2. INCOMPLETENESS

The methods of arithmetization and self–reference were originally used to prove incompleteness theorems for arithmetical theories. In this chapter we present the most important theorems of this type.

A sentence φ (in the language of S) is *undecidable* in S if $S\nvdash \varphi$ and $S\nvdash \neg\varphi$. S is *complete* if no sentence is undecidable in S, otherwise *incomplete*.

§1. Incompleteness. We begin with the first and most important result of the whole subject, Gödel's incompleteness theorem (for theories in L_A).

Theorem 1. Let φ be a Π_1 sentence such that
(G) $\quad Q\vdash \varphi \leftrightarrow \neg Pr_T(\varphi)$.
Then φ is true and $T\nvdash \varphi$. Thus, if T is Σ_1–sound, then also $T\nvdash \neg\varphi$.

Proof. Suppose $T\vdash \varphi$. Then, by Fact 7 (b), $Q\vdash Pr_T(\varphi)$. But then, by (G), $Q\vdash \neg\varphi$ and so $T\vdash \neg\varphi$. It follows that T is inconsistent, contrary to Convention 2. Thus, $T\nvdash \varphi$. By (G), φ is true. Thus, $\neg\varphi$ is a false Σ_1 sentence and so $T\nvdash \neg\varphi$ if T is Σ_1–sound. ■

Notice the close similarity between the proofs of Theorem 1, Lemma 1.2, and Theorem 1.3 (the liar paradox).

To derive the conclusion that $T\nvdash \neg\varphi$ in Theorem 1, we needed the assumption that T is Σ_1–sound. We can now see that this is stronger than mere consistency: $T + \neg\varphi$ is consistent but not Σ_1–sound. (Note that it does not follow from Theorem 1 that $T + \neg\varphi$ is incomplete.) Thus, the question arises if, assuming consistency only, there is a (Π_1) sentence which is undecidable in T. Our next result, known as Rosser's theorem, shows that the answer is affirmative.

Theorem 2. Let θ be a Π_1 sentence such that
(R) $\quad Q\vdash \theta \leftrightarrow \forall z(Prf_T(\theta,z) \rightarrow \exists u{\le}z Prf_T(\neg\theta,u))$.
Then θ is undecidable in T.

Proof. We first prove that $T\nvdash \theta$. Suppose, for *reductio ad absurdum*, $T\vdash \theta$ and let p be a proof of θ in T. Then, by Fact 7 (a),
(1) $\quad Q\vdash Prf_T(\theta,p)$.
Since T is consistent, we have $T\nvdash \neg\theta$. By Fact 7 (d), $Q\vdash \neg Prf_T(\neg\theta,q)$ for every q. But then, by Fact 1 (iv),

$$Q\vdash u \le p \rightarrow \neg Prf_T(\neg\theta,u).$$

Combining this with (1) we get

$$Q\vdash \exists z(Prf_T(\theta,z) \wedge \forall u{\le}z\neg Prf_T(\neg\theta,u)).$$

But then, by (R), $Q\vdash \neg\theta$ and so $T\vdash \neg\theta$, a contradiction. Thus, $T\nvdash \theta$ as desired.