The Second Incompleteness Theorem

How can we express the assertion that PA doesn’t prove its own consistency? Saying PA is inconsistent amounts to saying that PA ⊢ 0 = 1. So we can take the consistency statement ConPA to be the sentence ¬ProvPA(⌜0 = 1⌝), and then the following theorem does the job:

**Theorem inp.1.** Assuming PA is consistent, then PA does not derive ConPA.

It is important to note that the theorem depends on the particular representation of ConPA (i.e., the particular representation of ProvPA(y)). All we will use is that the representation of ProvPA(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 γPA be the Gödel sentence that we constructed in the proof of ?? (P2). We have shown “If PA is consistent, then PA does not derive γPA.” If we formalize this in PA, we have a proof of

\[ \text{Con}_{PA} \rightarrow \neg \text{Prov}_{PA}(\neg \gamma_{PA}). \]

Now suppose PA derives ConPA. Then it derives ¬ProvPA(⌜γPA⌝). But since γPA is a Gödel sentence, this is equivalent to γPA. So PA derives γPA.

But: we know that if PA is consistent, it doesn’t derive γPA! So if PA is consistent, it can’t derive ConPA.

To make the argument more precise, we will let γPA be the Gödel sentence for PA and use the derivability conditions (P1)-(P3) to show that PA derives ConPA → γPA. This will show that PA doesn’t derive ConPA. Here is a sketch
of the proof, in \textbf{PA}. (For simplicity, we drop the \textbf{PA} subscripts.)

\[
\begin{align*}
\gamma \iff \neg \text{Prov}(\gamma) & \quad \hspace{2cm} (1) \\
\gamma \to \neg \text{Prov}(\gamma) & \quad \hspace{2cm} \text{from eq. (1)} \\
\gamma \to (\text{Prov}(\neg \gamma) \to \bot) & \quad \hspace{2cm} \text{from eq. (2) by logic} \\
\text{Prov}(\neg \gamma \to (\text{Prov}(\neg \gamma) \to \bot)) & \quad \hspace{2cm} \text{by from eq. (1) by condition P1} \\
\text{Prov}(\neg \gamma) \to \text{Prov}(\neg \text{Prov}(\neg \gamma)) & \quad \hspace{2cm} \text{by P3} \\
\text{Prov}(\neg \gamma) \to \text{Prov}(\neg \bot) & \quad \hspace{2cm} \text{from eq. (5) by condition P2} \\
\text{Con} \to \neg \text{Prov}(\neg \gamma) & \quad \hspace{2cm} \text{contraposition of eq. (8) and } \text{Con} \equiv \neg \text{Prov}(\bot) \\
\text{Con} \to \gamma & \quad \hspace{2cm} \text{from eq. (1) and eq. (9) by logic}
\end{align*}
\]

The use of logic in the above just elementary facts from propositional logic, e.g., eq. (3) uses \( \vdash \neg \varphi \leftrightarrow (\varphi \to \bot) \) and eq. (8) uses \( \varphi \to (\psi \to \chi), \varphi \to \psi \vdash \varphi \to \chi \). The use of condition P2 in eq. (5) and eq. (6) relies on instances of P2, \( \text{Prov}(\varphi \to \psi) \to (\text{Prov}(\varphi) \to \text{Prov}(\psi)) \). In the first one, \( \varphi \equiv \gamma \) and \( \psi \equiv \text{Prov}(\neg \gamma) \to \bot \); in the second, \( \varphi \equiv \text{Prov}(G) \) and \( \psi \equiv \bot \).

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

\textbf{Theorem inp.2.} \textit{Let } \mathbf{T} \textit{ be any consistent, axiomatized theory extending } \mathbf{Q} \textit{ and let } \text{Prov}_T(y) \textit{ be any formula satisfying derivability conditions P1–P3 for } \mathbf{T}. \textit{Then } \mathbf{T} \textit{ does not derive } \text{Con}_T. \hspace{2cm} \text{thm:second-incompleteness-gen}

\textbf{Problem inp.1.} \textit{Show that } \mathbf{PA} \textit{ derives } \gamma_{\mathbf{PA}} \to \text{Con}_{\mathbf{PA}}. \hspace{2cm} \text{inc:inp-2in:2thm}

\textbf{digression} The moral of the story is that no “reasonable” consistent theory for mathematics can derive its own consistency statement. Suppose \( \mathbf{T} \) is a theory of mathematics that includes \( \mathbf{Q} \) and Hilbert’s “finitary” reasoning (whatever that may be). Then, the whole of \( \mathbf{T} \) cannot derive the consistency statement of \( \mathbf{T} \), 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, $\text{ZFC}$, Zermelo–Fraenkel set theory with the axiom of choice.

**Photo Credits**

**Bibliography**