Chapter udf

Incompleteness and Provability

inp.1 Introduction

inc:inp:int:sec

Hilbert thought that a system of axioms for a mathematical structure, such as the natural numbers, is inadequate unless it allows one to derive all true statements about the structure. Combined with his later interest in formal systems of deduction, this suggests that he thought that we should guarantee that, say, the formal systems we are using to reason about the natural numbers is not only consistent, but also *complete*, i.e., every statement in its language is either derivable or its negation is. Gödel's first incompleteness theorem shows that no such system of axioms exists: there is no complete, consistent, axiomatizable formal system for arithmetic. In fact, no "sufficiently strong," consistent, axiomatizable mathematical theory is complete.

A more important goal of Hilbert's, the centerpiece of his program for the justification of modern ("classical") mathematics, was to find finitary consistency proofs for formal systems representing classical reasoning. With regard to Hilbert's program, then, Gödel's second incompleteness theorem was a much bigger blow. The second incompleteness theorem can be stated in vague terms, like the first incompleteness theorem. Roughly speaking, it says that no sufficiently strong theory of arithmetic can prove its own consistency. We will have to take "sufficiently strong" to include a little bit more than **Q**.

The idea behind Gödel's original proof of the incompleteness theorem can be found in the Epimenides paradox. Epimenides, a Cretan, asserted that all Cretans are liars; a more direct form of the paradox is the assertion "this sentence is false." Essentially, by replacing truth with derivability, Gödel was able to formalize a sentence which, in a roundabout way, asserts that it itself is not derivable. If that sentence were derivable, the theory would then be inconsistent. Gödel showed that the negation of that sentence is also not derivable from the system of axioms he was considering. (For this second part, Gödel had to assume that the theory **T** is what's called "ω-consistent." ω-Consistency is related to consistency, but is a stronger property.[1] A few years after Gödel,

[1] That is, any ω-consistent theory is consistent, but not vice versa.

Rosser showed that assuming simple consistency of **T** is enough.)

The first challenge is to understand how one can construct a sentence that refers to itself. For every formula φ in the language of **Q**, let $\ulcorner\varphi\urcorner$ denote the numeral corresponding to ${}^{\#}\varphi^{\#}$. Think about what this means: φ is a formula in the language of **Q**, ${}^{\#}\varphi^{\#}$ is a natural number, and $\ulcorner\varphi\urcorner$ is a *term* in the language of **Q**. So every formula φ in the language of **Q** has a *name*, $\ulcorner\varphi\urcorner$, which is a term in the language of **Q**; this provides us with a conceptual framework in which formulas in the language of **Q** can "say" things about other formulas. The following lemma is known as the fixed-point lemma.

Lemma inp.1. *Let* **T** *be any theory extending* **Q***, and let* $\psi(x)$ *be any formula with only the variable* x *free. Then there is a sentence* φ *such that* $\mathbf{T} \vdash \varphi \leftrightarrow \psi(\ulcorner\varphi\urcorner)$.

The lemma asserts that given any property $\psi(x)$, there is a sentence φ that asserts "$\psi(x)$ is true of me," and **T** "knows" this.

How can we construct such a sentence? Consider the following version of the Epimenides paradox, due to Quine:

> "Yields falsehood when preceded by its quotation" yields falsehood when preceded by its quotation.

This sentence is not directly self-referential. It simply makes an assertion about the syntactic objects between quotes, and, in doing so, it is on par with sentences like

1. "Robert" is a nice name.
2. "I ran." is a short sentence.
3. "Has three words" has three words.

But what happens when one takes the phrase "yields falsehood when preceded by its quotation," and precedes it with a quoted version of itself? Then one has the original sentence! In short, the sentence asserts that it is false.

inp.2 The Fixed-Point Lemma

explanation

The fixed-point lemma says that for any formula $\psi(x)$, there is a sentence φ such that $\mathbf{T} \vdash \varphi \leftrightarrow \psi(\ulcorner\varphi\urcorner)$, provided **T** extends **Q**. In the case of the liar sentence, we'd want φ to be equivalent (provably in **T**) to "$\ulcorner\varphi\urcorner$ is false," i.e., the statement that ${}^{\#}\varphi^{\#}$ is the Gödel number of a false sentence. To understand the idea of the proof, it will be useful to compare it with Quine's informal gloss of φ as, "'yields a falsehood when preceded by its own quotation' yields a falsehood when preceded by its own quotation." The operation of taking an expression, and then forming a sentence by preceding this expression by its own quotation may be called *diagonalizing* the expression, and the result its diagonalization. So, the diagonalization of 'yields a falsehood when preceded

inc:inp:fix:sec

by its own quotation' is "'yields a falsehood when preceded by its own quotation' yields a falsehood when preceded by its own quotation." Now note that Quine's liar sentence is not the diagonalization of 'yields a falsehood' but of 'yields a falsehood when preceded by its own quotation.' So the property being diagonalized to yield the liar sentence itself involves diagonalization!

In the language of arithmetic, we form quotations of a formula with one free variable by computing its Gödel numbers and then substituting the standard numeral for that Gödel number into the free variable. The diagonalization of $\alpha(x)$ is $\alpha(\overline{n})$, where $n = {}^{\ulcorner}\alpha(x)^{\urcorner\#}$. (From now on, let's abbreviate $\overline{{}^{\ulcorner}\alpha(x)^{\urcorner\#}}$ as $\overline{{}^{\ulcorner}\alpha(x)^{\urcorner}}$.) So if $\psi(x)$ is "is a falsehood," then "yields a falsehood if preceded by its own quotation," would be "yields a falsehood when applied to the Gödel number of its diagonalization." If we had a symbol diag for the function $\mathrm{diag}(n)$ which computes the Gödel number of the diagonalization of the formula with Gödel number n, we could write $\alpha(x)$ as $\psi(\mathit{diag}(x))$. And Quine's version of the liar sentence would then be the diagonalization of it, i.e., $\alpha(\overline{{}^{\ulcorner}\alpha(x)^{\urcorner}})$ or $\psi(\mathit{diag}(\overline{{}^{\ulcorner}\psi(\mathit{diag}(x))^{\urcorner}}))$. Of course, $\psi(x)$ could now be any other property, and the same construction would work. For the incompleteness theorem, we'll take $\psi(x)$ to be "x is not derivable in $\mathbf{T}$." Then $\alpha(x)$ would be "yields a sentence not derivable in $\mathbf{T}$ when applied to the Gödel number of its diagonalization."

To formalize this in $\mathbf{T}$, we have to find a way to formalize diag. The function $\mathrm{diag}(n)$ is computable, in fact, it is primitive recursive: if n is the Gödel number of a formula $\alpha(x)$, $\mathrm{diag}(n)$ returns the Gödel number of $\alpha(\overline{{}^{\ulcorner}\alpha(x)^{\urcorner}})$. (Recall, $\overline{{}^{\ulcorner}\alpha(x)^{\urcorner}}$ is the standard numeral of the Gödel number of $\alpha(x)$, i.e., $\overline{{}^{\ulcorner}\alpha(x)^{\urcorner\#}}$). If diag were a function symbol in $\mathbf{T}$ representing the function diag, we could take φ to be the formula $\psi(\mathit{diag}(\overline{{}^{\ulcorner}\psi(\mathit{diag}(x))^{\urcorner}}))$. Notice that

$$\begin{aligned}\mathrm{diag}({}^{\ulcorner}\psi(\mathit{diag}(x))^{\urcorner\#}) &= {}^{\ulcorner}\psi(\mathit{diag}(\overline{{}^{\ulcorner}\psi(\mathit{diag}(x))^{\urcorner}}))^{\#}\\ &= {}^{\ulcorner}\varphi^{\urcorner\#}.\end{aligned}$$

Assuming $\mathbf{T}$ can derive

$$\mathit{diag}(\overline{{}^{\ulcorner}\psi(\mathit{diag}(x))^{\urcorner}}) = {}^{\ulcorner}\varphi^{\urcorner},$$

it can derive $\psi(\mathit{diag}(\overline{{}^{\ulcorner}\psi(\mathit{diag}(x))^{\urcorner}})) \leftrightarrow \psi({}^{\ulcorner}\varphi^{\urcorner})$. But the left hand side is, by definition, φ.

Of course, diag will in general not be a function symbol of $\mathbf{T}$, and certainly is not one of $\mathbf{Q}$. But, since diag is computable, it is *representable* in $\mathbf{Q}$ by some formula $\theta_{\mathrm{diag}}(x, y)$. So instead of writing $\psi(\mathit{diag}(x))$ we can write $\exists y\,(\theta_{\mathrm{diag}}(x, y) \wedge \psi(y))$. Otherwise, the proof sketched above goes through, and in fact, it goes through already in $\mathbf{Q}$.

inc:inp:fix: lem:fixed-point

Lemma inp.2. *Let $\psi(x)$ be any formula with one free variable x. Then there is a sentence φ such that $\mathbf{Q} \vdash \varphi \leftrightarrow \psi({}^{\ulcorner}\varphi^{\urcorner})$.*

Proof. Given $\psi(x)$, let $\alpha(x)$ be the formula $\exists y\,(\theta_{\mathrm{diag}}(x, y) \wedge \psi(y))$ and let φ be its diagonalization, i.e., the formula $\alpha(\overline{{}^{\ulcorner}\alpha(x)^{\urcorner}})$.

Since θ_{diag} represents diag, and $\text{diag}(\ulcorner \alpha(x) \urcorner) = \ulcorner \varphi \urcorner$, **Q** can derive

$$\theta_{\text{diag}}(\ulcorner \alpha(x) \urcorner, \ulcorner \varphi \urcorner) \tag{inp.1}$$

$$\forall y\, (\theta_{\text{diag}}(\ulcorner \alpha(x) \urcorner, y) \to y = \ulcorner \varphi \urcorner). \tag{inp.2}$$

inc:inp:fix: repdiag1 inc:inp:fix: repdiag2

Now we show that $\mathbf{Q} \vdash \varphi \leftrightarrow \psi(\ulcorner \varphi \urcorner)$. We argue informally, using just logic and facts derivable in **Q**.

First, suppose φ, i.e., $\alpha(\ulcorner \alpha(x) \urcorner)$. Going back to the definition of $\alpha(x)$, we see that $\alpha(\ulcorner \alpha(x) \urcorner)$ just is

$$\exists y\, (\theta_{\text{diag}}(\ulcorner \alpha(x) \urcorner, y) \wedge \psi(y)).$$

Consider such a y. Since $\theta_{\text{diag}}(\ulcorner \alpha(x) \urcorner, y)$, by eq. (inp.2), $y = \ulcorner \varphi \urcorner$. So, from $\psi(y)$ we have $\psi(\ulcorner \varphi \urcorner)$.

Now suppose $\psi(\ulcorner \varphi \urcorner)$. By eq. (inp.1), we have

$$\theta_{\text{diag}}(\ulcorner \alpha(x) \urcorner, \ulcorner \varphi \urcorner) \wedge \psi(\ulcorner \varphi \urcorner).$$

It follows that

$$\exists y\, (\theta_{\text{diag}}(\ulcorner \alpha(x) \urcorner, y) \wedge \psi(y)).$$

But that's just $\alpha(\ulcorner \alpha(x) \urcorner)$, i.e., φ. □

digression

You should compare this to the proof of the fixed-point lemma in computability theory. The difference is that here we want to define a *statement* in terms of itself, whereas there we wanted to define a *function* in terms of itself; this difference aside, it is really the same idea.

Problem inp.1. A formula $\varphi(x)$ is a *truth definition* if $\mathbf{Q} \vdash \psi \leftrightarrow \varphi(\ulcorner \psi \urcorner)$ for all sentences ψ. Show that no formula is a truth definition by using the fixed-point lemma.

inp.3 The First Incompleteness Theorem

inc:inp:1in: sec

We can now describe Gödel's original proof of the first incompleteness theorem. Let **T** be any computably axiomatized theory in a language extending the language of arithmetic, such that **T** includes the axioms of **Q**. This means that, in particular, **T** represents computable functions and relations.

We have argued that, given a reasonable coding of formulas and proofs as numbers, the relation $\text{Prf}_T(x, y)$ is computable, where $\text{Prf}_T(x, y)$ holds if and only if x is the Gödel number of a derivation of the formula with Gödel number y in **T**. In fact, for the particular theory that Gödel had in mind, Gödel was able to show that this relation is primitive recursive, using the list of 45 functions and relations in his paper. The 45th relation, xBy, is just $\text{Prf}_T(x, y)$ for his particular choice of **T**. Remember that where Gödel uses the word "recursive" in his paper, we would now use the phrase "primitive recursive."

Since $\mathrm{Prf}_T(x, y)$ is computable, it is representable in $\mathbf{T}$. We will use $\mathsf{Prf}_T(x, y)$ to refer to the formula that represents it. Let $\mathsf{Prov}_T(y)$ be the formula $\exists x\, \mathsf{Prf}_T(x, y)$. This describes the 46th relation, $\mathrm{Bew}(y)$, on Gödel's list. As Gödel notes, this is the only relation that "cannot be asserted to be recursive." What he probably meant is this: from the definition, it is not clear that it is computable; and later developments, in fact, show that it isn't.

Let $\mathbf{T}$ be an axiomatizable theory containing $\mathbf{Q}$. Then $\mathrm{Prf}_T(x, y)$ is decidable, hence representable in $\mathbf{Q}$ by a formula $\mathsf{Prf}_T(x, y)$. Let $\mathsf{Prov}_T(y)$ be the formula we described above. By the fixed-point lemma, there is a formula $\gamma_{\mathbf{T}}$ such that $\mathbf{Q}$ (and hence $\mathbf{T}$) derives

inc:inp:1in: eqn:qpf

$$\gamma_{\mathbf{T}} \leftrightarrow \neg \mathsf{Prov}_T(\ulcorner \gamma_{\mathbf{T}} \urcorner). \tag{inp.3}$$

Note that $\gamma_{\mathbf{T}}$ says, in essence, "$\gamma_{\mathbf{T}}$ is not derivable in $\mathbf{T}$."

inc:inp:1in: lem:cons-G-unprov

Lemma inp.3. *If $\mathbf{T}$ is a consistent, axiomatizable theory extending $\mathbf{Q}$, then $\mathbf{T} \nvdash \gamma_{\mathbf{T}}$.*

Proof. Suppose $\mathbf{T}$ derives $\gamma_{\mathbf{T}}$. Then there *is* a derivation, and so, for some number m, the relation $\mathrm{Prf}_T(m, {}^{\#}\gamma_{\mathbf{T}}{}^{\#})$ holds. But then $\mathbf{Q}$ derives the sentence $\mathsf{Prf}_T(\overline{m}, \ulcorner \gamma_{\mathbf{T}} \urcorner)$. So $\mathbf{Q}$ derives $\exists x\, \mathsf{Prf}_T(x, \ulcorner \gamma_{\mathbf{T}} \urcorner)$, which is, by definition, $\mathsf{Prov}_T(\ulcorner \gamma_{\mathbf{T}} \urcorner)$. By eq. (inp.3), $\mathbf{Q}$ derives $\neg \gamma_{\mathbf{T}}$, and since $\mathbf{T}$ extends $\mathbf{Q}$, so does $\mathbf{T}$. We have shown that if $\mathbf{T}$ derives $\gamma_{\mathbf{T}}$, then it also derives $\neg \gamma_{\mathbf{T}}$, and hence it would be inconsistent. □

inc:inp:1in: thm:oconsis-q

Definition inp.4. A theory $\mathbf{T}$ is *ω-consistent* if the following holds: if $\exists x\, \varphi(x)$ is any sentence and $\mathbf{T}$ derives $\neg\varphi(\overline{0})$, $\neg\varphi(\overline{1})$, $\neg\varphi(\overline{2})$, ... then $\mathbf{T}$ does not prove $\exists x\, \varphi(x)$.

Note that every ω-consistent theory is also consistent. This follows simply from the fact that if $\mathbf{T}$ is inconsistent, then $\mathbf{T} \vdash \varphi$ for every φ. In particular, if $\mathbf{T}$ is inconsistent, it derives both $\neg\varphi(\overline{n})$ for every n and also derives $\exists x\, \varphi(x)$. So, if $\mathbf{T}$ is inconsistent, it is ω-inconsistent. By contraposition, if $\mathbf{T}$ is ω-consistent, it must be consistent.

inc:inp:1in: lem:omega-cons-G-unref

Lemma inp.5. *If $\mathbf{T}$ is an ω-consistent, axiomatizable theory extending $\mathbf{Q}$, then $\mathbf{T} \nvdash \neg \gamma_{\mathbf{T}}$.*

Proof. We show that if $\mathbf{T}$ derives $\neg \gamma_{\mathbf{T}}$, then it is ω-inconsistent. Suppose $\mathbf{T}$ derives $\neg \gamma_{\mathbf{T}}$. If $\mathbf{T}$ is inconsistent, it is ω-inconsistent, and we are done. Otherwise, $\mathbf{T}$ is consistent, so it does not derive $\gamma_{\mathbf{T}}$ by Lemma inp.3. Since there is no derivation of $\gamma_{\mathbf{T}}$ in $\mathbf{T}$, $\mathbf{Q}$ derives

$$\neg \mathsf{Prf}_T(\overline{0}, \ulcorner \gamma_{\mathbf{T}} \urcorner), \neg \mathsf{Prf}_T(\overline{1}, \ulcorner \gamma_{\mathbf{T}} \urcorner), \neg \mathsf{Prf}_T(\overline{2}, \ulcorner \gamma_{\mathbf{T}} \urcorner), \ldots$$

and so does $\mathbf{T}$. On the other hand, by eq. (inp.3), $\neg \gamma_{\mathbf{T}}$ is equivalent to $\exists x\, \mathsf{Prf}_T(x, \ulcorner \gamma_{\mathbf{T}} \urcorner)$. So $\mathbf{T}$ is ω-inconsistent. □

Problem inp.2. Every ω-consistent theory is consistent. Show that the converse does not hold, i.e., that there are consistent but ω-inconsistent theories. Do this by showing that $\mathbf{Q} \cup \{\neg\gamma_{\mathbf{Q}}\}$ is consistent but ω-inconsistent.

Theorem inp.6. *Let* $\mathbf{T}$ *be any* ω*-consistent, axiomatizable theory extending* $\mathbf{Q}$*. Then* $\mathbf{T}$ *is not complete.* inc:inp:1in: thm:first-incompleteness

Proof. If $\mathbf{T}$ is ω-consistent, it is consistent, so $\mathbf{T} \nvdash \gamma_{\mathbf{T}}$ by Lemma inp.3. By Lemma inp.5, $\mathbf{T} \nvdash \neg\gamma_{\mathbf{T}}$. This means that $\mathbf{T}$ is incomplete, since it derives neither $\gamma_{\mathbf{T}}$ nor $\neg\gamma_{\mathbf{T}}$. □

inp.4 Rosser's Theorem

Can we modify Gödel's proof to get a stronger result, replacing "ω-consistent" with simply "consistent"? The answer is "yes," using a trick discovered by Rosser. Rosser's trick is to use a "modified" derivability predicate $\mathsf{RProv}_T(y)$ instead of $\mathsf{Prov}_T(y)$. inc:inp:ros: sec

Theorem inp.7. *Let* $\mathbf{T}$ *be any consistent, axiomatizable theory extending* $\mathbf{Q}$*. Then* $\mathbf{T}$ *is not complete.* inc:inp:ros: thm:rosser

Proof. Recall that $\mathsf{Prov}_T(y)$ is defined as $\exists x\, \mathsf{Prf}_T(x, y)$, where $\mathsf{Prf}_T(x, y)$ represents the decidable relation which holds iff x is the Gödel number of a derivation of the sentence with Gödel number y. The relation that holds between x and y if x is the Gödel number of a *refutation* of the sentence with Gödel number y is also decidable. Let $\mathrm{not}(x)$ be the primitive recursive function which does the following: if x is the code of a formula φ, $\mathrm{not}(x)$ is a code of $\neg\varphi$. Then $\mathrm{Ref}_T(x, y)$ holds iff $\mathrm{Prf}_T(x, \mathrm{not}(y))$. Let $\mathsf{Ref}_T(x, y)$ represent it. Then, if $\mathbf{T} \vdash \neg\varphi$ and δ is a corresponding derivation, $\mathbf{Q} \vdash \mathsf{Ref}_T(\ulcorner\delta\urcorner, \ulcorner\varphi\urcorner)$. We define $\mathsf{RProv}_T(y)$ as

$$\exists x\, (\mathsf{Prf}_T(x, y) \wedge \forall z\, (z < x \to \neg\mathsf{Ref}_T(z, y))).$$

Roughly, $\mathsf{RProv}_T(y)$ says "there is a proof of y in $\mathbf{T}$, and there is no shorter refutation of y." Assuming $\mathbf{T}$ is consistent, $\mathsf{RProv}_T(y)$ is true of the same numbers as $\mathsf{Prov}_T(y)$; but from the point of view of *provability* in $\mathbf{T}$ (and we now know that there is a difference between truth and provability!) the two have different properties. If $\mathbf{T}$ is *in*consistent, then the two do *not* hold of the same numbers! ($\mathsf{RProv}_T(y)$ is often read as "y is Rosser provable." Since, as just discussed, Rosser provability is not some special kind of provability—in inconsistent theories, there are sentences that are provable but not Rosser provable—this may be confusing. To avoid the confusion, you could instead read it as "y is shmovable.")

By the fixed-point lemma, there is a formula $\rho_{\mathbf{T}}$ such that

$$\mathbf{Q} \vdash \rho_{\mathbf{T}} \leftrightarrow \neg\mathsf{RProv}_T(\ulcorner\rho_{\mathbf{T}}\urcorner). \tag{inp.4}$$ inc:inp:ros: RT

In contrast to the proof of Theorem inp.6, here we claim that if $\mathbf{T}$ is consistent, $\mathbf{T}$ doesn't derive $\rho_{\mathbf{T}}$, and $\mathbf{T}$ also doesn't derive $\neg\rho_{\mathbf{T}}$. (In other words, we don't need the assumption of ω-consistency.)

First, let's show that $\mathbf{T} \nvdash \rho_T$. Suppose it did, so there is a derivation of ρ_T from T; let n be its Gödel number. Then $\mathbf{Q} \vdash \mathsf{Prf}_T(\overline{n}, \ulcorner \rho_T \urcorner)$, since Prf_T represents Prf_T in $\mathbf{Q}$. Also, for each $k < n$, k is not the Gödel number of a derivation of $\neg\rho_T$, since $\mathbf{T}$ is consistent. So for each $k < n$, $\mathbf{Q} \vdash \neg\mathsf{Ref}_T(\overline{k}, \ulcorner \rho_T \urcorner)$. By **??**, $\mathbf{Q} \vdash \forall z\,(z < \overline{n} \to \neg\mathsf{Ref}_T(z, \ulcorner \rho_T \urcorner))$. Thus,

$$\mathbf{Q} \vdash \exists x\,(\mathsf{Prf}_T(x, \ulcorner \rho_T \urcorner) \wedge \forall z\,(z < x \to \neg\mathsf{Ref}_T(z, \ulcorner \rho_T \urcorner))),$$

but that's just $\mathsf{RProv}_T(\ulcorner \rho_T \urcorner)$. By eq. (inp.4), $\mathbf{Q} \vdash \neg\rho_T$. Since $\mathbf{T}$ extends $\mathbf{Q}$, also $\mathbf{T} \vdash \neg\rho_T$. We've assumed that $\mathbf{T} \vdash \rho_T$, so $\mathbf{T}$ would be inconsistent, contrary to the assumption of the theorem.

Now, let's show that $\mathbf{T} \nvdash \neg\rho_T$. Again, suppose it did, and suppose n is the Gödel number of a derivation of $\neg\rho_T$. Then $\mathrm{Ref}_T(n, {}^{\#}\rho_T{}^{\#})$ holds, and since Ref_T represents Ref_T in $\mathbf{Q}$, $\mathbf{Q} \vdash \mathsf{Ref}_T(\overline{n}, \ulcorner \rho_T \urcorner)$. We'll again show that $\mathbf{T}$ would then be inconsistent because it would also derive ρ_T. Since

$$\mathbf{Q} \vdash \rho_T \leftrightarrow \neg\mathsf{RProv}_T(\ulcorner \rho_T \urcorner),$$

and since $\mathbf{T}$ extends $\mathbf{Q}$, it suffices to show that

$$\mathbf{Q} \vdash \neg\mathsf{RProv}_T(\ulcorner \rho_T \urcorner).$$

The sentence $\neg\mathsf{RProv}_T(\ulcorner \rho_T \urcorner)$, i.e.,

$$\neg\exists x\,(\mathsf{Prf}_T(x, \ulcorner \rho_T \urcorner) \wedge \forall z\,(z < x \to \neg\mathsf{Ref}_T(z, \ulcorner \rho_T \urcorner))),$$

is logically equivalent to

$$\forall x\,(\mathsf{Prf}_T(x, \ulcorner \rho_T \urcorner) \to \exists z\,(z < x \wedge \mathsf{Ref}_T(z, \ulcorner \rho_T \urcorner))).$$

We argue informally using logic, making use of facts about what $\mathbf{Q}$ derives. Suppose x is arbitrary and $\mathsf{Prf}_T(x, \ulcorner \rho_T \urcorner)$. We already know that $\mathbf{T} \nvdash \rho_T$, and so for every k, $\mathbf{Q} \vdash \neg\mathsf{Prf}_T(\overline{k}, \ulcorner \rho_T \urcorner)$. Thus, for every k it follows that $x \neq \overline{k}$. In particular, we have (a) that $x \neq \overline{n}$. We also have $\neg(x = \overline{0} \vee x = \overline{1} \vee \cdots \vee x = \overline{n-1})$ and so by **??**, (b) $\neg(x < \overline{n})$. By **??**, $\overline{n} < x$. Since $\mathbf{Q} \vdash \mathsf{Ref}_T(\overline{n}, \ulcorner \rho_T \urcorner)$, we have $\overline{n} < x \wedge \mathsf{Ref}_T(\overline{n}, \ulcorner \rho_T \urcorner)$, and from that $\exists z\,(z < x \wedge \mathsf{Ref}_T(z, \ulcorner \rho_T \urcorner))$. Since x was arbitrary we get, as required, that

$$\forall x\,(\mathsf{Prf}_T(x, \ulcorner \rho_T \urcorner) \to \exists z\,(z < x \wedge \mathsf{Ref}_T(z, \ulcorner \rho_T \urcorner))). \qquad \square$$

Problem inp.3. Two sets A and B of natural numbers are said to be *computably inseparable* if there is no decidable set X such that $A \subseteq X$ and $B \subseteq \overline{X}$ ($\overline{X}$ is the complement, $\mathbb{N} \setminus X$, of X). Let $\mathbf{T}$ be a consistent axiomatizable extension of $\mathbf{Q}$. Suppose A is the set of Gödel numbers of sentences provable in $\mathbf{T}$ and B the set of Gödel numbers of sentences refutable in $\mathbf{T}$. Prove that A and B are computably inseparable.

inp.5 Comparison with Gödel's Original Paper

inc:inp:gop:sec

It is worthwhile to spend some time with Gödel's 1931 paper. The introduction sketches the ideas we have just discussed. Even if you just skim through the paper, it is easy to see what is going on at each stage: first Gödel describes the formal system P (syntax, axioms, proof rules); then he defines the primitive recursive functions and relations; then he shows that xBy is primitive recursive, and argues that the primitive recursive functions and relations are represented in **P**. He then goes on to prove the incompleteness theorem, as above. In Section 3, he shows that one can take the unprovable assertion to be a sentence in the language of arithmetic. This is the origin of the β-lemma, which is what we also used to handle sequences in showing that the recursive functions are representable in **Q**. Gödel doesn't go so far to isolate a minimal set of axioms that suffice, but we now know that **Q** will do the trick. Finally, in Section 4, he sketches a proof of the second incompleteness theorem.

inp.6 The Derivability Conditions for PA

inc:inp:prc:sec

Peano arithmetic, or **PA**, is the theory extending **Q** with induction axioms for all formulas. In other words, one adds to **Q** axioms of the form

$$(\varphi(0) \wedge \forall x\,(\varphi(x) \to \varphi(x'))) \to \forall x\,\varphi(x)$$

for every formula φ. Notice that this is really a *schema*, which is to say, infinitely many axioms (and it turns out that **PA** is *not* finitely axiomatizable). But since one can effectively determine whether or not a string of symbols is an instance of an induction axiom, the set of axioms for **PA** is computable. **PA** is a much more robust theory than **Q**. For example, one can easily prove that addition and multiplication are commutative, using induction in the usual way. In fact, most finitary number-theoretic and combinatorial arguments can be carried out in **PA**.

Since **PA** is computably axiomatized, the derivability predicate $\mathrm{Prf}_{\mathbf{PA}}(x, y)$ is computable and hence represented in **Q** (and so, in **PA**). As before, we will take $\mathsf{Prf}_{\mathbf{PA}}(x, y)$ to denote the formula representing the relation. Let $\mathsf{Prov}_{\mathbf{PA}}(y)$ be the formula $\exists x\,\mathrm{Prf}_{\mathbf{PA}}(x, y)$, which, intuitively says, "y is derivable from the axioms of **PA**." The reason we need a little bit more than the axioms of **Q** is we need to know that the theory we are using is strong enough to derive a few basic facts about this derivability predicate. In fact, what we need are the following facts:

P1. If $\mathbf{PA} \vdash \varphi$, then $\mathbf{PA} \vdash \mathsf{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner)$.

P2. For all formulas φ and ψ,

$$\mathbf{PA} \vdash \mathsf{Prov}_{\mathbf{PA}}(\ulcorner\varphi \to \psi\urcorner) \to (\mathsf{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner) \to \mathsf{Prov}_{\mathbf{PA}}(\ulcorner\psi\urcorner)).$$

P3. For every formula φ,

$$\mathbf{PA} \vdash \mathsf{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner) \to \mathsf{Prov}_{\mathbf{PA}}(\ulcorner\mathsf{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner)\urcorner).$$

The only way to verify that these three properties hold is to describe the formula $\mathsf{Prov}_{\mathbf{PA}}(y)$ carefully and use the axioms of **PA** to describe the relevant formal derivations. Conditions (1) and (2) are easy; it is really condition (3) that requires work. (Think about what kind of work it entails ...) Carrying out the details would be tedious and uninteresting, so here we will ask you to take it on faith that **PA** has the three properties listed above. A reasonable choice of $\mathsf{Prov}_{\mathbf{PA}}(y)$ will also satisfy

P4. If $\mathbf{PA} \vdash \mathsf{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner)$, then $\mathbf{PA} \vdash \varphi$.

But we will not need this fact.

digression

Incidentally, Gödel was lazy in the same way we are being now. At the end of the 1931 paper, he sketches the proof of the second incompleteness theorem, and promises the details in a later paper. He never got around to it; since everyone who understood the argument believed that it could be carried out (he did not need to fill in the details.)

inp.7 The Second Incompleteness Theorem

inc:inp:2in:sec

How can we express the assertion that **PA** doesn't prove its own consistency? Saying **PA** is inconsistent amounts to saying that $\mathbf{PA} \vdash 0 = 1$. So we can take the consistency statement $\mathsf{Con}_{\mathbf{PA}}$ to be the sentence $\neg\mathsf{Prov}_{\mathbf{PA}}(\ulcorner 0 = 1\urcorner)$, and then the following theorem does the job:

inc:inp:2in:thm:second-incompleteness

Theorem inp.8. *Assuming* **PA** *is consistent, then* **PA** *does not derive* $\mathsf{Con}_{\mathbf{PA}}$.

It is important to note that the theorem depends on the particular representation of $\mathsf{Con}_{\mathbf{PA}}$ (i.e., the particular representation of $\mathsf{Prov}_{\mathbf{PA}}(y)$). All we will use is that the representation of $\mathsf{Prov}_{\mathbf{PA}}(y)$ satisfies the three derivability conditions, so the theorem generalizes to any theory with a derivability predicate having these properties.

It is informative to read Gödel's sketch of an argument, since the theorem follows like a good punch line. It goes like this. Let $\gamma_{\mathbf{PA}}$ be the Gödel sentence that we constructed in the proof of Theorem inp.6. We have shown "If **PA** is consistent, then **PA** does not derive $\gamma_{\mathbf{PA}}$." If we formalize this *in* **PA**, we have a proof of

$$\mathsf{Con}_{\mathbf{PA}} \to \neg\mathsf{Prov}_{\mathbf{PA}}(\ulcorner\gamma_{\mathbf{PA}}\urcorner).$$

Now suppose **PA** derives $\mathsf{Con}_{\mathbf{PA}}$. Then it derives $\neg\mathrm{Prov}_{\mathbf{PA}}(\ulcorner\gamma_{\mathbf{PA}}\urcorner)$. But since $\gamma_{\mathbf{PA}}$ is a Gödel sentence, this is equivalent to $\gamma_{\mathbf{PA}}$. So **PA** derives $\gamma_{\mathbf{PA}}$.

But: we know that if **PA** is consistent, it doesn't derive $\gamma_{\mathbf{PA}}$! So if **PA** is consistent, it can't derive $\mathsf{Con}_{\mathbf{PA}}$.

To make the argument more precise, we will let $\gamma_{\mathbf{PA}}$ be the Gödel sentence for **PA** and use the derivability conditions (P1)–(P3) to show that **PA** derives $\mathsf{Con}_{\mathbf{PA}} \to \gamma_{\mathbf{PA}}$. This will show that **PA** doesn't derive $\mathsf{Con}_{\mathbf{PA}}$. Here is a sketch

of the proof, in **PA**. (For simplicity, we drop the **PA** subscripts.)

$\gamma \leftrightarrow \neg\mathsf{Prov}(\ulcorner\gamma\urcorner)$ (inp.5) inc:inp:2in:G2-1

γ is a Gödel sentence

$\gamma \to \neg\mathsf{Prov}(\ulcorner\gamma\urcorner)$ (inp.6) inc:inp:2in:G2-2

from eq. (inp.5)

$\gamma \to (\mathsf{Prov}(\ulcorner\gamma\urcorner) \to \bot)$ (inp.7) inc:inp:2in:G2-3

from eq. (inp.6) by logic

$\mathsf{Prov}(\ulcorner\gamma \to (\mathsf{Prov}(\ulcorner\gamma\urcorner) \to \bot)\urcorner)$ (inp.8) inc:inp:2in:G2-4

by from eq. (inp.7) by condition P1

$\mathsf{Prov}(\ulcorner\gamma\urcorner) \to \mathsf{Prov}(\ulcorner(\mathsf{Prov}(\ulcorner\gamma\urcorner) \to \bot)\urcorner)$ (inp.9) inc:inp:2in:G2-5

from eq. (inp.8) by condition P2

$\mathsf{Prov}(\ulcorner\gamma\urcorner) \to (\mathsf{Prov}(\ulcorner\mathsf{Prov}(\ulcorner\gamma\urcorner)\urcorner) \to \mathsf{Prov}(\ulcorner\bot\urcorner))$ (inp.10) inc:inp:2in:G2-6

from eq. (inp.9) by condition P2 and logic

$\mathsf{Prov}(\ulcorner\gamma\urcorner) \to \mathsf{Prov}(\ulcorner\mathsf{Prov}(\ulcorner\gamma\urcorner)\urcorner)$ (inp.11) inc:inp:2in:G2-7

by P3

$\mathsf{Prov}(\ulcorner\gamma\urcorner) \to \mathsf{Prov}(\ulcorner\bot\urcorner)$ (inp.12) inc:inp:2in:G2-8

from eq. (inp.10) and eq. (inp.11) by logic

$\mathsf{Con} \to \neg\mathsf{Prov}(\ulcorner\gamma\urcorner)$ (inp.13) inc:inp:2in:G2-9

contraposition of eq. (inp.12) and $\mathsf{Con} \equiv \neg\mathsf{Prov}(\ulcorner\bot\urcorner)$

$\mathsf{Con} \to \gamma$

from eq. (inp.5) and eq. (inp.13) by logic

The use of logic in the above just elementary facts from propositional logic, e.g., eq. (inp.7) uses $\vdash \neg\varphi\leftrightarrow(\varphi\to\bot)$ and eq. (inp.12) uses $\varphi\to(\psi\to\chi), \varphi\to\psi \vdash \varphi\to\chi$. The use of condition P2 in eq. (inp.9) and eq. (inp.10) relies on instances of P2, $\mathsf{Prov}(\ulcorner\varphi \to \psi\urcorner) \to (\mathsf{Prov}(\ulcorner\varphi\urcorner) \to \mathsf{Prov}(\ulcorner\psi\urcorner))$. In the first one, $\varphi \equiv \gamma$ and $\psi \equiv \mathsf{Prov}(\ulcorner\gamma\urcorner) \to \bot$; in the second, $\varphi \equiv \mathsf{Prov}(\ulcorner G\urcorner)$ and $\psi \equiv \bot$.

The more abstract version of the second incompleteness theorem is as follows:

Theorem inp.9. *Let* **T** *be any consistent, axiomatized theory extending* **Q** *and let* $\mathsf{Prov}_T(y)$ *be any formula satisfying derivability conditions P1–P3 for* **T**. *Then* **T** *does not derive* Con_T. inc:inp:2in:thm:second-incompleteness-gen

Problem inp.4. Show that **PA** derives $\gamma_{\mathbf{PA}} \to \mathsf{Con}_{\mathbf{PA}}$.

digression

The moral of the story is that no "reasonable" consistent theory for mathematics can derive its own consistency statement. Suppose **T** is a theory of mathematics that includes **Q** and Hilbert's "finitary" reasoning (whatever that may be). Then, the whole of **T** cannot derive the consistency statement of **T**, and so, a fortiori, the finitary fragment can't derive the consistency statement

of **T** either. In that sense, there cannot be a finitary consistency proof for "all of mathematics."

There is some leeway in interpreting the term "finitary," and Gödel, in the 1931 paper, grants the possibility that something we may consider "finitary" may lie outside the kinds of mathematics Hilbert wanted to formalize. But Gödel was being charitable; today, it is hard to see how we might find something that can reasonably be called finitary but is not formalizable in, say, **ZFC**, Zermelo–Fraenkel set theory with the axiom of choice.

inp.8 Löb's Theorem

inc:inp:lob: sec

The Gödel sentence for a theory **T** is a fixed point of $\neg\mathsf{Prov}_T(y)$, i.e., a sentence γ such that

$$\mathbf{T} \vdash \neg\mathsf{Prov}_T(\ulcorner\gamma\urcorner) \leftrightarrow \gamma.$$

It is not derivable, because if $\mathbf{T} \vdash \gamma$, (a) by derivability condition (1), $\mathbf{T} \vdash \mathsf{Prov}_T(\ulcorner\gamma\urcorner)$, and (b) $\mathbf{T} \vdash \gamma$ together with $\mathbf{T} \vdash \neg\mathsf{Prov}_T(\ulcorner\gamma\urcorner) \leftrightarrow \gamma$ gives $\mathbf{T} \vdash \neg\mathsf{Prov}_T(\ulcorner\gamma\urcorner)$, and so **T** would be inconsistent. Now it is natural to ask about the status of a fixed point of $\mathsf{Prov}_T(y)$, i.e., a sentence δ such that

$$\mathbf{T} \vdash \mathsf{Prov}_T(\ulcorner\delta\urcorner) \leftrightarrow \delta.$$

If it were derivable, $\mathbf{T} \vdash \mathsf{Prov}_T(\ulcorner\delta\urcorner)$ by condition (1), but the same conclusion follows if we apply modus ponens to the equivalence above. Hence, we don't get that **T** is inconsistent, at least not by the same argument as in the case of the Gödel sentence. This of course does not show that **T** *does* derive δ.

We can make headway on this question if we generalize it a bit. The left-to-right direction of the fixed point equivalence, $\mathsf{Prov}_T(\ulcorner\delta\urcorner) \to \delta$, is an instance of a general schema called a *reflection principle*: $\mathsf{Prov}_T(\ulcorner\varphi\urcorner) \to \varphi$. It is called that because it expresses, in a sense, that **T** can "reflect" about what it can derive; basically it says, "If **T** can derive φ, then φ is true," for any φ. This is true for sound theories only, of course, and this suggests that theories will in general not derive every instance of it. So which instances can a theory (strong enough, and satisfying the derivability conditions) derive? Certainly all those where φ itself is derivable. And that's it, as the next result shows.

inc:inp:lob: thm:lob

Theorem inp.10. *Let* **T** *be an axiomatizable theory extending* **Q**, *and suppose* $\mathsf{Prov}_T(y)$ *is a formula satisfying conditions P1–P3 from section inp.7. If* **T** *derives* $\mathsf{Prov}_T(\ulcorner\varphi\urcorner) \to \varphi$, *then in fact* **T** *derives* φ.

Put differently, if $\mathbf{T} \nvdash \varphi$, then $\mathbf{T} \nvdash \mathsf{Prov}_T(\ulcorner\varphi\urcorner) \to \varphi$. This result is known as Löb's theorem.

The heuristic for the proof of Löb's theorem is a clever proof that Santa Claus exists. (If you don't like that conclusion, you are free to substitute any other conclusion you would like.) Here it is: explanation

1. Let X be the sentence, "If X is true, then Santa Claus exists."

2. Suppose X is true.

3. Then what it says holds; i.e., we have: if X is true, then Santa Claus exists.

4. Since we are assuming X is true, we can conclude that Santa Claus exists, by modus ponens from (2) and (3).

5. We have succeeded in deriving (4), "Santa Claus exists," from the assumption (2), "X is true." By conditional proof, we have shown: "If X is true, then Santa Claus exists."

6. But this is just the sentence X. So we have shown that X is true.

7. But then, by the argument (2)–(4) above, Santa Claus exists.

A formalization of this idea, replacing "is true" with "is derivable," and "Santa Claus exists" with φ, yields the proof of Löb's theorem. The trick is to apply the fixed-point lemma to the formula $\mathsf{Prov}_T(y) \to \varphi$. The fixed point of that corresponds to the sentence X in the preceding sketch.

Proof of Theorem inp.10. Suppose φ is a sentence such that $\mathbf{T}$ derives $\mathsf{Prov}_T(\ulcorner\varphi\urcorner) \to \varphi$. Let $\psi(y)$ be the formula $\mathsf{Prov}_T(y) \to \varphi$, and use the fixed-point lemma to find a sentence θ such that $\mathbf{T}$ derives $\theta \leftrightarrow \psi(\ulcorner\theta\urcorner)$. Then each of the following

is derivable in $\mathbf{T}$:

inc:inp:lob: L-1
$$\theta \leftrightarrow (\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \varphi) \tag{inp.14}$$
θ is a fixed point of $\psi(y)$

inc:inp:lob: L-2
$$\theta \to (\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \varphi) \tag{inp.15}$$
from eq. (inp.14)

inc:inp:lob: L-3
$$\mathsf{Prov}_T(\ulcorner\theta \to (\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \varphi)\urcorner) \tag{inp.16}$$
from eq. (inp.15) by condition P1

inc:inp:lob: L-4
$$\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \mathsf{Prov}_T(\ulcorner\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \varphi\urcorner) \tag{inp.17}$$
from eq. (inp.16) using condition P2

inc:inp:lob: L-5
$$\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to (\mathsf{Prov}_T(\ulcorner\mathsf{Prov}_T(\ulcorner\theta\urcorner)\urcorner) \to \mathsf{Prov}_T(\ulcorner\varphi\urcorner)) \tag{inp.18}$$
from eq. (inp.17) using P2 again

inc:inp:lob: L-6
$$\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \mathsf{Prov}_T(\ulcorner\mathsf{Prov}_T(\ulcorner\theta\urcorner)\urcorner) \tag{inp.19}$$
by derivability condition P3

inc:inp:lob: L-7
$$\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \mathsf{Prov}_T(\ulcorner\varphi\urcorner) \tag{inp.20}$$
from eq. (inp.18) and eq. (inp.19)

inc:inp:lob: L-8
$$\mathsf{Prov}_T(\ulcorner\varphi\urcorner) \to \varphi \tag{inp.21}$$
by assumption of the theorem

inc:inp:lob: L-9
$$\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \varphi \tag{inp.22}$$
from eq. (inp.20) and eq. (inp.21)

inc:inp:lob: L-10
$$(\mathsf{Prov}_T(\ulcorner\theta\urcorner) \to \varphi) \to \theta \tag{inp.23}$$
from eq. (inp.14)

inc:inp:lob: L-11
$$\theta \tag{inp.24}$$
from eq. (inp.22) and eq. (inp.23)

inc:inp:lob: L-12
$$\mathsf{Prov}_T(\ulcorner\theta\urcorner) \tag{inp.25}$$
from eq. (inp.24) by condition P1

φ from eq. (inp.21) and eq. (inp.25) □

With Löb's theorem in hand, there is a short proof of the second incompleteness theorem (for theories having a derivability predicate satisfying conditions P1–P3): if $\mathbf{T} \vdash \mathsf{Prov}_T(\ulcorner\bot\urcorner) \to \bot$, then $\mathbf{T} \vdash \bot$. If $\mathbf{T}$ is consistent, $\mathbf{T} \nvdash \bot$. So, $\mathbf{T} \nvdash \mathsf{Prov}_T(\ulcorner\bot\urcorner) \to \bot$, i.e., $\mathbf{T} \nvdash \mathsf{Con}_{\mathbf{T}}$. We can also apply it to show that δ, the fixed point of $\mathsf{Prov}_T(x)$, is derivable. For since

$$\mathbf{T} \vdash \mathsf{Prov}_T(\ulcorner\delta\urcorner) \leftrightarrow \delta$$

in particular

$$\mathbf{T} \vdash \mathsf{Prov}_T(\ulcorner\delta\urcorner) \to \delta$$

and so by Löb's theorem, $\mathbf{T} \vdash \delta$.

Problem inp.5. Let **T** be a computably axiomatized theory, and let Prov_T be a derivability predicate for **T**. Consider the following four statements:

1. If $T \vdash \varphi$, then $T \vdash \mathsf{Prov}_T(\ulcorner\varphi\urcorner)$.
2. $T \vdash \varphi \to \mathsf{Prov}_T(\ulcorner\varphi\urcorner)$.
3. If $T \vdash \mathsf{Prov}_T(\ulcorner\varphi\urcorner)$, then $T \vdash \varphi$.
4. $T \vdash \mathsf{Prov}_T(\ulcorner\varphi\urcorner) \to \varphi$

Under what conditions are each of these statements true?

inp.9 The Undefinability of Truth

inc:inp:tar: sec

The notion of *definability* depends on having a formal semantics for the language of arithmetic. We have described a set of formulas and sentences in the language of arithmetic. The "intended interpretation" is to read such sentences as making assertions about the natural numbers, and such an assertion can be true or false. Let $\mathfrak{N}$ be the structure with domain $\mathbb{N}$ and the standard interpretation for the symbols in the language of arithmetic. Then $\mathfrak{N} \vDash \varphi$ means "φ is true in the standard interpretation."

Definition inp.11. A relation $R(x_1, \ldots, x_k)$ of natural numbers is *definable* in $\mathfrak{N}$ if and only if there is a formula $\varphi(x_1, \ldots, x_k)$ in the language of arithmetic such that for every $n_1, \ldots, n_k$, $R(n_1, \ldots, n_k)$ if and only if $\mathfrak{N} \vDash \varphi(\overline{n}_1, \ldots, \overline{n}_k)$.

Put differently, a relation is definable in $\mathfrak{N}$ if and only if it is representable in the theory **TA**, where $\mathbf{TA} = \{\varphi : \mathfrak{N} \vDash \varphi\}$ is the set of true sentences of arithmetic. (If this is not immediately clear to you, you should go back and check the definitions and convince yourself that this is the case.)

Lemma inp.12. *Every computable relation is definable in* $\mathfrak{N}$*.*

Proof. It is easy to check that the formula representing a relation in **Q** defines the same relation in $\mathfrak{N}$. □

Now one can ask, is the converse also true? That is, is every relation definable in $\mathfrak{N}$ computable? The answer is no. For example:

Lemma inp.13. *The halting relation is definable in* $\mathfrak{N}$*.*

Proof. Let H be the halting relation, i.e.,

$$H = \{\langle e, x\rangle : \exists s\, T(e, x, s)\}.$$

Let θ_T define T in $\mathfrak{N}$. Then

$$H = \{\langle e, x\rangle : \mathfrak{N} \vDash \exists s\, \theta_T(\overline{e}, \overline{x}, s)\},$$

so $\exists s\, \theta_T(z, x, s)$ defines H in $\mathfrak{N}$. □

Problem inp.6. Show that $Q(n) \Leftrightarrow n \in \{^{\ulcorner}\varphi^{\urcorner} : \mathbf{Q} \vdash \varphi\}$ is definable in arithmetic.

What about **TA** itself? Is it definable in arithmetic? That is: is the set $\{^{\ulcorner}\varphi^{\urcorner} : \mathfrak{N} \vDash \varphi\}$ definable in arithmetic? Tarski's theorem answers this in the negative.

inc:inp:tar: thm:tarski

Theorem inp.14. *The set of true sentences of arithmetic is not definable in arithmetic.*

Proof. Suppose $\theta(x)$ defined it, i.e., $\mathfrak{N} \vDash \varphi$ iff $\mathfrak{N} \vDash \theta(^{\ulcorner}\varphi^{\urcorner})$. By the fixed-point lemma, there is a formula φ such that $\mathbf{Q} \vdash \varphi \leftrightarrow \neg\theta(^{\ulcorner}\varphi^{\urcorner})$, and hence $\mathfrak{N} \vDash \varphi \leftrightarrow \neg\theta(^{\ulcorner}\varphi^{\urcorner})$. But then $\mathfrak{N} \vDash \varphi$ if and only if $\mathfrak{N} \vDash \neg\theta(^{\ulcorner}\varphi^{\urcorner})$, which contradicts the fact that $\theta(y)$ is supposed to define the set of true statements of arithmetic. □

Tarski applied this analysis to a more general philosophical notion of truth. Given any language L, Tarski argued that an adequate notion of truth for L would have to satisfy, for each sentence X,

> 'X' is true if and only if X.

Tarski's oft-quoted example, for English, is the sentence

> 'Snow is white' is true if and only if snow is white.

However, for any language strong enough to represent the diagonal function, and any linguistic predicate $T(x)$, we can construct a sentence X satisfying "X if and only if not $T(\text{'}X\text{'})$." Given that we do not want a truth predicate to declare some sentences to be both true and false, Tarski concluded that one cannot specify a truth predicate for all sentences in a language without, somehow, stepping outside the bounds of the language. In other words, a the truth predicate for a language cannot be defined in the language itself.

Photo Credits

Bibliography