APPENDIX 2

THE GÖDEL INCOMPLETENESS THEOREMS

We sketch proofs of Gödel's theorems, obtaining along the way an important result of Tarski on the undefinability of mathematical truth.

We begin by setting up a *formal language* in which arithmetical statements can be written down in a precise way. This language will be denoted by the symbol **L** and called the *language of arithmetic*. As is the case with any language, we must first specify its *alphabet*: that of **L** consists of the following *symbols*:

Arithmetical variables: $x_1,\ x_2,\ x_3, \ldots$
Arithmetical constants: $\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots$
Arithmetical operation symbols: $+, \times$
Equality symbol: $=$
Logical operators: $\wedge$ (*conjunction, "and"*)
$\vee$ (*disjunction, "or"*)
$\neg$ (*negation, "not"*),
$\rightarrow$ (*implication, "if … then"*)
$\leftrightarrow$ (*bi-implication, "is equivalent to"*)
$\forall$ (*universal quantifier, "for all"*)
$\exists$ (*existential quantifier, "there exists"*)
Punctuation symbols: (,) , [,], commas, etc.

Expressions of **L**—*arithmetical expressions*—are built up by stringing together finite sequences of symbols. As in any language, only certain expressions of **L** will be deemed meaningful or *well-formed*. These are the *terms,* the arithmetical counterparts of nouns, and the *formulas*, the arithmetical counterparts of declarative assertions.

Arithmetical terms, or simply *terms,* are specified by means of the following rules:

(i) Any arithmetical variable or constant standing by itself is a term.

(ii) If $\boldsymbol{t}$ and $\boldsymbol{u}$ are terms, so too are the expressions $\boldsymbol{t + u}, \boldsymbol{t \times u}$.

(iii) An expression is a term when, and only when, it follows that it is one from finitely many applications of clauses (i) and (ii).

Thus, for example, each of the expressions $\mathbf{6}$, $x_1 + \mathbf{2}$ and $(x_1 \times x_5) + \mathbf{27}$ is a term.

Arithmetical formulas, or simply *formulas,* are specified by means of the following rules:

(a) For any terms ***t, u,*** the expression $\boldsymbol{t = u}$ is a formula.

(b) If A and B are formulas, so too are all the expressions $\neg A$, $A \wedge B$, $A \vee B$, $A \rightarrow B$, $A \leftrightarrow B$, $\forall x_n A$, $\exists x_n A$, for any numerical variable x_n.

(c) An expression is a formula when, and only when, it follows that it is one from finitely many applications of clauses (a) and (b).

Thus, for instance, each of the following expressions is a formula: $x_1 + \mathbf{2} = \mathbf{3}$, $\exists x_1(x_1 + \mathbf{2} = \mathbf{3})$, $\forall x_1 \forall x_5(x_1 \times x_5 = \mathbf{3})$.

Terms assume *numerical values* and formulas *truth values* (truth or falsehood) when their constituent symbols are interpreted in the natural way. In this natural interpretation, each arithmetical variable is assigned some arbitrary but fixed integer as value, and then each arithmetical constant $\boldsymbol{n}$ is interpreted as the corresponding integer n, the arithmetical operation symbols + and × as addition and multiplication of integers, respectively, the equality symbol as identity of integers, and finally the logical operators as the corresponding logical particles of ordinary discourse. For example, if we assign the values 3 to x_1 and 7 to x_5 , then the resulting values assigned to the three terms above are, respectively, 6; 3 + 2, i.e., 5; and (3 × 7) + 27, i.e., 48. Under the same assignment of values, the resulting truth values assigned to the three formulas above are, respectively,

THE TRUTH VALUE OF THE STATEMENT $3 + 2 = 3$, I.E., *FALSEHOOD;*
THE TRUTH VALUE OF THE STATEMENT *THERE IS A NUMBER WHOSE SUM WITH* 2 *EQUALS* 3, I.E., *TRUTH* ;
THE TRUTH VALUE OF THE STATEMENT *THE PRODUCT OF ANY PAIR OF NUMBERS EQUALS* 3, I.E., *FALSEHOOD.*

In place of the clumsy locution "the truth value of A is truth (or falsehood)" we shall usually employ the phrase "A is *true* (or *false*)." Note that then a formula A is true exactly when its negation $\neg A$ is false.

We observe that, while the truth or falsehood of the first of these formulas is dependent on the values assigned to the variables occurring in it (in this case, just x_1), the truth values of the second two are *independent* of the values assigned to such variables. Formulas having this independence property are called *sentences*: they may be regarded as making simple declarative assertions—either true or false— about the system of natural numbers. Formally speaking, a sentence is a formula in which each occurrence of a variable x is accompanied by the occurrence of a corresponding "quantifier" expression of the form $\forall x$ or $\exists x$.

Occurrences of variables in formulas not accompanied by a corresponding quantifier expression are called *free* occurrences: for example, the occurrence of the variable x_1 in the first formula above is free, but those in the second and third formulas are not. We write $A(x_1 \ldots, x_n)$ for any formula A in which at most the variables $x_1, \ldots, x_n$ have free occurrences, and, for any natural numbers $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n$, we write $A(\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)$ for the formula (evidently a sentence) obtained by substituting $\boldsymbol{m}_1$ for $x_1, \ldots, \boldsymbol{m}_n$ for x_n at each of the latter's free occurrences in A. Thus, for example, if $A(x_1, x_2, x_3, x_4)$ is the formula

$$\exists x_1(x_1 + x_2 = \mathbf{4} \ \wedge x_3 \times x_4 = \mathbf{7}),$$

then $A($**1, 5, 7, 8**$)$ is the sentence

$$\exists x_1(x_1 + \mathbf{5} = \mathbf{4} \ \wedge \ \mathbf{7} \times \mathbf{8} = \mathbf{7}).$$

We shall employ similar notational conventions for terms.

We next assign *numerical labels* to the symbols of the language of arithmetic in the following way. Suppose that its symbols, excluding variables and constants, are k in number. To these symbols we assign, in some initially arbitrary but subsequently fixed manner, the labels 0, 1, ..., $k - 1$. Then to each numerical variable x_n we assign the label $k + 2n$ and to each numerical constant $\boldsymbol{n}$ the label $k + 2n + 1$. In this way each symbol s is assigned a label which we shall denote by s^*.

Finally, each expression $s_1 s_2 ... s_n$ is assigned the *code number*

$$2^{s^*} \times 3^{s^*} \times ... \times p_n^{s^*},$$

where p_n is the n^{th} prime number. In this way each expression is assigned a unique positive integer as its code, and, conversely, every positive integer is the code of some unique expression.

We shall use the symbol A_n to denote the arithmetical expression with code number n.

Let P be a property of natural numbers: we write $P(m)$ to indicate that the number m has the property P. Similarly, if R is a relation among natural numbers, we write $R(m_1, ..., m_n)$ to indicate that the numbers $m_1, ..., m_n$ stand in the relation R. We shall often use the term "relation" to cover properties as well.

A relation R among natural numbers is called *arithmetically definable* if there is an arithmetical formula $A(x_1, ..., x_n)$ such that, for all numbers $m_1, ..., m_n$, we have

$$R(m_1, ..., m_n) \text{ iff } [1] \text{ the sentence } A(\boldsymbol{m_1, ..., m_n}) \text{ is true.}$$

In this case we say that the relation R *is defined by* the formula A. We extend this concept to arithmetical *expressions* by saying that a property of (or a relation among) such expressions is arithmetically *definable* if the corresponding property of (or relation among) their *code numbers* is so definable**.**

Now it can be shown without much difficulty that the property of being (the code number of) an arithmetical *formula*, or a *sentence,* is arithmetically definable. But what about the property of being a *true* sentence? We shall establish the remarkable result that this property is *not* arithmetically definable**.**

Since the assignment of code numbers to arithmetical expressions is evidently a wholly mechanical process, it is possible to compute, for any given formula[2] $A_m(x_1)$

[1] We use "iff" as an abbreviation for the phrase "if and only if", that is, "is equivalent to".

[2] Recall that A_m stands for the formula with code number m. In writing $A_m(x_1)$ we are, accordingly, assuming that A_m has free occurrences of at most the variable x_1.

with code number *m,* and any number *n,* the code number of the sentence $A_m(\boldsymbol{n})$. This computation is in turn arithmetically representable in the sense that one can construct an arithmetical term[3] $\boldsymbol{s}(x_1, x_2)$ with the property that, for any numbers *m, n, p*,

the sentence $\boldsymbol{s}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{p}$ *is true iff p is the code number of the sentence* $A_m(\boldsymbol{n})$.

Now let *S* any collection of arithmetical sentences. We proceed to prove the

Arithmetical Truth Theorem. Suppose that *S* satisfies the following conditions:

(i)Each member of *S* is true.
(ii) The property of being (the code number of) a member of *S* is arithmetically definable.

Then there is a *true* sentence *G* of **L** such that neither *G* nor its negation ¬*G* are members of *S.*

Proof. By assumption (ii) there is a formula $T(x_1)$ of **L** such that, for all numbers *n*,

$T(\boldsymbol{n})$ is true iff *n* is the code number of a sentence in *S*

iff A_n is in *S*.

Write $B(x_1)$ for the formula $\neg T(\boldsymbol{s}(x_1, x_1))$ (i.e., the result of substituting $\boldsymbol{s}(x_1, x_1)$ for all free occurrences of x_1 in $\neg T(x_1)$) and suppose that *B* has code number *m*. Then

B is the sentence A_m.

Next, let *p* be the natural number such that

$$\boldsymbol{p} = \boldsymbol{s}(m, m)$$

is a true sentence. Then, *by definition, p* is the code number of the sentence $A_m(\boldsymbol{m})$.

Now write *G* for $A_m(\boldsymbol{m})$. Then *p* is the code number of *G*, or, in other words,

G is A_p.

Thus we have

G is true iff $A_m(\boldsymbol{m})$.is true

iff $B(\boldsymbol{m})$ is true

[3]Here "s" stands for "substitution."

iff $\neg T(\boldsymbol{s}(\boldsymbol{m}, \boldsymbol{m}))$ is true

iff $T(\boldsymbol{p})$ is false

iff A_p is not in S

iff G is not in S.

We see from this G asserts of itself that it is not in S. It follows that G is true, for, if it were false, it would follow from the above that it was in S, and hence true by assumption (i). Since G is now true, again by the above it cannot be a member of S. Finally, since $\neg G$ must now be false, it cannot be a member of S since by assumption every member of the latter is true. The proof is complete.

By taking S in this theorem to be the collection of *all true arithmetical sentences*, we immediately obtain

Tarski's Theorem on the Undefinability of Truth. The property of being a true arithmetical sentence is not arithmetically definable.

We observe that the relevant sentence G in Tarski's theorem asserts "I am not in the set of true sentences", i.e., "I am false". Thus, like the sentence in the Liar paradox, G asserts its own falsehood.

Next, one can formulate the notion of a *proof* from a set of arithmetical sentences S and that of a formula *provable* from S in such a way that:

(1) if each member of S is true, so is each sentence provable from S;
(2) if the property of being a member of S is arithmetically definable, so is the property of being a sentence provable from S.

Then from the Arithmetical Truth Theorem one infers

Gödel's First Incompleteness Theorem (weak form). Let S be a set of true arithmetical sentences and suppose that the property of being a member of S is arithmetically definable. Then S is *incomplete*, i.e. there is a (true) arithmetical sentence G such that neither it nor its negation are provable from S.

To prove this we define $\overline{S}$ to be the set of all sentences provable from S. Then by (1) and (2) above, $\overline{S}$ consists of true sentences and the property of being a member of $\overline{S}$ is arithmetically definable. Accordingly we may apply the Arithmetical Truth Theorem to $\overline{S}$: this yields a (true) sentence G such that neither G nor $\neg G$ are members of $\overline{S}$, in other words, neither are provable from S.

The sentence G here will be seen to assert, not its own falsehood, but its own *unprovability* from S.

The import of this theorem may be stated in the following way. Suppose we think of our set S as a possible set of *axioms* for arithmetic, from which one might hope to be able to infer (at least in principle) all arithmetical truths. Then the theorem shows that one will *never* be able to construct within any language like **L** a set S of axioms for arithmetic which is *sound,* in the sense that each of its members is true, arithmetically *definable*, so that we know what to put in it, and *complete*, so that it can be used to prove or refute[4] any arithmetical sentence. It is possible for S to possess *any two* of these properties, *but not all three at once*.

By refining this argument one can significantly strengthen its conclusion. Let us call S *consistent* if no formula of the form $A \wedge \neg A$ (a *contradiction*) is provable from S. Let R be a relation defined by a formula A. We say that R is *S-definite* if, for any natural numbers $m_1, ..., m_n$, we have

$$R(m_1, ..., m_n) \text{ iff } A(\boldsymbol{m}_1, ..., \boldsymbol{m}_n) \text{ is provable from } S$$
$$\textit{not } R(m_1, ..., m_n) \text{ iff } \neg A(\boldsymbol{m}_1, ..., \boldsymbol{m}_n) \text{ is provable from } S.$$

Let Q be the *substitution relation* among natural numbers, that is, the relation which obtains among those triples m_1, m_2, m_3 of numbers for which the sentence $\boldsymbol{s}(\boldsymbol{m}_1, \boldsymbol{m}_2) = \boldsymbol{m}_3$ is true. Then one can prove

Gödel's First Incompleteness Theorem (strong form). Suppose that S is consistent, the property of being a member of S is arithmetically definable, and the property Q and the property of being a formula provable from S are both S-definite. Then S is incomplete.

We sketch the proof of this theorem. As before, we take G to be an arithmetical sentence which asserts its own unprovability from S. For any formula A, let us write $\vdash_S A$ for "A is provable from S", and $\nvdash_S A$ for "A is not provable from S."

Suppose that $\vdash_S G$. Then because provability from S is S-definite, it follows that

$$\vdash_S \text{ "}G \text{ is provable from } S\text{"}.$$

But the assertion "G is provable from S" is essentially just $\neg G$ (since G is essentially "G is unprovable from S"), so we get

$$\vdash_S \neg G,$$

contradicting the supposed consistency of S. Therefore $\nvdash_S G$.

Now suppose that $\vdash_S \neg G$. Then since S is consistent, it follows that $S \nvdash_S G$ and because provability from S is S-definite, we get

[4]We say that a sentence is *refutable* if its negation is provable.

$$\vdash_S \text{“}G \text{ is unprovable from } S\text{”}.$$

Noting again that G is essentially the assertion “G is unprovable from S", it then follows that

$$\vdash_S G,$$

contradicting the consistency of S. Hence also $\nvdash_S \neg G$, and we are done.

The advantage of this strong form of the incompleteness theorem is that in it the “external” requirement that all the members of S be *true* has been replaced by the much weaker “internal” requirement that S be merely *consistent*. We may sum it up by saying that *any consistent definable set of axioms for arithmetic must be incomplete.*

In sketching the proof of this last theorem we established the implication

$$\textit{if } S \text{ is consistent, } \textit{then } \nvdash_S G. \qquad (*)$$

Now the assertion “S is consistent” can be expressed as an arithmetical sentence Con_S in the following way. Let n_0 be the code number of some demonstrably false sentence, $\mathbf{0 = 1}$, say, and write $P(x_1)$ for the arithmetical formula defining the property of being the code number of a sentence provable from S. Then Con_S may be taken to be the sentence

$$\neg P(n_0).$$

It turns out that the proof of the implication (*) can be written down formally in the language of arithmetic. Recalling yet again that G is essentially the assertion “$S \nvdash G$ ”, this yields a proof from S of the arithmetical sentence

$$Con_S \rightarrow G,$$

Now suppose that

$$\vdash_S Con_S.$$

Then since, as we have seen,

$$\vdash_S Con_S \rightarrow G,$$

it would follow that

$$\vdash_S G.$$

But, by the First Incompleteness Theorem, if S is consistent, then $\nvdash_S G$. This is a contradiction, and we infer

Gödel's Second Incompleteness Theorem. Under the same conditions as the strong form of the First Incompleteness Theorem, the arithmetical sentence Con_S expressing the consistency of S is not provable from S.

In other words, the consistency of any arithmetically definable consistent system of axioms for arithmetic is not demonstrable in the system itself. Thus the consistency of arithmetic—assuming that it is indeed consistent—can only be demonstrated by appeal to procedures which *transcend arithmetic,* that is, in which the infinite figures in some essential way. This discovery dealt a shattering blow to Hilbert's program for establishing the consistency of mathematics by "finitistic" means.