then the property of numbering a proof in the theory which extends S with axioms in $\mathfrak{L}$ of wffs is $\mathfrak{K}$-definable. [By inspection of the details of the arithmetization of syntax.]

(iii) The property of numbering an S-theorem is definable in Mostowski's sense (but can be shown later not to be recursive). And if the property of belonging to class $\mathfrak{L}$ of wffs is definable, then the property of numbering a theorem of the theory which extends S with axioms in $\mathfrak{L}$ is definable. [Unbounded existential quantifications of $\mathfrak{K}$-definable relations are definable, but not in general $\mathfrak{K}$-definable.]

These crucial results are proved straightforwardly, groundwork having already been done in the detailed presentation of the arithmetized syntax.

(4) The last section of Ch. V introduces the notion of a recursively enumerable set (defined of course via Mostowski's notion of recursivitity), shows that any recursive set is r.e., that that if a set and its complement are r.e., then it is recursive, and that the theorems of extension of S with a recursive set of axioms are r.e.

(D) So now we turn to the pivotal Chapter VI, 'Proofs of Incompleteness Theorems'.

(1) As before $\mathfrak{K}$ is a consistent set of S-wffs [he allows open sentences], and 'closed' [i.e. including the S-axioms and closed under implication]. We get the elegantly unifying

Theorem I: *the set of sentences in $\mathfrak{K}$ is not $\mathfrak{K}$-definable.*

Proof. Suppose otherwise. Then take the property n has when $\varphi_n(\mathsf{\overline{n}})$ is not among the sentences in $\mathfrak{K}$ (where φ_n is the expression with number n). Then this property too is $\mathfrak{K}$-definable, so there is some one-free variable expression φ_d which $\mathfrak{K}$-defines it, so (by definition)

(i) if $\varphi_n(\mathsf{\overline{n}})$ is not among the sentences in $\mathfrak{K}$, then $\varphi_d(\mathsf{\overline{n}})$ is in $\mathfrak{K}$,

(ii) if $\varphi_n(\mathsf{\overline{n}})$ is among the sentences in $\mathfrak{K}$, then $\varphi_d(\mathsf{\overline{n}})$ is not in $\mathfrak{K}$.

But $\varphi_d(\mathsf{\overline{n}})$ is a sentence, and putting $n = d$ gives us a contradiction.

Corollaries. (a) The set *Tr* of true S-sentences is not *Tr*-definable (definable in Mostowski's sense, i.e. expressible). (b) The set *Th* of S-theorems is not *Th*-definable (recursive in Mostowski's sense, i.e. capturable). (c) The complement of *Th* is not r.e.

(2) Mostowski now proves a negative version of the Diagonalization Lemma. With modernized notation, suppose ψ is any open wff with one-free variable. Then

Theorem II: *There is a sentence γ such that $S \vdash \neg\gamma \leftrightarrow \psi(\overline{\ulcorner\gamma\urcorner})$.*

Proof. Let $\sigma(x)$ be a recursive definition of the function *sub* that maps n to the number of $\varphi_n(\mathsf{\overline{n}})$. Suppose ψ is any open wff with one-free variable. Let p be the number of the open wff $\neg\psi(\sigma(\mathsf{x}))$, and let γ be the wff $\neg\psi(\sigma(\mathsf{\overline{p}}))$.

Then $sub(p) = \ulcorner\gamma\urcorner$, whence $S \vdash \sigma(\mathsf{\overline{p}}) = \overline{\ulcorner\gamma\urcorner}$. Hence $S \vdash \psi(\sigma(\mathsf{\overline{p}})) \leftrightarrow \psi(\overline{\ulcorner\gamma\urcorner})$. Whence $S \vdash \neg\gamma \leftrightarrow \psi(\overline{\ulcorner\gamma\urcorner})$. (And more generally, as Mostowski prefers to put it, $\neg\gamma \leftrightarrow \psi(\overline{\ulcorner\gamma\urcorner})$ is in any $\mathfrak{K}$.)

But note, Mostowski doesn't use this lemma in quite the modern way – he doesn't use it in his syntactic version of the incompleteness theorem noted in (4) below – though his seems to be the first really clear statement of a version of the general Diagonalization Lemma.

(3) First, semantic, proof of incompleteness for S. Use the obvious argument that the set *Th* is definable but (by Theorem I) *Tr* isn't so either some theorems aren't truths or some truths aren't theorems. The first is ruled out by S's soundness. So there are unprovable truths whose negations, being false, are also unprovable.

Theorem II allows us to sharpen the result and give us an example of an unprovable sentence. Take the recursive *Prf* relation, and suppose that the wff Prf *Th*-defines it (i.e. captures it). Note by Theorem II applied to $\exists\mathsf{y}\mathsf{Prf}(\mathsf{y},\mathsf{x})$ we have a wff γ such that $S \vdash \neg\gamma \leftrightarrow \exists\mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\ulcorner\gamma\urcorner})$. Since, as we are assuming, S is sound we have either $\neg\gamma \wedge \exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\ulcorner\gamma\urcorner})$ is true or $\gamma \wedge \neg\exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\ulcorner\gamma\urcorner})$ is true. The first is ruled out otherwise we'd have $\exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\ulcorner\gamma\urcorner})$ true, so γ provable contradicting the soundness of S. So γ is true, and $\neg\exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\ulcorner\gamma\urcorner})$ true, therefore γ is unprovable: and its negation being false, that is unprovable too. So γ us undecidable, hence its provable equivalent $\forall \mathsf{y}\neg\mathsf{Prf}(\mathsf{y}, \overline{\ulcorner\gamma\urcorner})$ is undecidable. And as Mostowski remarks, that is the universal quantification of something recursive (so, as we would now say, is Π_1).

(4) Second, syntactic, proof à la Gödel. Theorem I tells us that Th is not recursive. So, in particular, the wff $\exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \mathsf{x})$ does not Th-define it. Hence for some n we must have either

n numbers a theorem, but $S \nvdash \exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\mathsf{n}})$, or
n doesn't number a theorem, but (a) $S \nvdash \neg\exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\mathsf{n}})$.

But the first is ruled out, since if n numbers a theorem, there is some m such that $Prf(m, n)$ so $S \vdash \mathsf{Prf}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$ whence $S \vdash \exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\mathsf{n}})$ after all.

So n doesn't number a theorem. So for all m, not-$Prf(m, n)$ so $S \vdash \neg\mathsf{Prf}(\overline{\mathsf{m}}, \overline{\mathsf{n}})$. Hence, on pain of ω-inconsistency, (b) $S \nvdash \exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\mathsf{n}})$.

Whence from (a) and (b), $\neg\exists \mathsf{y}\mathsf{Prf}(\mathsf{y}, \overline{\mathsf{n}})$ is undecidable. (Mostowski remarks that the construction in Theorem II yields a suitable value for n, but that is not needed for this proof.)

(5) Third, syntactic, proof à la Rosser. As usual, this is harder work, and Mostowski's presentation doesn't really add anything of interest (over and above the interest in his slight twist on Gödel's proof).

(E) Still in Chapter VI, Mostowski now has a section 'Generalizations'. In particular, he notes

(1) *No recursively enumerable set of sentences (including the S axioms and closed under implication) is complete* – so any recursively axiomatized extension of S is incomplete. [Rerun the Rosser argument, using the Rosser-provability predicate defined in terms of the wff that captures the enumerating function.]

(2) '*No definable closed and ω-consistent class $\mathfrak{K}$ is complete.*'

(3) These two results are 'not comparable' for (a) there exists a definable closed ω-consistent set of sentences which is not recursively enumerable, and (b) there exists a recursively enumerable and consistent closed set of sentences which is ω-consistent. [Mostowski's proof of (a) is interesting. Take the open sentences $\varphi(\mathsf{x})$ which are such that for all n, $S \vdash \varphi(\overline{\mathsf{n}})$. Form $\mathfrak{L}$ the set of sentences $\forall \mathsf{x}\varphi(\mathsf{x})$, and then take $\mathfrak{K}$ to be its deductive closure in S. $\mathfrak{L}$ is definable [expressible], and hence so is $\mathfrak{K}$; everything in $\mathfrak{L}$ and hence in $\mathfrak{K}$ is true, so the latter is ω-consistent. But $\mathfrak{K}$ isn't r.e., or else we could do a Gödel construction and find an unprovable-in-$\mathfrak{K}$ wff which is of the form $\forall \mathsf{x}\varphi(\mathsf{x})$ where each instance of $\varphi(\overline{\mathsf{n}})$ is S-provable.]

(F) Just for the record, in the three-part Appendix which finishes the 115 page book, Mostowski discusses

(1) The Second Theorem. Here, the general project of arithmetizing the proof of part of the First Theorem is described.

(2) The truth of undecidable sentences (as obtained the proof à la Gödel in Ch. VI). Mostowski considers extending S to a system S_1 which adds enough set theory to formalize the semantics for S, which gives us an extended theory that can prove the standard Gödel sentence for S – but is, of course, still incomplete.

(3) Speed up. Very roughly, there are sentences provable in S which have vastly shorter proofs in S_1. Less roughly, for any primitive recursive f, there is a proof of size g_1 (measured by Gödel number) in S_1 of a wff which is already provable in S but whose smallest proof in S whose size $g > f(g_1)$. [Could be worth returning to this discussion, as the proof looks somewhat different to the one I give for a closely related result in *IGT* §21.7.]

14. Kleene 1952: the pivotal textbook

In his note 'The writing of *Introduction to Metamathematics*' 1991, Kleene notes that up to 1985, about 17,500 copies of the English version of his wonderful text were sold, as were more thousands of various translations (including a sold-out first print run of 8000 of the Russian translation). So this is the book with a quite pivotal influence on the education of later logicians (including textbook writers!) – and on their understanding of the incompleteness theorems in particular, for Gödelian incompleteness is indeed a central theme of the book.

(A) In Ch. IV, Kleene presents a formal system (he doesn't seem to label it) which is a version of first-order Peano Arithmetic. The underlying logic is done Frege-Hilbert style, with $\&, \vee, \supset, \neg$ as connectives and both $\forall, \exists$ as basic (Kleene later introduces natural-deduction style rules as derived rules). Kleene treats the identity rules as part of the 'postulates for number theory', and doesn't give Leibniz's Law as basic, but rather derives its application for complex arithmetical predicates from simpler assumptions, as Mostowski also did. His other arithmetical primitives are '0', '+', '×', with '′' used for successor. A sign '<' is introduced in the familiar way by stipulating that $\mathsf{x} < \mathsf{y}$ abbreviates $\exists \mathsf{z}(\mathsf{z}' + \mathsf{x} = \mathsf{y})$.

Ch. V explains the notion of a formal dedication, Ch. VI explores the background propositional logic, Ch. VII the background predicate logic. Then Ch. VIII starts work on 'Formal number theory':

§38 As noted, the identity laws for Kleene count as part of his number theory, so this section proves that identity behaves as you'd expect. Then '<' is introduced by the usual definition.

§39 Outlines proofs of some expected results about addition, multiplication and order.

§40 Derives the least number principle, legitimacy of course-of-values induction, etc.

§41 (1) Introduces the idea of a formal wff capturing (Kleene: 'numeralwise expressing') a numerical property as contrasted with expressing it ('express[ing] under the interpretation of the symbolism'): stresses that capturing does *not* require that 'various general properties of the predicate should be formal provable'. Shows that $\mathsf{x} < \mathsf{y} =_{\text{def}} \exists \mathsf{z}(\mathsf{z}' + \mathsf{x} = \mathsf{y})$ captures $x < y$, and proves that formal restricted quantifiers defined using '<' behave as expected.

(2) (Kleene notes the rest of the section is skippable until §49.) Defines what it is for a formal predicate to capture a function as a function (as *IGT* puts it, and we put it in §1 above) or 'numeralwise represent' the function (as Kleene puts it). Thus, a two-place formal predicate $\varphi(\mathsf{x}, \mathsf{y})$ *captures* the one-place numerical function f as a function in the formal theory T just in case for all m,

(a) if $f(m) = n$, then $T \vdash \varphi(\overline{\mathsf{m}}, \overline{\mathsf{n}})$,

(c) $T \vdash \exists! \mathsf{x} \varphi(\overline{\mathsf{m}}, \mathsf{x})$.

(Clause (b) in our original statement in §1 is redundant.) Capturing as a function entails plain capturing (as defined in §1 above) but not vice versa. Kleene doesn't here say *why* he is later going to want to work with capturing as a function rather than plain capturing – i.e. *why* is it important whether T 'knows' the capturing predicate is functional?

(3) The basic successor, addition and multiplication functions can be captured as functions. Logic-free equations are all provable when true, refutable when false. Likewise for Boolean combinations of equations. Quotient and remainder are capturable as functions.

Hence (though not here announced as such) the β-function $rm(c, (i'.d)')$ is capturable as a function.

N.B. Kleene knows that his full system isn't needed for this since he has knows of Robinson's work, of which only an abstract had been published in 1952. (But this is a last-minute addition, not really integrated into the book.)

(B) Ch. VIII, §42, proves Gödel's Theorem, assuming a Lemma eventually to be proved in §52. In more detail, this section includes:

(1) Brief informal presentation of incompleteness theorem via the fact that 'a closed formula can be found which, interpreted by a person who knows [about the used enumeration of wffs], asserts its own provability'.

(2) Very informal characterization of Gödel numbering.

(3) The Lemma: (i) There is a formula (in *IGT* notation) Gdl which captures *Gdl* the relation that bolds when m numbers the proof of the result of substituting n for the free variable in the formula number n. (i) There is also a formula $\overline{\mathsf{Gdl}}$ which captures the relation $\overline{Gdl}$ that bolds when m numbers the proof of the negation of the result of substituting n for the free variable in the formula number n. [Note, it is only the capturing of a *relation* that is needed at this point.]

(4) The usual Gödel sentence $\forall \mathsf{x} \neg \mathsf{Gdl}(\mathsf{x}, \overline{\mathsf{p}})$ where p is the number of $\forall \mathsf{x} \neg \mathsf{Gdl}(\mathsf{x}, \mathsf{y})$. The incompleteness theorem is proved in a standard laborious way (as in Gödel 1934). [Nothing hereabouts on Diagonal Lemma.]

(5) The usual Rosser sentence, constructed from Gdl and $\overline{\mathsf{Gdl}}$. Rosser's improved incompleteness theorem is proved in standard direct way [e.g. as sketched in §7 (C) above)].

(6) Informal remarks on Second Theorem.

(7) Remarks on phenomenon of ω-incompleteness.

(8) Very brief note on consistency extensions.

(9) Finally in the section, Kleene notes 'Thus far we have Gödel's theorem only for our particular formal system (except for the last remarks, which refer to a succession of systems). The question arises now whether it may not depend on some peculiarities of the present formalization of logic, and might be avoided in some other. In the next chapters, besides completing the proof of the required lemma for Gödel's theorem, we shall reach a standpoint from which we can discuss these questions for formal systems in general, with the formal system studied here as an example (§§60, 61).'

(C) Ch. IX gives an extended treatment of primitive recursive functions. It starts as you'd expect, with §43 primitive recursive functions informally defined, and then §44 a more careful presentation. §45 defines relative primitive recursiveness and introduces characteristic functions (Kleene 'representing function's), with 0 as true. The section also defines primitive recursive predicates. Then Kleene shows that bounded quantifiers, definition by cases, etc. preserve primitive recursiveness: various functions related to prime number representation are also primitive recursive. §46 shows definition by 'course of values' recursion also yields primitive recursive functions. §47 is skippable (starred section, an aside on the notion of 'primitive recursive uniformly'). Then we are on to stuff that matters more specifically to the presentation of Gödel's theorem:

§48 Gödel's β-function. This is briskly and cleanly presented – no explicit mention of Chinese Remainder Theorem [worth checking if better explained than in *IGT*?].

§49 The key section proving, first, that all primitive recursive functions are arithmetic (definable from successor, addition and multiplication with the usual logical operators). Then 'By translating from the intuitive arithmetical symbolism into the formal symbolism, we obtain a formula $\mathsf{P}(\mathsf{x}_1, \ldots, \mathsf{x}_r, \mathsf{w})$ which expresses $f(x_1, \ldots, x_r) = w$ under the interpretation of the formal system.'

Then Kleene shows that any primitive recursive function $f(x_1, \ldots, x_r)$ can be captured-as-a-function by the relevant constructed wff $\mathsf{P}(\mathsf{x}_1, \ldots, \mathsf{x}_r, \mathsf{w})$. The structure of the proof is the expected induction on a constructional history of f, and the troublesome case is proving the second, uniqueness, condition for capturing-as-a-function for the case where a function is defined by primitive recursion from known primitive recursive functions. This is a rather rebarbatively dense half-page on p. 244. So this is what we'd ideally like to skirt around if possible. [Question: How much more difficult does Kleene make things for himself by requiring capturing-as-a-function rather than plain capturing? What *exactly* do we gain by taking his route?]

(D) Ch. X is on the arithmetization of syntax (or as Kleene puts it, better perhaps, the arithmetization of metamathematics). The approach is rather sophisticated.

§50 Kleene first treats the syntax of his formal system as a 'generalized arithmetic' – its basic symbols are the 'zeros' and its constructors – such as the three-place constructor that takes us from a quantifier-former '$\forall$', variable 'x' and wff 'φ' to '$\forall\mathsf{x}\varphi$' – are the 'successor' functions. In a more modern parlance, we have here a recursive datatype (see e.g. Forster 2003), over which we can run inductive arguments and definitions by recursion.

§51 Definitions by recursion of various metamathematical properties of items in the recursive datatype. E.g. of 'x is a numeral' or 'y is a term' or 'z is an axiom'.

§52 Gödel numbering based on powers of primes – but a more complex version than is common, but with a certain elegance. So take e.g. $\exists\mathsf{y}\,\neg\mathsf{y} = \mathsf{0}$. In our new perspective, this comes from the three-place constructor applied to '$\exists$', 'y' and '$\neg\mathsf{y} = \mathsf{0}$', and gets the value $2^{\ulcorner\exists\urcorner}\cdot 3^{\ulcorner\mathsf{y}\urcorner}\cdot 5^{\ulcorner\neg\mathsf{y} = \mathsf{0}\urcorner}$, and now e.g. the number $\ulcorner\neg\mathsf{y} = \mathsf{0}\urcorner$, being the number of the two-place constructor that applies negation to the wff $\mathsf{y} = \mathsf{0}$, gets the number $2^{\ulcorner\neg\urcorner}\cdot 3^{\ulcorner\mathsf{y} = \mathsf{0}\urcorner}$, and $\ulcorner\mathsf{y} = \mathsf{0}\urcorner$ is in turn $2^{\ulcorner=\urcorner}\cdot 3^{\ulcorner\mathsf{y}\urcorner}\cdot 5^{\ulcorner\mathsf{0}\urcorner}$. And so it goes. This means repeatedly taking prime factorization of a Gödel number, and factorizing exponents, etc., directly mirrors repeatedly digging down levels of syntactic structure.

Passing then from syntactic entities to their Gödel numbers, we can move from syntactic properties to corresponding numerical properties. And the properties corresponding to those defined by recursion in §51 are fairly seen to be primitive recursive. For example, take the property of being a term – i.e. either being 0, or being a variable, or the successor of a term or the addition/multiplication of two terms. This goes over to the numerical property of numbering a term being either the number for 0, or the number for a variable, or the number for the addition/multiplication of two terms. And then that definition by 'course of values' recursion can be turned into a primitive recursion. (As Kleene puts it, 'The gist of this proof is that the primitive recursions in the generalized arithmetic become course-of-values recursions in the ordinary arithmetic, since the Gödel numbering preserves the order relationships although it destroys the relationships of immediate succession.')

Finally in this section, Kleene shows that relations Gdl and $\overline{Gdl}$ are primitive recursive, and that the property of numbering a theorem of his system is expressible (not capturable) by a wff which expresses the existential quantification of a primitive recursive relation. So that gives us the Lemma needed for his proofs of completeness.

(E) The final section of Chapter X (on inductive and recursive definitions) is starred as skippable. Chapter XI turns to the discussion of general recursive functions, introduced Herbrand-Gödel style in §§54, 55 and shown to go beyond the primitive recursive. §56 'arithmetizes' the formalism for general recursive functions. §57 introduces the unbounded minimization operator, and proves μ-recursiveness and general recursiveness are equivalent, proves enumeration theorem etc. §58 proves the normal form theorem, etc. In §59, Kleene shows that general recursiveness and capturability in his formal system are equivalent (and both the same as what he calls 'reckonability'). The next two sections then return to the incompleteness theorem.

§60 Discusses generalizing the incompleteness theorem beyond the formal system that has been the topic of the previous seven chapters, and revisits the 'generalized theorem' of Kleene 1943.

In fact, Theorem XIII Part iii on p. 304 is pretty much the reconstructed result that I extracted towards the end of §11 above. Part ii of the theorem, given earlier on p. 303, is a version which uses a semantic assumption of 'correctness' instead of ω-consistency. (Suppose S is 'correct and complete' for property $\forall x \neg T(\overline{n}, \overline{n}, x)$, i.e. S is such that for all n, $S \vdash \forall \mathsf{x} \neg \mathsf{T}(\bar{\mathsf{n}}, \bar{\mathsf{n}}, \mathsf{x})$ iff $\forall x \neg T(n, n, x)$. But we can choose f as before so that $S \vdash \forall \mathsf{x} \neg \mathsf{T}(\bar{\mathsf{f}}, \bar{\mathsf{f}}, \mathsf{x})$ iff $\exists x \neg T(f, f, x)$. Contradiction follows.)

To repeat what I said before in discussing Kleene 1943, note that following Gödel's original construction, undecidable Gödel sentences for different theories will be built in a uniform way but from different predicates Prf which capture the different Prf relations for different theories. Changing a theory by adding new axioms will involve changing the underlying form of the corresponding Gödel sentence. In Kleene's proof, however, the relational predicate T can stay fixed as we change the theory, and we keep the same form of undecidable sentence built from T even as we add axioms, this time just changing the relevant index f. (Kleene 1991 later puts it like this: 'Thus, for each such formal system [formally axiomatized, can capture the T-relation, etc.] we have as a formally undecidable proposition ... a respective value of one preassigned predicate ..., a result which Gödel did not have in 1931.')

Kleene also mentions what we would now call the possibility of Σ_1 complete theories and a number of other results.

§61 Kleene now presents a 'symmetric [generalized] form of Gödel's theorem' which – roughly – stands to Rosser's theorem as the generalized result from 1943 stands to Gödel's original. Essentially the trick is to Rosserize the $T(x, x, y)$ relation, in two different ways to give two distinct primitive recursive $W_0(x, y)$ and $W_1(x, y)$, where $\forall x \neg(\exists y W_0(x, y) \wedge \exists y W_1(x, y))$. Then, so long as our theory can capture these two relations, prove the formal analogue of the disjointness condition we've just given, and is consistent, we can find an f such that neither the formula $\forall \mathsf{y} \neg \mathsf{W}_0(\bar{\mathsf{f}}, \mathsf{y})$

> (which is true) nor its negation is provable (and symmetrically, there's a number g such that neither $\forall \mathsf{y} \neg \mathsf{W}_1(\overline{\mathsf{g}}, \mathsf{y})$ nor its negation is provable).
>
> Kleene 1991 reports that 'Kreisel wrote me ... that Gödel on several occasions remarked on this symmetric form as being a very significant improvement of his incompleteness results.' But I wonder why? This isn't immediately clear. You might think that the *significant* improvement is already there in Kleene's earlier generalization, in the point that the form of the undecidable sentence it produces stays fixed (modulo the coding scheme) as a theory is expanded with new axioms – we only have to adjust a numerical index. So what does the symmetric version add, other than that the argument can be Rosserized to avoid an assumption of ω-consistency? [But we'll return to this question later.]

I've recounted what's in the relevant parts of *Introduction to Metamathematics* in rather patchy detail. But Kleene writes so clearly, he usually leaves little need for running commentary. And, even though I'd not read this closely for decades, there is a sense of familiarity about much of what he writes (modulo relatively superficial presentational matters), as aspects of his discussion evidently shape many later textbooks. The book remains a really impressive achievement.

15. Tarski, Mostowski, and Robinson 1953: what it takes

We know by this stage in the history that it is enough for a theory to be properly axiomatized, consistent, and capture enough (primitive) recursive functions/relations to be incomplete. But just what does it take to capture (enough) primitive recursive functions and relations?

First-order Peano Arithmetic does the trick. But what about cut-down subsystems? This is the problem that was cracked by Raphael Robinson; he shows that the finitely axiomatized subsystem Q can capture all primitive recursive functions, and indeed capture as functions (so any consistent properly axiomatized extension is incomplete too); and an infinitely axiomatized subsystem of *that*, dubbed R, in fact suffices.

These days, that's all very familiar. But it's worth going back to the second essay of the three that comprise Tarski, Mostowski, and Robinson, 1953 to see how the main result about the adequacy of Q and R is proved there. (As noted, Kleene in his 1952 knows of this result, but it clearly came to his notice too late to be properly woven into his book.)

(A) Before turning to the adequacy of Q and R, we should pause over the paper's quite elegant first theorem. Fix a sensible system of Gödel numbering. Define the diagonal function $D(n)$ for theory T to be the function that maps n to the $\varphi_n(\overline{\mathsf{n}})$ where the T-wff φ_n has number n. Let Th be the property of numbering a T-theorem. Then, by assumption, for all m, n,

> Theorem: *If T is consistent, it can't both capture D as a function, and also capture Th.*

Proof: Suppose $\delta(\mathsf{x}, \mathsf{y})$ captures D as a function and $\theta(\mathsf{x})$ captures Th in T. Then, by definition,

(a) if $D(m) = n$ then $T \vdash \delta(\overline{\mathsf{m}}, \overline{\mathsf{n}})$.
(b) for any m, $T \vdash \exists! \mathsf{y} \delta(\overline{\mathsf{m}}, \mathsf{y})$.
(c) if $Th(n)$, then $T \vdash \theta(\overline{\mathsf{n}})$
(d) if not-$Th(n)$, then $T \vdash \neg\theta(\overline{\mathsf{n}})$

Now consider the wff $\forall \mathsf{y}(\delta(\mathsf{x}, \mathsf{y}) \rightarrow \neg\theta(\mathsf{y}))$, and suppose this has number d, i.e. is $\varphi_d(\overline{\mathsf{x}})$. And let the wff $\varphi_d(\overline{\mathsf{d}})$, i.e. $\forall \mathsf{y}(\delta(\overline{\mathsf{d}}, \mathsf{y}) \rightarrow \neg\theta(\mathsf{y}))$, have number g. So $D(d) = g$, and hence by (a) we have

(e) $T \vdash \delta(\overline{\mathsf{d}}, \overline{\mathsf{g}})$

Suppose $T \vdash \varphi_d(\overline{\mathsf{d}})$, i.e. $T \vdash \forall \mathsf{y}(\delta(\overline{\mathsf{d}}, \mathsf{y}) \to \neg\theta(\mathsf{y}))$. Then $T \vdash \neg\theta(\overline{\mathsf{g}})$. Suppose alternatively $T \nvdash \varphi_d(\overline{\mathsf{d}})$. Then not-$Th(g)$. So by (d) $T \vdash \neg\theta(\overline{\mathsf{g}})$. Hence, either way, we get

(f) $T \vdash \neg\theta(\overline{\mathsf{g}})$.

Now, from (b) we trivially have

(g) $T \vdash \exists! \mathsf{y}\delta(\overline{\mathsf{d}}, \mathsf{y})$.

And now simple logic plus (e), (g) and (f) give

(h) $T \vdash \forall \mathsf{y}(\delta(\overline{\mathsf{d}}, \mathsf{y}) \to \neg\theta(\mathsf{y}))$.

But that's to say $T \vdash \varphi_d(\overline{\mathsf{d}})$, so $Th(g)$, so by (c),

(i) $T \vdash \theta(\overline{\mathsf{g}})$.

And that with (f) makes T inconsistent.

Which is all quite neat, though – if we didn't know the motivating background behind this sort of construction – it would seem a bit tricksy. Three comments:

(i) Note that (b) – the condition that ensures that δ *captures D as a function* – is essential to the proof. Here, then, we do have a proof that turns on the use of a stronger notion of capturing.

(ii) Note too that some previous constructions in our history used a *functional* wff sub to capture diagonalizing substitution, so the construction could start by going from some wff $\varphi(\mathsf{x})$ to a wff $\varphi(\mathsf{sub}(\mathsf{x}, \mathsf{x}))$. The alternative trick of instead going from $\varphi(\mathsf{x})$ to $\forall \mathsf{y}(\delta(\mathsf{x}, \mathsf{y}) \to \varphi(\mathsf{y}))$, using a relational wff δ to capture the diagonalizing substitution function, a little trick much used in later discussions, is foregrounded [for the first time?] here.

(iii) Since D is capturable in any nice theory, we can conclude that Th isn't. But Th for any consistent complete theory is a decidable property, so capturable. So nice theories aren't complete. But TM&R don't pursue questions of incompleteness here: they focus is, of course, on undecidability.

(B) Let's turn, then, to the question of capturing (primitive) recursive functions as functions in Q and R.

(1) The seven-axiom Robinson arithmetic Q (the version where order relations aren't basic but are introduced by definition) can be taken as familiar (though note TM&R's proof of the independence of the axioms).

The theory R is the theory which has as axioms

(i) all true atomic additions and multiplications for numerals,

(ii) $\overline{\mathsf{m}} \neq \overline{\mathsf{n}}$ whenever $m \neq n$,

(iii) $\mathsf{x} \leq \overline{\mathsf{n}} \leftrightarrow \mathsf{x} = \mathsf{0} \lor \mathsf{x} = \overline{\mathsf{1}} \lor \mathsf{x} = \overline{\mathsf{2}} \lor \ldots \lor \mathsf{x} = \overline{\mathsf{n}}$ for each n, and

(iv) $\mathsf{x} \leq \overline{\mathsf{n}} \lor \overline{\mathsf{n}} \leq \mathsf{x}$ for each n.

where $\leq$ is defined in the usual way. It is routine to show that R is contained in Q. So it is enough to show that R captures every recursive function as a function.

(2) One way of characterizing recursive functions is as those got from initial functions, via repeated application of composition, recursion (implemented by 'for' loops) and minimization (implemented by 'do until' loops). Intuitively, 'for' loops can be treated as a special kind of 'do until' loop, so we'd expect that you can also characterize recursive functions is as those got from some initial functions, via composition and minimization as constructors. And indeed, this can be done if we are allowed enough initial functions. There are different ways of implementing this idea, but Julia Robinson discovered one in her 1950 paper 'General recursive functions'. Namely: all recursive functions can be defined using successor, sum, 'excess over a square' [the difference between a number and the largest square

no larger than it'], composition and a form of minimization [given a function g, this minimization yields the function that given m returns the least n such that $g(n) = m$, assuming g takes every value].

Robinson's construction is pretty ingenious – after all, it isn't exactly obvious even how to even get multiplication – but the very ingenuity required makes the construction rather unrevealing.

(3) TM&R use Julia Robinson's clever result: so it is enough for them to show that the initial functions can be captured-as-functions in R, and that composition and the required form of minimization take us capturable to capturable functions. [NB, no struggle with the β-function is needed, as when we try more directly to represent definitions by recursion.]

This is relatively straightforward: see pp. 56–59 of their essay. For example, they announce that if G captures the one-place g as a function and likewise H captures h, then $\forall \mathsf{z}(\mathsf{H}(\mathsf{x},\mathsf{z}) \wedge \mathsf{G}(\mathsf{z},\mathsf{y}))$ captures $g \circ h$. And $\mathsf{G}(\mathsf{y},\mathsf{x}) \wedge \forall \mathsf{z}(\mathsf{G}(\mathsf{z},\mathsf{x}) \to \mathsf{y} \leq \mathsf{z})$ captures the function that given m returns the least n such that $g(n) = m$. It is then just a matter of checking the details and noting that the proofs only use facts known to R. No additional trickery is required.

(4) Note however that the wffs here are built up with universal quantifiers. Thus if G is Π_1 (because defined by composition, say), then minimising gives a Π_2 function. As definitional chains get longer, the quantifier complexity of the constructed capturing wff increases (or so it seems). So TM&R's method of showing R and Q are adequate to capture the recursive functions won't deliver the canonical result that Σ_1 wffs suffice (which is what we need to get the consequent key result that there are Π_1 Gödel sentences). [So in this respect, TM&R don't ideally deliver what we want?]

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.

Forster, T., 2003. *Logic, Induction and Sets.* London Mathematical Society Student Texts. Cambridge: Cambridge 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.

Kleene, S. C., 1991. The writing of *Introduction to Metamathematics.* In T. Drucker (ed.), *Perspectives on the history of mathematical logic*, pp. 161–168. Boston: Birkhäuser.

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

Robinson, J., 1950. General recursive functions. *Proceedings of the American Mathematical Society*, 1: 703–718.

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.

Tarski, A., Mostowski, A., and Robinson, R., 1953. *Undecidable Theories.* Amsterdam: North-Holland Publishing Co.

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.