The Gödel sentence for a theory $T$ is a fixed point of $\neg \text{Prov}_T(x)$, i.e., a sentence $\gamma$ such that

$$T \vdash \neg \text{Prov}_T(\gamma) \leftrightarrow \gamma.$$  

It is not provable, because if $T \vdash \gamma$, (a) by provability condition (1), $T \vdash \text{Prov}_T(\gamma)$, and (b) $T \vdash \gamma$ together with $T \vdash \neg \text{Prov}_T(\gamma) \leftrightarrow \gamma$ gives $T \vdash \neg \text{Prov}_T(\gamma)$, and so $T$ would be inconsistent. Now it is natural to ask about the status of a fixed point of $\text{Prov}_T(x)$, i.e., a sentence $\delta$ such that

$$T \vdash \text{Prov}_T(\delta) \leftrightarrow \delta.$$  

If it were provable, $T \vdash \text{Prov}_T(\delta)$ 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 prove $\delta$.

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

**Theorem inp.1.** Let $T$ be an axiomatizable theory extending $Q$, and suppose $\text{Prov}_T(y)$ is a formula satisfying conditions P1–P3 from ?? If $T$ proves $\text{Prov}_T(\neg \varphi) \rightarrow \varphi$, then in fact $T$ proves $\varphi$.

Put differently, if $T \not\vdash \varphi$, then $T \not\vdash \text{Prov}_T(\neg \varphi) \rightarrow \varphi$. 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:

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 provable,” and “Santa Claus exists” with $\varphi$, yields the proof of L"ob's theorem. The trick is to apply the fixed-point lemma to the formula $\text{Prov}_T(y) \rightarrow \varphi$. The fixed point of that corresponds to the sentence $X$ in the preceding sketch.

Proof. Suppose $\varphi$ is a sentence such that $T$ proves $\text{Prov}_T(\neg\varphi) \rightarrow \varphi$. Let $\psi(y)$ be the formula $\text{Prov}_T(y) \rightarrow \varphi$, and use the fixed-point lemma to find a sentence $\theta$ such that $T$ proves $\theta \leftrightarrow \psi(\theta)$. Then each of the following is provable in $T$:

\[
\begin{align*}
\theta & \leftrightarrow (\text{Prov}_T(\neg\theta) \rightarrow \varphi) \tag{1} \\
& \quad \text{is a fixed point of $\psi$} \tag{L-1}
\end{align*}
\]

\[
\begin{align*}
\theta & \rightarrow (\text{Prov}_T(\neg\theta) \rightarrow \varphi) \tag{2} \\
& \quad \text{from eq. (1)} \tag{L-2}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\theta) & \rightarrow (\text{Prov}_T(\neg\theta) \rightarrow \varphi) \tag{3} \\
& \quad \text{from eq. (2) by condition P1} \tag{L-3}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\theta) & \rightarrow \text{Prov}_T(\text{Prov}_T(\neg\theta) \rightarrow \varphi) \tag{4} \\
& \quad \text{from eq. (3) using condition P2} \tag{L-4}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\theta) & \rightarrow (\text{Prov}_T(\text{Prov}_T(\neg\theta) \rightarrow \varphi)) \tag{5} \\
& \quad \text{from eq. (4) using P2 again} \tag{L-5}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\theta) & \rightarrow \text{Prov}_T(\text{Prov}_T(\neg\theta) \neg\varphi) \tag{6} \\
& \quad \text{by provability condition P3} \tag{L-6}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\theta) & \rightarrow \text{Prov}_T(\neg\varphi) \tag{7} \\
& \quad \text{from eq. (5) and eq. (6)} \tag{L-7}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\varphi) & \rightarrow \varphi \tag{8} \\
& \quad \text{by assumption of the theorem} \tag{L-8}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\theta) & \rightarrow \varphi \tag{9} \\
& \quad \text{from eq. (7) and eq. (8)} \tag{L-9}
\end{align*}
\]

\[
\begin{align*}
(\text{Prov}_T(\neg\theta) \rightarrow \varphi) & \rightarrow \theta \tag{10} \\
& \quad \text{from eq. (1)} \tag{L-10}
\end{align*}
\]

\[
\begin{align*}
\theta & \tag{11} \\
& \quad \text{from eq. (9) and eq. (10)} \tag{L-11}
\end{align*}
\]

\[
\begin{align*}
\text{Prov}_T(\neg\theta) & \tag{12} \\
& \quad \text{from eq. (11) by condition P1} \tag{L-12}
\end{align*}
\]

\[
\begin{align*}
\varphi & \tag{13} \\
& \quad \text{from eq. (8) and eq. (12)} \tag{L-13}
\end{align*}
\]

\[\square\]
With Löb’s theorem in hand, there is a short proof of the first incompleteness theorem (for theories having a provability predicate satisfying conditions P1–P3: if $T \vdash \text{Prov}_T(\neg \perp) \rightarrow \perp$, then $T \vdash \perp$. If $T$ is consistent, $T \nvdash \perp$. So, $T \nvdash \text{Prov}_T(\neg \perp) \rightarrow \perp$, i.e., $T \nvdash \text{Con}_T$. We can also apply it to show that $\delta$, the fixed point of $\text{Prov}_T(x)$, is provable. For since

$$T \vdash \text{Prov}_T(\neg \delta) \leftrightarrow \delta$$

in particular

$$T \vdash \text{Prov}_T(\neg \delta) \rightarrow \delta$$

and so by Löb’s theorem, $T \vdash \delta$.

**Problem inp.1.** Let $T$ be a computably axiomatized theory, and let $\text{Prov}_T$ be a provability predicate for $T$. Consider the following four statements:

1. If $T \vdash \varphi$, then $T \vdash \text{Prov}_T(\neg \varphi)$.
2. $T \vdash \varphi \rightarrow \text{Prov}_T(\neg \varphi)$.
3. If $T \vdash \text{Prov}_T(\neg \varphi)$, then $T \vdash \varphi$.
4. $T \vdash \text{Prov}_T(\neg \varphi) \rightarrow \varphi$

Under what conditions are each of these statements true?

**Photo Credits**

**Bibliography**