Gödel’s Incompleteness Theorems

First published Mon Nov 11, 2013; substantive revision Thu Apr 2, 2020

Gödel's two incompleteness theorems are among the most important results in modern logic, and have deep implications for various issues. They concern the limits of provability in formal axiomatic theories. The first incompleteness theorem states that in any consistent formal system \( F \) within which a certain amount of arithmetic can be carried out, there are statements of the language of \( F \) which can neither be proved nor disproved in \( F \). According to the second incompleteness theorem, such a formal system cannot prove that the system itself is consistent (assuming it is indeed consistent). These results have had a great impact on the philosophy of mathematics and logic. There have been attempts to apply the results also in other areas of philosophy such as the philosophy of mind, but these attempted applications are more controversial. The present entry surveys the two incompleteness theorems and various issues surrounding them. (See also the entry on Kurt Gödel for a discussion of the incompleteness theorems that contextualizes them within a broader discussion of his mathematical and philosophical work.)

1. Introduction
   1.1 Outline
   1.2 Some Formalized Theories
   1.3 The Relevance of the Church-Turing Thesis

2. The First Incompleteness Theorem
   2.1 Preliminaries
   2.2 Representability
   2.3 Arithmetization of the Formal Language
   2.4 Diagonalization, or, “Self-reference”
   2.5 The First Incompleteness Theorem—Proof Completed
   2.6 Incompleteness and Non-standard Models

3. The Second Incompleteness Theorem
   3.1 Preliminaries
   3.2 Derivability Conditions
   3.3 Feferman’s Alternative Approach to the Second Theorem

4. Results Related to the Incompleteness Theorems
   4.1 Tarski’s Theorem on the Undefinability of Truth
   4.2 The Undecidability Results
   4.3 Reflection principles and Löb’s Theorem
   4.4 Hilbert’s Tenth Problem and the MRDP Theorem
   4.5 Concrete Cases of Unprovable Statements

5. The History and Early Reception of the Incompleteness Theorems

6. Philosophical Implications—Real and Alleged
   6.1 Philosophy of Mathematics
   6.2 Self-evident and Analytical Truths
   6.3 ‘Gödelian’ Arguments against Mechanism
   6.4 Gödel and Benacerraf on Mechanism and Platonism
1. Introduction

1.1 Outline

Gödel's incompleteness theorems are among the most important results in modern logic. These discoveries revolutionized the understanding of mathematics and logic, and had dramatic implications for the philosophy of mathematics. There have also been attempts to apply them in other fields of philosophy, but the legitimacy of many such applications is much more controversial.

In order to understand Gödel's theorems, one must first explain the key concepts essential to it, such as “formal system”, “consistency”, and “completeness”. Roughly, a formal system is a system of axioms equipped with rules of inference, which allow one to generate new theorems. The set of axioms is required to be finite or at least decidable, i.e., there must be an algorithm (an effective method) which enables one to mechanically decide whether a given statement is an axiom or not. If this condition is satisfied, the theory is called “recursively axiomatizable”, or, simply, “axiomatizable”. The rules of inference (of a formal system) are also effective operations, such that it can always be mechanically decided whether one has a legitimate application of a rule of inference at hand. Consequently, it is also possible to decide for any given finite sequence of formulas, whether it constitutes a genuine derivation, or a proof, in the system—given the axioms and the rules of inference of the system.

A formal system is complete if for every statement of the language of the system, either the statement or its negation can be derived (i.e., proved) in the system. A formal system is consistent if there is no statement such that the statement itself and its negation are both derivable in the system. Only consistent systems are of any interest in this context, for it is an elementary fact of logic that in an inconsistent formal system every statement is derivable, and consequently, such a system is trivially complete.

Gödel established two different though related incompleteness theorems, usually called the first incompleteness theorem and the second incompleteness theorem. “Gödel’s theorem” is sometimes used to refer to the conjunction of these two, but may refer to either—usually the first—separately. Accommodating an improvement due to J. Barkley Rosser in 1936, the first theorem can be stated, roughly, as follows:

**First incompleteness theorem**

Any consistent formal system \(\mathcal{F}\) within which a certain amount of elementary arithmetic can be carried out is incomplete; i.e., there are statements of the language of \(\mathcal{F}\) which can neither be proved nor disproved in \(\mathcal{F}\).

Gödel's theorem does not merely claim that such statements exist: the method of Gödel's proof explicitly produces a particular sentence that is neither provable nor refutable in \(\mathcal{F}\); the “undecidable” statement can be found mechanically from a specification of \(\mathcal{F}\). The sentence in question is a relatively simple statement of number theory, a purely universal arithmetical sentence.

A common misunderstanding is to interpret Gödel's first theorem as showing that there are truths that cannot be proved. This is, however, incorrect, for the incompleteness theorem does not deal with provability in any absolute sense, but only concerns derivability in some particular formal system or another. For any statement \(\neg (A)\) unprovable in a particular formal system \(\mathcal{F}\), there are, trivially, other formal systems in which \(\neg (A)\) is provable (take \(\neg (A)\) as an axiom). On the other hand, there is the extremely powerful standard axiom system of Zermelo-Fraenkel set theory (denoted as **ZF**, or, with the axiom of choice, **ZFC**; see the section on the axioms of **ZFC** in the entry on set theory), which is more than sufficient for the derivation of all ordinary mathematics. Now there are, by Gödel's first theorem, arithmetical truths that are not provable even in **ZFC**. Proving them would thus require a formal system that incorporates methods going beyond **ZFC**. There is thus a sense in which such truths are not provable using today's “ordinary” mathematical methods and axioms, nor can they be proved in a way that mathematicians would today regard as unproblematic and conclusive.

Gödel's second incompleteness theorem concerns the limits of consistency proofs. A rough statement is:

**Second incompleteness theorem**

For any consistent system \(\mathcal{F}\) within which a certain amount of elementary arithmetic can be carried out, the consistency of \(\mathcal{F}\) cannot be proved in \(\mathcal{F}\) itself.

In the case of the second theorem, \(\mathcal{F}\) must contain a little bit more arithmetic than in the case of the first theorem, which holds under very weak conditions. It is important to note that this result, like the first incompleteness theorem, is a theorem about formal provability, or derivability (which is always relative to some formal system; in this case, to \(\mathcal{F}\))
1.2 Some Formalized Theories

The existence of incomplete theories is hardly surprising. Take any theory, even a complete one (see below for examples), and drop some axiom; unless the axiom is redundant, the resulting system is incomplete. The incompleteness theorems, however, deal with a much more radical kind of incompleteness phenomenon. Unlike the above sort of trivially incomplete theories, which can be easily completed, there is no way of completing the relevant theories; all their extensions, inasmuch as they are still formal systems and hence axiomatizable, are also incomplete. They remain, so to speak, eternally incomplete and can never be completed. They are “essentially incomplete”.

In the first and loose statements of the incompleteness theorems given above, the vague requirement that “a certain amount of elementary arithmetic can be carried out” occurred. It is time to make this more precise.

1.2.1 Arithmetical Theories

The weakest standard system of arithmetic that is usually considered in connection with incompleteness and undecidability is so-called Robinson arithmetic (due to Raphael M. Robinson; see Tarski, Mostowski and Robinson 1953), standardly denoted as \( \mathsf{Q} \). As axioms, it has the following seven assumptions:

\[
\begin{align*}
& \neg(0 = x'), \\
& x' = y' \rightarrow x = y, \\
& \exists y (x = y'), \\
& x + 0 = x, \\
& x + y' = (x + y)', \\
& x \times 0 = 0, \\
& x \times y' = (x \times y) + x.
\end{align*}
\]

The intended interpretation of “\( x' \)” is the successor function, and obviously, of \(+\) and \(\times\), the addition and the multiplication functions, respectively. “\( 0 \)” is the only constant and denotes the number zero.

Adding to these elementary axioms the axiom scheme of induction:

\[
\exists y \left[ (\phi(0) \land \forall x [\phi(x) \rightarrow \phi(x')]) \rightarrow \forall x \phi(x) \right].
\]

results in (first order) Peano Arithmetic (PA). Note that unlike \( \mathsf{Q}, \) PA contains infinitely many axioms, because all (infinitely many) instances of the induction scheme, one corresponding to every formula \( \phi(x) \) (with at least one free variable) of the language, are taken as axioms. But it is a routine mechanical task to check whether a given sentence is an instance of this scheme. PA is generally taken as the standard first-order system of arithmetic.

Another natural and much-studied arithmetical system, which lies between \( \mathsf{Q} \) and PA in strength, is Primitive Recursive Arithmetic (PRA). It contains not just the above axioms of \( \mathsf{Q} \) governing successor, addition and multiplication, but also defining axioms for all primitive recursive functions (see the entry on recursive functions), and the application of the induction scheme is restricted to quantifier-free formulas (i.e., \( \forall \phi(x) \)) is not allowed to contain any (unbounded) quantifiers.

However, essentially the same system is obtained if one takes just the axioms of \( \mathsf{Q} \) and the induction scheme restricted to, roughly, purely existential formulas (in technical terms, \( \forall \Sigma^0_1 \)-formulas; see below) (this was first showed by Parsons 1970). Moreover, \( \forall \Sigma^0_1 \)-induction can be shown (Paris and Kirby 1978) to be equivalent to the induction scheme restricted to (roughly) purely universal formulas \( \forall \Pi^0_1 \)-formulas). PRA can also be formulated as a “logic-free” equational calculus. PRA, or something equivalent to it, is sufficient for developing the theory of syntax for formalized theories. It is often taken as the unproblematic background theory in which various other systems, whose legitimacy may be more controversial, are studied.

A much stronger system than \( \mathsf{PA} \), important in the foundations of mathematics, which will be mentioned now and then below, is second-order arithmetic \( \mathsf{PA}^2 \) (also often denoted by \( \mathsf{Z}_2 \)). It is more than sufficient for developing all ordinary analysis and algebra. Its language is a two-sorted first-order language (see the entry on second-order and higher-order logic), i.e., it contains two sorts of variables, number variables \( \forall x, \exists x \) (or \( \forall x, y, z \)) and property variables \( \forall X, \exists X \) (or \( \forall X, Y, Z \)) where properties are extensionally conceived. As axioms it includes, in addition to the basic axioms of \( \mathsf{PA} \), all instances of the second-order comprehension scheme:

\[
\exists \forall \phi(x) \exists X \forall x [x \subseteq X \land \forall \phi(x) \rightarrow \phi(x)]
\]

where \( \phi(x) \) can be any formula of the language of \( \mathsf{PA}^2 \) in which \( X \) doesn't occur free. (It should be mentioned that \( \mathsf{PA}^2 \) can also be formulated by adding the primitive notion of set membership \( \{x | \phi(x) \} \) to the language, regarding the variables \( \forall X, Y, Z \), as explicitly ranging over sets, and reformulating second-order comprehension as \( \forall \exists x \forall \phi(x) \exists X [x \subseteq X \land \forall \phi(x) \rightarrow \phi(x)] \).

\( \mathsf{PA}^2 \) is a very strong theory. Via the method of interpretations (see below), one can show that it is proof-theoretically as strong as the Zermelo-Fraenkel set theory \( \mathsf{ZFC} \) without the power-set axiom, call it \( \mathsf{ZFC} - \mathsf{Pow} \) (whereas the
standard, first-order PA is similarly proof-theoretically equivalent to ZFC without the axiom of infinity, ZFC–Inf. (Cf. the section on the axioms of ZFC in the entry on set theory.)

Obviously, it is assumed that our formal systems are also equipped with a system of rules of inferences (and possibly some logical axioms), usually some standard system of classical logic (though the incompleteness theorems do not essentially presuppose classical logic, but also apply to systems with, e.g., intuitionistic logic). The above standard systems all come with classical logic. The standard notation \( \vdash \) is used to express (in the meta-level) that \( A \) is derivable in \( F \), that is, that there is a proof of \( A \) in \( F \), or, in other words, that \( A \) is a theorem of \( F \). Accordingly, \( \not\vdash A \) means that \( A \) is not derivable in \( F \).

To summarize: when it is said, in the context of the incompleteness theorems, that “a certain amount of elementary arithmetic can be carried out” in a system, this usually means that it contains PRA or at least Q. For the first incompleteness theorem, Q is sufficient; for the standard proofs of the second theorem, something like PRA, at a minimum, is needed. There is a version of the second incompleteness theorem for Q (see Bezbahor & Shepherdson 1976) but there has been some debate on whether the relevant statement in Q can really be taken to express consistency, Q being so weak (see Kreisel 1958; Bezbahor & Shepherdson 1976; Pudlák 1996; Franks 2009).

1.2.2 Theories not Formulated in the Language of Arithmetic

Of course, there are many important and interesting theories in mathematics which are not even formulated in the language of arithmetic. However, the applicability of the incompleteness theorems can be dramatically extended outside the language of first-order arithmetic and its extensions, when it is noted that all that is needed is that weak theories such as Q, or PRA, can be interpreted in the system in question. Most importantly, this involves various systems of set theory. For example, the incompleteness theorems hold for ZFC–Inf (i.e., ZFC without the axiom of infinity) and all its extensions, however strong (as long as they are axiomatizable.)

Roughly, a theory \( \vdash (T_1) \) is interpretable in another theory \( \vdash (T_2) \) if the primitive concepts and the range of the variables of \( \vdash (T_1) \) are definable in \( \vdash (T_2) \) so that it is possible to translate every theorem of \( \vdash (T_1) \) into a theorem of \( \vdash (T_2) \). One should not misunderstand such interpretations as providing anything like intuitive synonymy. Two theories may have radically different intended subject matter and yet, as formal systems, one may be interpretable in another. (As an illustration: a simple theory of ancestors may be, taken as a formal system, interpreted in arithmetic; obviously this does not mean that grandmothers and such are really numbers.) What is significant is that interpretability preserves certain elementary formal properties of theories, most importantly, consistency: if \( \vdash (T_1) \) is interpretable in \( \vdash (T_2) \) and \( \vdash (T_2) \) is consistent, \( \vdash (T_1) \) is also consistent. And any system in which Q can be interpreted is guaranteed to be essentially incomplete. For any such theory in which Q is interpretable, the incompleteness could be proved also directly; for example, in various theories of set theory, one can code formulas and derivations (instead of numbers) by sets, “Gödel sets”, and proceed then as usual (see, e.g., Fitting 2007). However, for most purposes, it is just much simpler to establish the interpretability of Q in the theory at issue.

In sum, it is said that “a certain amount of elementary arithmetic can be carried out within a system”, what is meant is either that the system is an axiomatizable extension of Q, or that Q can be interpreted in it. (In the case of the standard proofs of the second incompleteness theorem, substitute PRA for Q.)

1.2.3 Some Exceptions: Complete Theories

On the other hand, not all theories of arithmetic are incomplete. The theory of only addition of natural numbers but without multiplication (often called “Presburger arithmetic”), for example, is complete (and decidable) (Presburger 1929), as is the theory of multiplication of the positive integers (Skolem 1930). These theories are, though, very weak. But in any case, at least a theory which deals with both addition and multiplication is needed. More interestingly, the natural first-order theory of arithmetic of real numbers (with both addition and multiplication), the so-called theory of real closed fields \( \mathbf{RCF} \), is both complete and decidable, as was shown by Tarski (1948); he also demonstrated that the first-order theory of Euclidean geometry is complete and decidable. Thus, one should keep in mind that there are some non-trivial and interesting theories to which Gödel’s theorems do not apply.

1.3 The Relevance of the Church-Turing Thesis

Gödel originally only established the incompleteness of a particular though very comprehensive formalized theory P, a variant of Russell’s type-theoretical system PM (for Principia Mathematica, see the sections on Paradoxes and Russell’s Type Theories in the entries on type theory and Principia Mathematica), and all extension of P with the same language, whose set of axioms is primitive recursive. He also suggested, though did not demonstrate, that the proof could be adapted to apply also to the standard axiom systems of set theory such as ZFC. Though it turns out that Gödel in fact already had a very general result, it was, at the time, unclear just how general this really was (see also Section 5).

What was still missing was an analysis of the intuitive notion of decidability, needed in the characterization of the notion of an arbitrary formal system. Recall that the set of axioms and the proof relation of a formalized system are required to be decidable. Mathematicians and logicians have implicitly used the intuitive notion of a decision method since antiquity, and as long as one asked for a positive solution, it was sufficient that one presented a concrete method that intuitively struck everyone as a mechanical method. For the general limitative results, such as the general incompleteness theorems, or the undecidability results (see 4.2), however, a precise mathematical explication of the
but not strongly representable (the key example being the set of statements provable in the system).

representability for functions. As the incompleteness results in particular teach us, there are sets which are only weakly representable.

It is obvious how all these notions are generalized to many-place relations. There are also related notions of set membership.

\[ \mathbf{n} \in S \Leftrightarrow F \vdash A(\underline{n}). \]

A set \( S \) of natural numbers is strongly representable in \( \mathcal{L}(F) \) if there is a formula \( \forall x A(x) \) such that for every natural number \( n \):

\[ F \vdash \neg A(n). \]

\[ \mathbf{n} \in S \Rightarrow F \vdash A(\underline{n}); \quad \mathbf{n} \not\in S \Rightarrow F \vdash \neg A(\underline{n}), \]

\[ \mathbf{n} \in S \Leftrightarrow F \vdash \exists x A(x). \]

2.2 Representability

Gödel's proof also requires the notion of representability of sets and relations in a formal system \( \mathcal{L}(F) \). More precisely, two related notions are needed.

A set \( \{S\} \) of natural numbers is strongly representable in \( \mathcal{L}(F) \) if there is a formula \( \forall x A(x) \) of the language of \( \mathcal{L}(F) \) with one free variable \( x \) such that for every natural number \( \mathbf{n} \):

\[ \begin{align} \mathbf{n} & \in S \Rightarrow \exists x A(x) \wedge \exists x \neg A(x); \\
\neg \mathbf{n} & \in S \Rightarrow \exists x A(x). \end{align} \]

A set \( \{S\} \) of natural numbers is weakly representable in \( \mathcal{L}(F) \) if there is a formula \( \exists x A(x) \) such that for every natural number \( \mathbf{n} \):

\[ \begin{align} \mathbf{n} & \in S \Rightarrow \exists x A(x); \\
\neg \mathbf{n} & \in S \Rightarrow \exists x \neg A(x). \end{align} \]

It is obvious how all these notions are generalized to many-place relations. There are also related notions of representability for functions. As the incompleteness results in particular teach us, there are sets which are only weakly representable (the key example being the set of statements provable in the system).
The property of being provable in $F$ can then be defined as $\exists x \Prf_F (x, y)$. Let us abbreviate this formula, which (weakly or strongly) represents the set in question, and usually such a formula is used to represent $\Prov_F (S)$.

Both notions of representability—strong and weak—must be clearly distinguished from mere definability (in the standard sense of the word). A set $\Prov_F (S)$ is definable in the language of arithmetic if there is a formula $\Prov_F (S(x))$ in the language such that $\Prov_F (S(\text{underline}(n)))$ is true in the standard structure of natural numbers (the intended interpretation) if and only if $\Prov_F (S(\text{boldsymbol}(n) \in S))$. There are many sets which can be defined in the language of arithmetic but not (even weakly) represented in any $\Prov_F (S)$, such as the set of consistent formulas, the set of sentences unprovable in the system $\Prov_F (S)$, or the set of Diophantine equations with no solutions (see below).

### 2.3 Arithmetization of the Formal Language

The next essential step of Gödel's proof is to take the language of a formal system, which is always precisely defined (this is part of being a formal system), and fix a correspondence of a certain kind between the expressions of that language and the system of natural numbers—a coding, "arithmetization", or "Gödel numbering", of the language. There are many possible ways of accomplishing this, and the details do not really matter (for some more details of one quite standard approach, see the supplementary document Gödel Numbering). The essential point is that the chosen mapping is effective: it is always possible to pass, purely mechanically, from an expression to its code number, and from a number to the corresponding expression. Today, when most of us are familiar with computers and the fact that so many things can be coded by zeros and ones, the possibility of such an arithmetization is hardly surprising.

Roughly, one proceeds as follows: First, the primitive symbols of the language are paired with distinct natural numbers, "symbol numbers". A little number theory then suffices to code sequences of numbers by single numbers. Consequently, well-formed formulas, as sequences of primitive symbols, are each assigned a unique number. Finally, derivations, or proofs, of the system, being sequences of formulas, are arithmetized, and are also assigned specific numbers. Such a code, the "Gödel number" of a formula $A$, is denoted as $\ulcorner A\urcorner$, and similarly for derivations.

In this way, syntactical properties, relations and operations are reflected in arithmetic: for example, $\neg(x)$ is the arithmetical function that sends the Gödel number of a formula to the Gödel number of its negation; in other words, $\Prov_F (\ulcorner \neg A \urcorner) = \Prov_F (\ulcorner A \urcorner)$; similarly, $(x, y)$ is the function which maps the Gödel numbers of a pair of formulas to the Gödel number of the implication of the formulas: $\Prov_F (\ulcorner A \urcorner, \ulcorner B \urcorner) = \Prov_F (\ulcorner A \rightarrow B \urcorner)$; and so on. There is an arithmetical formula, call it $\Prov_F (x)$, which is true of $\Prov_F (\boldsymbol{n})$ if $\Prov_F (\boldsymbol{n})$ is a Gödel number of a well-formed formula of the system. There is also an arithmetical formula $M(x, y, z)$ which is true exactly if one has a valid application of the rule of inference modus ponens for some formulas $\Prov_F (\ulcorner A \urcorner)$ and $\Prov_F (\ulcorner B \urcorner)$ with $\Prov_F (\ulcorner x = \ulcorner A \urcorner)$ and $\Prov_F (\ulcorner y = \ulcorner B \urcorner)$ and $\Prov_F (\ulcorner z = \ulcorner A \urcorner)$; etc. In this way, all the syntactic properties and operations can be simulated at the level of numbers, and moreover they are strongly representable in all theories which contain $\Prov_F (S)$.

As it is decidable (by the definition of formal systems) whether a given sequence of formulas constitutes a proof of a given sentence, according to the rules of the chosen formal system $\Prov_F (S)$, the binary relation "$x(y)" is (the Gödel number of) a proof of a formula (with the Gödel number) $\Prov_F (S(y))$ can be strongly represented in all systems containing $\Prov_F (S)$, and thus in $\Prov_F (S)$ in particular. Let us denote the formula which strongly represents this relation in $\Prov_F (S)$ itself as $\Prov_F (\ulcorner A \urcorner, \ulcorner B \urcorner)$. In this way, we have:

The Representability Theorem

In any consistent formal system which contains $\Prov_F (S)$:

1. A set (or relation) is strongly representable if and only if it is recursive;
2. A set (or relation) is weakly representable if and only if it is recursively enumerable.

In sum, we have:

Both notions of representability—strong and weak—must be clearly distinguished from mere definability (in the standard sense of the word). A set $\Prov_F (S)$ is definable in the language of arithmetic if there is a formula $\Prov_F (S(x))$ in the language such that $\Prov_F (S(\text{underline}(n)))$ is true in the standard structure of natural numbers (the intended interpretation) if and only if $\Prov_F (S(\text{boldsymbol}(n) \in S)).$ There are many sets which can be defined in the language of arithmetic but not (even weakly) represented in any $\Prov_F (S)$, such as the set of consistent formulas, the set of sentences unprovable in the system $\Prov_F (S)$, or the set of Diophantine equations with no solutions (see below).
It is always possible to choose the provability predicate \( \Prov_F (x) \) to be a \( \Sigma^{0}_1 \)-formula.

### 2.4 Diagonalization, or, “Self-reference”

The next and perhaps somewhat surprising ingredient of Gödel’s proof is the following important lemma (we still assume that \( \forall (F) \) is a formal system which contains \( Q \)):

**The Diagonalization Lemma**

Let \( \forall (A(x)) \) be an arbitrary formula of the language of \( \forall (F) \) with only one free variable. Then a sentence \( \forall (D) \) can be mechanically constructed such that

\[
[ F \vdash D \iff A(\ulcorner D\urcorner). \]

(For a sketch of the proof; see Supplement: The Diagonalization Lemma)

In the literature, this lemma is sometimes also called “the self-referential lemma” or “the fixed point lemma”. It has many important applications beyond the incompleteness theorems.

It is often said that given a property denoted by \( \forall (A(x)) \), the sentence \( \forall (D) \) is a self-referential sentence which “says of itself” that it has the property \( \forall (A) \). Such figures of speech may be heuristically useful, but they are also easily misleading and suggest too much. For example, note that the lemma only provides a (provable) material equivalence between \( \forall (D) \) and \( \forall (A(\ulcorner D\urcorner)) \) (which states that both sides must have the same truth-value) and does not claim any sort of sameness of meaning. In particular, \( \forall (D) \) and \( \forall (A(\ulcorner D\urcorner)) \) are by no means identical—and neither are \( \ulcorner D\urcorner \) and \( \ulcorner A(\ulcorner D\urcorner)\urcorner \).

### 2.5 The First Incompleteness Theorem—Proof Completed

To complete the proof, the Diagonalization Lemma is applied to the negated provability predicate \( \neg \Prov_F (x) \):

This gives a sentence \( \forall (G_F) \) such that

\[
[ \neg \Prov_F (\ulcorner G_F\urcorner). \]

Thus, it can be shown, even inside \( \forall (F) \), that \( \forall (G_F) \) is true if and only if it is not provable in \( \forall (F) \).

It is not difficult to show that \( \forall (G_F) \) is neither provable nor disprovable in \( \forall (F) \), if \( \forall (F) \) only is 1-consistent.

For the first half, assume that \( \forall (G_F) \) were provable. Then, by the weak representability of provability-in-\( \forall (F) \) by \( \forall (\Prov_F(x)) \), \( \forall (F) \) would also prove \( \forall (\Prov_F (\ulcorner G_F\urcorner)) \). However, because \( \forall (F) \) in fact also proves the equivalence (G), i.e., \( [ F \vdash \forall (G_F) \iff \neg \Prov_F (\ulcorner G_F\urcorner) ] \), \( \forall (F) \) would then prove \( \forall (\neg G_F) \) too. But this would mean that \( \forall (F) \) is inconsistent. In sum, if \( \forall (F) \) is consistent, then \( \forall (G_F) \) is not provable in \( \forall (F) \).

For this first half, the assumption of the simple consistency of \( \forall (F) \) suffices.

For the second half, it has to be assumed that \( \forall (F) \) is 1-consistent (if \( \forall (\Prov_F (\ulcorner G_F\urcorner)) \) has been chosen such that it is a \( \forall (\Sigma^{0}_1) \)-sentence; otherwise, the more general assumption of \( \forall (\omega) \)-consistency is needed).

Assume that \( [ \forall (F) \vdash \neg \Prov_F (\ulcorner G_F\urcorner) ] \). Then, \( \forall (F) \) cannot prove \( \forall (G_F) \), for otherwise \( \forall (F) \) would be simply inconsistent. Hence no natural number \( \forall (\text{boldsymbol}[n]) \) is the Gödel number of a proof of \( \forall (G_F) \), and because the proof relation is strongly representable, for all \( \forall (\text{boldsymbol}[n]) \), \( [ \forall (F) \vdash \neg \Prov_F (\ulcorner G_F\urcorner) ] \).

If also \( [ \forall (F) \exists x \Prov_F (x, \ulcorner G_F\urcorner) ] \), \( \forall (F) \) does not prove \( \forall (\exists x \Prov_F (x, \ulcorner G_F\urcorner)) \), in other words, by the definition of \( \forall (\Prov_F (x)) \), \( \forall (F) \) does not prove \( \forall (\Prov_F (\ulcorner G_F\urcorner)) \).

By the key equivalence (G), \( \forall (F) \) also does not prove \( \forall (\neg G_F) \).

**Gödel’s First Incompleteness Theorem**

Assume \( \forall (F) \) is a formalized system which contains Robinson arithmetic \( Q \). Then a sentence \( \forall (G_F) \) of the language of \( \forall (F) \) can be mechanically constructed from \( \forall (F) \) such that:

i. If \( \forall (F) \) is consistent, then \( \forall (F) \not\vdash G_F \).

ii. If \( \forall (F) \) is 1-consistent, then \( \forall (F) \not\vdash \neg G_F \).

Such an independent, or “undecidable” (that is, neither provable nor refutable in \( \forall (F) \)) statement \( \forall (G_F) \) in \( \forall (F) \) is often called “the Gödel sentence” of \( \forall (F) \).

In fact, in favourable circumstances, it can be shown that \( \forall (G_F) \) is true, provided that \( \forall (F) \) is indeed consistent. This is the case if, for example, the provability predicate \( \forall (\Prov_F (x)) \) has been chosen as a \( \forall (\Sigma^{0}_1) \)-formula: The Gödel sentence is then provably equivalent to the universal formula \( \forall (\forall x \Prov_F (x, \ulcorner G_F\urcorner)) \).

Such formulas can be proved false whenever they in fact are false: if false, there would be a number \( \forall (\text{boldsymbol}[n]) \) such that \( [ \forall (F) \vdash \Prov_F (\ulcorner n\urcorner, \ulcorner G_F\urcorner) ] \) (this holds already in \( Q \)). This, however, would contradict the incompleteness theorem. Therefore, \( \forall (G_F) \) cannot be false, and must be true. For this reason, the Gödel sentence is often called “true but unprovable”.

One should not get confused here: “Gödel’s theorem” is the general incompleteness result of Gödel which concerns a large class of formal systems, while the “Gödel sentence” is the constructed, formally undecidable sentence which varies from one formal system to another. This is why it is important to include the subscript \( F \) in \( \forall (G_F) \).

Furthermore, one should not confuse the two different senses of “undecidable” in this context. On the one hand, a
where $G_F$ is the Gödel sentence for $F$ provided by the first theorem. If $\Cons(F)$ were provable in $F$, so
\[ F \vdash \Cons(F) \rightarrow G_F, \]
This gives:
\[ (\text{case (i) above}) \text{ can then presumably be formalized} \]
\[ \neg \Prov_F (\ulcorner \bot \urcorner) \). Let us abbreviate this formula by $\Cons(F)$. The proof of the first part of the first incompleteness theorem

3. The Second Incompleteness Theorem

2.6 Incompleteness and Non-standard Models

It is illuminating to reflect on the first incompleteness theorem also from the model theoretic perspective—though the theorem itself does not in any way require this. Namely, it is possible to conclude that any theory $\forall(F)$ satisfying the conditions of the theorem must possess, in addition to the intended interpretation or “standard model” (in the case of arithmetical theories, the structure of natural numbers), non-intended interpretations or “non-standard models”—that no such theory can rule out the latter and fix uniquely the intended interpretation. Namely, if there are independent statements such as $\forall(G_F, F)$ must have both models which satisfy $\forall(G_F)$ and models which rather satisfy $\forall(\neg G_F)$. As $\forall(\neg G_F)$ is equivalent to $\forall(\exists x \Prf_F (x, \ulcorner \text{G_F} \urcorner))$, the latter models must possess entities which satisfy the formula $\forall(\Prf_F (x, \ulcorner \text{G_F} \urcorner))$. And yet we know (because $\forall(\Prf_F (x, y))$ strongly represents the proof relation) that for any numeral $\forall(\underline{n}, F)$ can prove $\forall(\neg \Prf_F (\underline{n}, F))$. Therefore, no natural number $\forall(\text{boldsymbol{n}})$ can witness the formula. It follows that any such non-standard model must contain, in addition to natural numbers (denotations of the numerals $\forall(\underline{n})$), “infinite” non-natural numbers after the natural numbers.

The study of non-standard models did not start with Gödel’s results—Skolem, in particular, was already aware of them earlier in a different context (he had discovered that first-order theories of set theory have unnaturally small, namely, countable models, in Skolem 1922; cf. the entry on Skolem’s paradox)—but the first incompleteness theorem elucidates the existence of non-standard models in the context of arithmetic, while the nonstandard models elucidate the first incompleteness theorem. Non-standard models have since then become a rich research area in mathematical logic (see, e.g., Boolos & Jeffrey 1989: Ch. 17; Kaye 1991).

3. The Second Incompleteness Theorem

3.1 Preliminaries

Informally, the reasoning leading to the second incompleteness theorem is relatively simple. Given the arithmetized provability predicate, it is also easy to present an arithmetized consistency statement: pick some manifestly inconsistent formula (in arithmetical theories, a standard choice is $\forall(\underline{0} = \underline{1}))$; let us denote it by $\forall(\bot)$; (the arithmetized counterpart of) the consistency of the system can then be defined as $\forall(\neg \Prov_F (\ulcorner \bot \urcorner))$. Let us abbreviate this formula by $\forall(\Cons(F))$. The proof of the first part of the first incompleteness theorem (i.e., the case (i) above) can then presumably be formalized inside $\forall(F)$ (in practice this would certainly be intricate). This gives:
\[ \forall(F) \vdash \forall(\Cons(F)) \rightarrow \forall(G_F, F), \]
where $\forall(G_F)$ is the Gödel sentence for $\forall(F)$ provided by the first theorem. If $\forall(\Cons(F))$ were provable in $\forall(F)$, so
Gödel's second incompleteness theorem

Assume \( \langle F \rangle \) is a consistent formalized system which contains elementary arithmetic. Then \( \langle F \not\vdash \Cons(F) \rangle \) cannot be provable in \( \langle F \rangle \) either.

There is a question of philosophical importance that should be mentioned here: As it stands, Gödel's second incompleteness theorem only establishes the unprovability of one sentence, \( \langle \Cons(F) \rangle \). But does this sentence really express that \( \langle F \rangle \) is consistent? (Compare this with the remark above that \( \langle G_F \rangle \) does not, strictly speaking, express its own unprovability.) Furthermore, might there not be other sentences which are provable and also express the consistency of \( \langle F \rangle \)?

Giving a rigorous proof of the second theorem in a more general form that covers all such sentences, however, has turned out to be very complicated. The basic reason for this is that, unlike in the first theorem, not just any, merely extensionally adequate provability predicate works for the formalization of the consistency claim. The manner of presentation makes all the difference. For example, Rosser's provability predicate mentioned above would not do; one can prove the "consistency" of \( \langle F \rangle \) in \( \langle F \rangle \), if consistency is expressed in terms of Rosser's provability predicate. One must thus add some further conditions for the provability predicate in order for the proof of the second incompleteness theorem to go through. Following Feferman (1960), it is customary to say that whereas the first theorem and its relatives are extensional results, the second theorem is intensional: it must be possible to think that \( \langle \Cons(F) \rangle \) in some sense expresses the consistency of \( \langle F \rangle \)—that it really means that \( \langle F \rangle \) is consistent.

3.2 Derivability Conditions

The proof of the second incompleteness theorem requires that the provability predicate in \( \langle F \rangle \) satisfies a number of conditions which are used in the details of the proof. There are several different sets of conditions that will do.

The first detailed proof of the second incompleteness theorem appeared in (Hilbert & Bernays 1939) (mainly written by Bernays), though only for one specific theory, PA. It uses a rather awkward set of conditions for the provability predicate. These were more technical lemmas for the needs of a particular proof and not any sort of analysis of "natural" provability predicates. A much more elegant, now standard list of "derivability conditions" was presented by Löb (1955)—though their intended use was somewhat different (see below).

Löb's Derivability Conditions

\[
\begin{align}
(D1) \ &\vdash F \vdash A \lor \vdash \Prov_F (\ulcorner A \urcorner). \\
(D2) \ &\vdash F \vdash \Prov_F (\ulcorner A \urcorner) \land \Prov_F (\ulcorner A \rightarrow \Prov_F (\ulcorner A \urcorner)) \rightarrow \Prov_F (\ulcorner A \urcorner). \\
(D3) \ &\vdash F \vdash \Prov_F (\ulcorner A \land B \urcorner) \leftrightarrow (\Prov_F (\ulcorner A \urcorner) \land \Prov_F (\ulcorner B \urcorner)).
\end{align}
\]

(D1) is simply a restatement of the requirement from the proof of the first theorem that provability is weakly representable. Roughly put, (D2) requires that the whole demonstration of (D1), for the candidate provability predicate \( \langle \Prov_F \rangle \), can itself be formalized inside \( \langle F \rangle \). Finally, (D3) requires that the provability predicate is closed under Modus Ponens.

If the arithmetized provability predicate indeed satisfies these conditions, the second theorem can be proved. Let \( G_F \) once again be the Gödel sentence for \( \langle F \rangle \) given by the first theorem. It is not too difficult to show, using the derivability conditions, that:

\[
\langle F \not\vdash G_F \rightarrow \vdash \Cons(F) \rangle
\]

This immediately yields the unprovability of \( \langle \Cons(F) \rangle \), given the first incompleteness theorem.

Furthermore, Jeroslow (1973) demonstrated, with an ingenious trick, that it is in fact possible to establish the second theorem without (D3). However, in some other cases (e.g., when proving Löb's theorem; see below), and in Provability Logic, all three conditions are still needed.

3.3 Feferman's Alternative Approach to the Second Theorem

Under the assumption that a provability predicate for a theory satisfies the derivability conditions (or, by Jeroslow's trick, at least D1 and D2) it is relatively easy to prove the relevant case of the second incompleteness theorem. However, in practice one has to establish whether a proposed arithmetized provability predicate really satisfies the conditions case by case, and typically this is long and tedious.

This drawback, among other things (see Feferman 1997), led Solomon Feferman in the late 1950s to look for an alternative line of attack to the second theorem (see Feferman 1960). Feferman approaches the issue in two steps:

First, he isolates the formulas \( \langle \Prov_{FOL}(\langle x \rangle) \rangle \) which arithmetize some standard notion of derivability in first-order logic in order to allow us to fix one chosen formula for provability in logic. How the set of non-logical axioms of the system at issue are presented is left open at this stage. Secondly, Feferman looks for a suitable constraint for presenting the axioms. Among the formulas of the language of arithmetic, he isolates what he calls PR- and RE-formulas; the former correspond to the canonical primitive recursive (PR) definitions in arithmetic, and the latter to existential generalizations of the former. Every recursively enumerable (RE) set can be defined by a formula of the latter sort; these are just the \( \langle \Sigma^0_1 \rangle \)-formulas. These two classes are easy to discriminate purely by their syntactical form. (In fact, by the MRDP Theorem (see below), one could—which instead of RE-formulas—focus on even simpler class of existentially quantified Diophantine equations.)
We have above noted the important fact that in all arithmetical theories \( F \) containing \( \mathcal{Q} \), a set is strongly representable in \( F \) if and only if it is recursive, and a set is recursively enumerable if and only if it is weakly representable. Furthermore, one can always take the formula weakly or strongly representing the set to be a \( \Sigma^{0}_1 \)-formula (i.e., \( \langle \Sigma^{0}_1 \rangle \)-formula; and, by MRDP Theorem, even an existentially quantified Diophantine equation).

It is then natural to require that the set of non-logical axioms of the system at issue is represented by such a formula. If the arithmetized definition of the set of Gödel numbers of axioms reflects how the axioms, if infinite, are inductively defined, the resulting formula will be \( \langle \Sigma^{0}_1 \rangle \). (For theories which are axiomatizable with finitely many axioms, there is a unique representation of the axioms in the form of a list, and consequently, a unique consistency statement relative to \( \langle \Prov_{FOL}(x) \rangle \).) In contrast to determining whether the derivability conditions are satisfied, it is a relatively routine task to determine that a given formula which formalizes the axioms is indeed of the required form \( \langle \Sigma^{0}_1 \rangle \).

Now the version of the second incompleteness theorem presented in Feferman 1960 is:

A variant of second incompleteness theorem (Feferman 1960)

Let \( \langle \mathcal{P} \rangle \) be a consistent extension of \( \mathcal{PA} \), and let \( \langle \Ax_F(x) \rangle \) be a \( \langle \Sigma^{0}_1 \rangle \)-formula which weakly represents the axioms of \( \langle \mathcal{F} \rangle \), and \( \langle \Cons(F) \rangle \) be a consistency statement constructed from \( \langle \Ax_F(x) \rangle \) and \( \langle \Prov_{FOL}(x) \rangle \). Then \( \langle \Cons(F) \rangle \) is not provable in \( \langle \mathcal{F} \rangle \).

For still different approaches to the second incompleteness theorem, see Feferman 1982, 1989a; Visser 2011. For some philosophical complications concerning the second theorem, see Detlefsen 1979, 1986, 1990, 2001; Auerbach 1985, 1992; Roeper 2003; Franks 2009 (see also section on incompleteness in the entry on Hilbert's program).

4. Results Related to the Incompleteness Theorems

4.1 Tarski’s Theorem on the Undefinability of Truth

Gödel first arrived at the incompleteness results (see Section 5 below) by noting that truth (of the language of a system) must be undefinable in the system, a result conventionally credited to Tarski (there are certain real virtues in Tarski’s way of presenting the issue; see Gómez Torrente 2004). Let us now view the result in the context of Tarski’s approach to truth.

Tarski clearly distinguished the object language, i.e., the language the truth of whose sentences is at stake, and the metalanguage in which the former is discussed. He also required (see the entry on Tarski’s truth definitions) that any satisfactory definition of truth \( \langle \True(x) \rangle \) for the object language should satisfy his “Convention T”, that is, it should have as its consequence all equivalences (“T-equivalences”) of the form

\[
\tag{T} \True(\ulcorner A \urcorner) \leftrightarrow B,
\]

where \( \ulcorner A \urcorner \) is a name of a sentence of the object language, and \( B \) its translation in the metalanguage. If the metalanguage is identical with the object language, or is an extension of the object language, \( \langle B \rangle \) is simply \( A \) itself, and the T-equivalences are of the form:

\[
\langle \True(\ulcorner A \urcorner) \rangle \leftrightarrow \langle A \rangle.
\]

What the undefinability theorem shows is that the object language and the metalanguage cannot coincide, but must be distinct.

Tarski’s Undefinability Theorem

Let \( \langle \mathcal{F} \rangle \) be a consistent formalized system that contains a sufficient amount of arithmetic. Then there is no formula \( \langle \Tr(x) \rangle \) in the language of \( \langle \mathcal{F} \rangle \) such that for every sentence \( \langle A \rangle \) of the language of \( \langle \mathcal{F} \rangle \):

\[
\lnot F \vdash \Tr(\ulcorner A \urcorner) \leftrightarrow \langle A \rangle.
\]

The idea of the proof: If there were such a formula of the language of \( \langle \mathcal{F} \rangle \), an easy application of the Diagonalization Lemma to its negation would result in the paradoxical sentence \( \langle \neg L \rangle \) (for “Liar”; see the Liar paradox), such that:

\[
\lnot F \vdash \neg \Tr(\ulcorner \neg L \urcorner) \leftrightarrow \langle \neg L \rangle.
\]

which, together with the T-equivalences, which were assumed to be derivable, would quickly give an explicit contradiction, thus contradicting the assumption that \( \langle \mathcal{F} \rangle \) is consistent.

Similarly, it can be proved that the set of true sentences of \( \langle \mathcal{F} \rangle \) is not definable in the intended interpretation of \( \langle \mathcal{F} \rangle \)—in the now standard sense of “definability” (see above).

4.2 The Undecidability Results

The tools used in proving Gödel’s theorems also provide various important undecidability results. A theory is called decidable if the set of its theorems (sentences derivable in it) is decidable, that is (by the Church-Turing thesis) recursive. Otherwise, the theory is undecidable. Informally, being decidable means that there is a mechanical procedure which enables one to decide whether an arbitrary given sentence (of the language of the theory) is a theorem or not.

If a theory is complete, it is decidable (proof sketch: given a sentence \( \langle A \rangle \), systematically generate the theorems of the theory; by completeness, eventually either \( \langle A \rangle \) or \( \langle \neg A \rangle \) will be produced in a finite time). The converse, though,
Exactly which instances of the reflection scheme are actually provable in the system? Löb's Theorem gives a precise

required to be a $\Sigma^{0}_1$-sentence.)

example, is the Reflection Principle restricted to $\Sigma^{0}_1$-sentences (i.e., the sentence $A$ in the scheme is

The scheme can also be restricted. Equivalent to the assumption of 1-consistency, or $\Sigma^{0}_1$-soundness, for

The paper of Martin Hugo Löb (1955), augmented by comments of a referee, brought substantial advances on various

First, it introduces the now standard Löb derivability conditions discussed previously in the context of the second

The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of

\[ \tag{Ref} \Prov_F(\ulcorner A\urcorner) \rightarrow A. \]

4.3 Reflection Principles and Löb’s Theorem

Heuristically, one may view the Gödel sentence $\langle G_F \rangle$ as expressing its own unprovability—saying “I am not

Church’s Theorem

First-order predicate logic is undecidable.

(This undecidability result was first established by Church 1936a, b; the method of deriving it via the undecidability of $Q$

Subsequently, a number of theories and problems from different areas of mathematics have been shown to be

undecidable (see, e.g., Davis 1977; Murawski 1999: Ch 3).

A theory $\langle F \rangle$ is called essentially undecidable if every consistent extension of it in the language of $\langle F \rangle$ is undecidable. The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of the axioms of $Q$. Then for any sentence $\langle B \rangle$ of the language of arithmetic,

\[ \tag{A} \Prov_F(\ulcorner A\urcorner) \rightarrow A. \]

\[ \tag{D} F \vdash G^B \leftrightarrow \neg B(\ulcorner G^B\urcorner). \]

As before, it follows that $\langle F \not\vdash G^B \rangle$. It has been assumed that $\langle B(x) \rangle$ strongly represents the set of theorems, so this entails $\langle F \vdash \neg B(\ulcorner G^B\urcorner) \rangle$, and therefore, by (D), $\langle F \not\vdash G^B \rangle$, a contradiction. Therefore, $\langle F \rangle$ must be undecidable.

A theory $\langle F \rangle$ is called essentially undecidable if every consistent extension of it in the language of $\langle F \rangle$ is undecidable. The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of

\[ \tag{A} \Prov_F(\ulcorner A\urcorner) \rightarrow A. \]

\[ \tag{D} F \vdash G^B \leftrightarrow \neg B(\ulcorner G^B\urcorner). \]

As before, it follows that $\langle F \not\vdash G^B \rangle$. It has been assumed that $\langle B(x) \rangle$ strongly represents the set of theorems, so this entails $\langle F \vdash \neg B(\ulcorner G^B\urcorner) \rangle$, and therefore, by (D), $\langle F \not\vdash G^B \rangle$, a contradiction. Therefore, $\langle F \rangle$ must be undecidable.

A theory $\langle F \rangle$ is called essentially undecidable if every consistent extension of it in the language of $\langle F \rangle$ is undecidable. The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of

\[ \tag{A} \Prov_F(\ulcorner A\urcorner) \rightarrow A. \]

\[ \tag{D} F \vdash G^B \leftrightarrow \neg B(\ulcorner G^B\urcorner). \]

As before, it follows that $\langle F \not\vdash G^B \rangle$. It has been assumed that $\langle B(x) \rangle$ strongly represents the set of theorems, so this entails $\langle F \vdash \neg B(\ulcorner G^B\urcorner) \rangle$, and therefore, by (D), $\langle F \not\vdash G^B \rangle$, a contradiction. Therefore, $\langle F \rangle$ must be undecidable.

A theory $\langle F \rangle$ is called essentially undecidable if every consistent extension of it in the language of $\langle F \rangle$ is undecidable. The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of

\[ \tag{A} \Prov_F(\ulcorner A\urcorner) \rightarrow A. \]

\[ \tag{D} F \vdash G^B \leftrightarrow \neg B(\ulcorner G^B\urcorner). \]

As before, it follows that $\langle F \not\vdash G^B \rangle$. It has been assumed that $\langle B(x) \rangle$ strongly represents the set of theorems, so this entails $\langle F \vdash \neg B(\ulcorner G^B\urcorner) \rangle$, and therefore, by (D), $\langle F \not\vdash G^B \rangle$, a contradiction. Therefore, $\langle F \rangle$ must be undecidable.

A theory $\langle F \rangle$ is called essentially undecidable if every consistent extension of it in the language of $\langle F \rangle$ is undecidable. The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of

\[ \tag{A} \Prov_F(\ulcorner A\urcorner) \rightarrow A. \]

\[ \tag{D} F \vdash G^B \leftrightarrow \neg B(\ulcorner G^B\urcorner). \]

As before, it follows that $\langle F \not\vdash G^B \rangle$. It has been assumed that $\langle B(x) \rangle$ strongly represents the set of theorems, so this entails $\langle F \vdash \neg B(\ulcorner G^B\urcorner) \rangle$, and therefore, by (D), $\langle F \not\vdash G^B \rangle$, a contradiction. Therefore, $\langle F \rangle$ must be undecidable.

A theory $\langle F \rangle$ is called essentially undecidable if every consistent extension of it in the language of $\langle F \rangle$ is undecidable. The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of

\[ \tag{A} \Prov_F(\ulcorner A\urcorner) \rightarrow A. \]

\[ \tag{D} F \vdash G^B \leftrightarrow \neg B(\ulcorner G^B\urcorner). \]

As before, it follows that $\langle F \not\vdash G^B \rangle$. It has been assumed that $\langle B(x) \rangle$ strongly represents the set of theorems, so this entails $\langle F \vdash \neg B(\ulcorner G^B\urcorner) \rangle$, and therefore, by (D), $\langle F \not\vdash G^B \rangle$, a contradiction. Therefore, $\langle F \rangle$ must be undecidable.

A theory $\langle F \rangle$ is called essentially undecidable if every consistent extension of it in the language of $\langle F \rangle$ is undecidable. The above proof sketch in fact establishes that $Q$ is essentially undecidable. (There are some very weak theories that

Recall that $Q$ has only finitely many axioms and let $\langle A, Q \rangle$ stand for the single sentence consisting of the conjunction of
It is often stated that before the celebrated Paris-Harrington theorem (see below), no such natural independent
misleading: interesting as they may be, they don’t have the generality of Gödel’s theorems proper, but only provide
are called variants of Gödel’s theorem, or their proofs of independence alternative proofs of Gödel’s theorem, but this is
statements which are independent of
1989b). Some well known, natural examples are listed below, beginning with some quite natural mathematical
undecidable in some standard theories (though, just how natural even these are has been disputed; see Feferman
e.g., in
there are any simple and natural mathematical statements which are likewise undecidable in chosen basic theories,
important result in this direction was achieved. Call an equation “an exponential Diophantine equation” if it involves also
exponentiation, as well as addition and multiplication (that is, one can have both constants and variables as
exponents); naturally, the focus is still in the integer solutions. Davis, Putnam, and Robinson (1961), showed that the
problem of solvability of exponential Diophantine equations is undecidable. In 1970, Yuri Matiyasevich added the final
missing piece, and demonstrated that the problem of the solvability of Diophantine equations is undecidable. Hence the
overall result is often called MRDP Theorem (for an exposition, see, e.g., Davis 1973; Matiyasevich 1993).

The tenth on Hilbert’s famous list of important open problem in mathematics from 1900 asks for a decision method for
the so-called Diophantine equations. Despite the unfamiliar term “Diophantine,” what is at issue here is truly
elementary. Consider any equation with one or more variables and with integer coefficients, which involves only
addition and multiplication, such as \( (x^2 + y^2 = 2) \), or \( (3x^2 + 5y^2 + 2xy = 0) \). If real-number solutions are sought,
one usually speaks simply about an “equation”. However, in number theory, typically a solution is sought consisting only
of integers. That makes a great difference. The former of the above equations has infinitely many solutions among real
numbers, but only four among integers. The equation \( (x^2 + y^2 = 3) \) also has infinitely many real solutions but no
integer solutions. When the focus is on the integer solutions, one talks about “Diophantine equations” (after the ancient
number theorist Diophantus of Alexandria).

For a positive solution of Hilbert’s tenth problem, it would have sufficed to present a particular concrete method which
would have intuitively been a “mechanical” decision method. However, Turing’s pioneering analysis of the notion of
decision method brought into focus the possibility of a negative solution. Beginning in the early 1950s, Julia Robinson
and Martin Davis worked on this problem, later joined by Hilary Putnam. As a result of their collaboration, the first
important result in this direction was achieved. Call an equation “an exponential Diophantine equation” if it involves also
exponentiation, as well as addition and multiplication (that is, one can have both constants and variables as
exponents); naturally, the focus is still in the integer solutions. Davis, Putnam, and Robinson (1961), showed that the
problem of solvability of exponential Diophantine equations is undecidable. In 1970, Yuri Matiyasevich added the final
missing piece, and demonstrated that the problem of the solvability of Diophantine equations is undecidable. Hence the
overall result is often called MRDP Theorem (for an exposition, see, e.g., Davis 1973; Matiyasevich 1993).

The essential technical achievement was that all semi-decidable (recursively enumerable) sets can be given a
Diophantine representation, i.e., they can be represented by a simple formula of the form \( \exists x_1 \ldots \exists x_n \exists t (s(t, \ldots ,x_n) = t) \), where \( \exists (s = t) \) is a Diophantine equation. More exactly, for any given recursively enumerable set \( \exists (S) \), there is a
Diophantine equation \( \exists (s(y, x_1 , \ldots ,x_n) = t(y, x_1 , \ldots ,x_n)) \) such that \( \exists (\text{boldsymbol}[n] \in S) \) if and only if \( \exists x_1 \ldots \exists x_n \exists (\text{boldsymbol}[n], x_1 , \ldots ,x_n) = t(\text{boldsymbol}[n], x_1 , \ldots ,x_n) \).

As there are semi-decidable (recursively enumerable) sets which are not decidable (recursive), the general conclusion
follows immediately:

**MRDP Theorem**

There is no general method for deciding whether or not a given Diophantine equation has a solution.

This also provides an elegant variant of the incompleteness theorems dealing with Diophantine equations:

**Corollary**

For any 1-consistent axiomatizable formal system \( \exists (F) \) there are Diophantine equations which have no solutions but
cannot be proved in \( \exists (F) \) to have no solutions.

(The question of avoiding the requirement of 1-consistency here is tricky; see Dyson, Jones and Shepherson 1982.)

**4.5 Concrete Cases of Unprovable Statements**

The undecidable sentences provided by Gödel’s proofs are (if written out) extremely complicated formulas with no
intuitive significance, construed only for the purposes of the incompleteness proofs. The question then arises whether
there are any simple and natural mathematical statements which are likewise undecidable in chosen basic theories,
e.g., in \( \exists (PA) \). There are now various specific statements with clear mathematical content which are known to be
undecidable in some standard theories (though, just how natural even these are has been disputed; see Feferman
1989b). Some well known, natural examples are listed below, beginning with some quite natural mathematical
statements which are independent of \( \exists (PA) \), and proceeding to more and more powerful theories. Sometimes such results are
called variants of Gödel’s theorem, or their proofs of independence alternative proofs of Gödel’s theorem, but this is
misleading: interesting as they may be, they don’t have the generality of Gödel’s theorems proper, but only provide
statements independent of a particular theory.

It is often stated that before the celebrated Paris-Harrington theorem (see below), no such natural independent
Arithmetic and other standard theories were complete. Apparently Gödel was also impressed by Brouwer, who in his mathematics were not decidable. Gödel himself had mentioned the possibility of an undecidable problem concerning and von Neumann, in contrast to the dominant spirit in Hilbert's program, had considered it possible that logic and the possibility of incompleteness in the context of set theory was discussed by Bernays and Tarski already in 1928, 5. The History and Early Reception of the Incompleteness Theorems

lacking a truth-value. Sometimes Paul Cohen's celebrated result that the Continuum Hypothesis (CH) is independent of ZFC (Cohen 1963, 1964; see the entry on independence and large cardinals). However, this case is very different. In all the above independence results the relevant statements are still theorems of mathematics, taken as shown to be true (the last case, which requires large cardinal axioms that go beyond ZFC, is more controversial; still, at least many set-theoricians find such axioms plausible). And with the first incompleteness theorem itself, the truth of the unprovable statement easily follows, given that the assumption of the consistency of the system is indeed correct. However, in the case of Cohen's result, there is absolutely no indication whether CH should be considered true, false, or perhaps lacking a truth-value.

5. The History and Early Reception of the Incompleteness Theorems

Gödel's results were certainly surprising, but some sort of incompleteness phenomenon was not totally unexpected. The possibility of incompleteness in the context of set theory was discussed by Bernays and Tarski already in 1928, and von Neumann, in contrast to the dominant spirit in Hilbert's program, had considered it possible that logic and mathematics were not decidable. Gödel himself had mentioned the possibility of an undecidable problem concerning real numbers in his thesis in 1929 (see Dawson 1985). Hilbert (1928), on the other hand, had assumed that Peano Arithmetic and other standard theories were complete. Apparently Gödel was also impressed by Brouwer, who in his
The correctness of Gödel's theorems remained the subject of lively debate throughout the 1930s (see Dawson 1985). In much later in (Post 1941).

On the other hand, it is fair to say that Emil Post had in some respects anticipated Gödel's discoveries. He obtained abstract versions of incompleteness results apparently already in 1922. In particular, he observed that his methods would provide a statement undecidable in Principia Mathematica. These results were, however, based on Post's own version of the "Church-Turing thesis", with which he was dissatisfied, and his work was left unpublished. It was reported much later in (Post 1941).

The reception of Gödel's results was mixed. Some important figures in the field of logic and the foundations of mathematics quite quickly assimilated the results and understood their relevance, but there was also quite a lot of misunderstanding and resistance (for detailed accounts of the reception, see Dawson 1985; Mancosu 1999).

Gödel revealed his results to Carnap in Vienna on 26 August 1930, and announced his result (the first theorem) in a casual discussion remark in the famous Königsberg Conference on September 7, 1930. John von Neumann, who was in the audience and was at the time working in the context of Hilbert's program, immediately understood the great importance of the result. On 20 November he wrote a letter to Gödel on a "remarkable" corollary of Gödel's result he had discovered: the unprovability of consistency (the second theorem). Meanwhile Gödel himself, however, had found the same idea and had already sent the final version of his article, which now contained also a statement of the second incompleteness theorem, for publication. The article was published in January 1931 (Gödel 1931; helpful introductions to Gödel's original paper are Kleene 1986 and Zach 2005). The word quickly started to spread of these results which apparently had great importance for the foundations of mathematics—though views on what really was the moral varied. Paul Bernays, perhaps the most important collaborator of Hilbert, showed great interest in the results, though he first had difficulties in understanding them properly. His active correspondence with Gödel also shows that Gödel was already at the time fully aware of the undefinability of truth.

As to the Diagonalization Lemma, actually Gödel himself originally demonstrated only a special case of it, that is, only for the provability predicate. The general lemma was apparently first discovered by Carnap 1934 (see Gödel 1934, 1935). Still more general versions, for formulas with free variables, were presented in Ehrenfeucht & Feferman 1960 and Montague 1962 (see Smoryński 1981).

The original approach focused on his specific though very comprehensive system $\mathbf{P}$ and its (primitive recursive) extensions, some doubted the generality of Gödel's results. Alonzo Church, for example, in a letter to Gödel in July 1932, suggested that Gödel's results would not apply to his system of $\lambda$-conversion (the system was later proved to be inconsistent by Kleene and Rosser). Gödel was anxious to generalize his discoveries, and extended the results to a wider class of systems in papers in 1932 and 1934. He also suggested that his methods would be applicable to standard systems of set theory (however, it was only after the satisfactory characterization of decidability and the Church-Turing thesis a few years later that it was possible to give a fully general formulation of the incompleteness theorems (see above); this was first done in Kleene 1936). The eminent set-theorist Ernst Zermelo directed some rather harsh criticism towards Gödel's work, but the two also corresponded on the topic. Zermelo seems to have had serious difficulties in understanding the relevant concepts and results.

In March 1933, Gödel received a letter from Paul Finsler, from Zürich, who suggested that he had already earlier (in Finsler 1926) done closely related work but with a more general relevance. Gödel replied that Finsler's system was not really defined at all. In his heated response, Finsler claimed that it was not necessary to be able to study a system for it to be sharply defined, and that there was no difference in principle between his ideas and Gödel's. In retrospect, it is quite clear that the approaches of Finsler and Gödel were very different: for Gödel's work, the notion of formalized system was essential, whereas Finsler rejected the very notion as artificially restrictive. In fact, it is far from clear that Finsler's ideas make any sense—whatever vague analogies there may be between them and Gödel's proof.

In the following years, Gödel continued his work on the problem of the consistency of arithmetic. He attempted to advance Hilbert's program by attempting to prove the consistency of arithmetic. However, Gödel's results showed that such a proof could not be achieved within the framework of arithmetic. Gödel's work contributed significantly to the development of intuitionistic logic.
6. Philosophical Implications—Real and Alleged

6.1 Philosophy of Mathematics

Of the various fields of philosophy, Gödel's theorems are obviously most immediately relevant for the philosophy of mathematics. To begin with, they pose, at least *prima facie*, serious problems for Hilbert's program (this issue is discussed in some detail in the section on the impact of incompleteness in the entry on *Hilbert's Program*). Then again, they have important consequences for *intuitionism* (see the entry on *intuitionism in the philosophy of mathematics*) (see also Gödel 1933, 1941; Raatikainen 2005).

There has been some dispute on the issue as to whether Gödel's theorems conclusively refute *logicism* (see the entry on *logicism*). Henkin (1962) and Musgrave (1977), for example, argue it does; Sternfeld (1976) and Rodríguez-Consuegra (1993) disagree (see also Hellman 1981; Raatikainen 2005).

Gödel himself developed an argument against the conventionalist philosophy of mathematics of logical positivism, and of Carnap's in particular, based on the incompleteness results (Gödel 1953/9). It is discussed in Goldfarb and Ricketts 1992; Ricketts 1995; Goldfarb 1995; Crocco 2003; Awodey & Carus 2003, 2004; Tennant 2008.

6.2 Self-evident and analytical truths

One can also give more general epistemological interpretations of Gödel's theorems. Quine and Ullian (1978), for example, consider the traditional philosophical picture that all truths could be proved by self-evident steps from self-evident truths and observation. They then point out that even the truths of elementary number theory are presumably not in general derivable by self-evident steps from self-evident truths (Quine & Ullian 1978: 64–65). Hilary Putnam (1975) in turn submits that, under a certain natural understanding of “analytic”, there must be, by Gödel's theorems, synthetic truths in mathematics. In fact, Gödel himself made remarks in a very similar spirit that even the theory of integers is demonstrably non-analytic (Gödel 1944).

6.3 ‘Gödelian’ arguments against mechanism

There have been repeated attempts to apply Gödel's theorems to demonstrate that the powers of the human mind outrun any mechanism or formal system. Such a Gödelian argument against mechanism was considered, if only in order to refute it, already by Turing in the late 1940s (see Piccinini 2003). An unqualified anti-mechanist conclusion was drawn from the incompleteness theorems in a widely read popular exposition, *Gödel's Theorem*, by Nagel and Newman (1958). Shortly afterwards, J.R. Lucas (1961) famously proclaimed that Gödel's incompleteness theorem proves that Mechanism is false, that is, that minds cannot be explained as machines.

He stated that

> given any machine which is consistent and capable of doing simple arithmetic, there is a formula it is incapable of producing as being true … but which we can see to be true.

More recently, very similar claims have been put forward by Roger Penrose (1989, 1994). John Searle (1997) has joined the discussion and partly defended Penrose against his critics. Crispin Wright (1994, 1995) has endorsed related ideas from an intuitionistic point of view (for criticism, see Dettlesfen 1995). They all insist that Gödel's theorems imply that the human mind infinitely surpasses the power of any finite machine or formal system.

These Gödelian anti-mechanist arguments are, however, problematic, and there is wide consensus that they fail. The standard response to this argument goes along the following lines (this objection goes back to Putnam 1960; see also Boolos 1968, Shapiro 1998): The argument assumes that for any formalized system, or a finite machine, there exists the Gödel sentence which is unprovable in that system, but which the human mind can see to be true. Yet Gödel's theorem has in reality a conditional form, and the alleged truth of the Gödel sentence of a system depends on the assumption of the consistency of the system. The anti-mechanist's argument thus also requires that the human mind can always see whether or not a given formalized theory is consistent. However, this is highly implausible (cf. Davis 1990). Lucas, Penrose and others have attempted to reply to such criticism (see, e.g., Lucas 1996; Penrose 1995, 1997). For detailed criticism of Penrose, see Boolos 1990; Davis 1990, 1993; Feferman 1995; Lindström 2001; Pudlák 1999; Shapiro 2003; many of these considerations are also relevant for what Lucas says).

6.4 Gödel and Benacerraf on Mechanism and Platonism
Interestingly, Gödel himself also presented an anti-mechanist argument although it was more cautious and only published posthumously (in his *Collected Works*, Vol. III, in 1995). That is, in his 1951 Gibbs lecture, Gödel drew the following disjunctive conclusion from the incompleteness theorems:

either … the human mind (even within the realm of pure mathematics) infinitely surpasses the power of any finite machine, or else there exist absolutely unsolvable diophantine problems.

Gödel speaks about this statement as a “mathematically established fact” (Gödel 1951; for more discussion on Gödel’s disjunctive claim, see, e.g., Shapiro 1998). According to Gödel, the second alternative

seems to disprove the view that mathematics is only our own creation … that mathematical objects and facts … exist objectively and independently of our mental acts and decisions.

Gödel was nonetheless inclined to deny the possibility of absolutely unsolvable problems, and although he did believe in mathematical Platonism, his reasons for this conviction were different, and he did not maintain that the incompleteness theorems alone establish Platonism. Thus Gödel believed in the first disjunct, that the human mind infinitely surpasses the power of any finite machine. Still, this conclusion of Gödel follows, as Gödel himself clearly explains, only if one denies, as does Gödel, the possibility of humanly unsolvable problems. It is not a necessary consequence of incompleteness theorems.

Now Gödel was, unlike the later advocates of the so-called Gödelian anti-mechanist argument, sensitive enough to admit that both mechanism and the alternative that there are humanly absolutely unsolvable problems are consistent with his incompleteness theorems. His fundamental reasons for disliking the latter alternative are much more philosophical. Gödel thought in a somewhat Kantian way that human reason would be fatally irrational if it asked questions it could not answer (for critical discussion, see Kreisel 1967; Boolos 1995; Raatikainen 2005).

As a reaction to Lucas’ argument, but before the publication of Gödel’s Gibbs Lecture, Paul Benacerraf (1967) put forward more qualified conclusions that interestingly resemble some ideas of Gödel. He argued that it is consistent with all the facts that I am indeed a Turing machine, but that I cannot ascertain which one. For some critical discussion, see Chihara 1972 and Hanson 1971.

6.5 Mysticism and the existence of God?

Sometimes quite fantastic conclusions are drawn from Gödel’s theorems. It has been even suggested that Gödel’s theorems, if not exactly prove, at least give strong support for mysticism or the existence of God. These interpretations seem to assume one or more misunderstandings which have already been discussed above: it is either assumed that Gödel provided an absolutely unprovable sentence, or that Gödel’s theorems imply Platonism, or anti-mechanism, or both.

For more discussion about the philosophical aspects of the incompleteness theorems, see Raatikainen 2005 and Franzén 2005.

Further reading

A standard reference for the incompleteness theorems is:


There are several introductory textbooks in mathematical logic which give a good exposition of the incompleteness theorems and related topics; for example:


Two books that are dedicated to the incompleteness theorems are:


Another useful book on the incompleteness theorems and related topics is:


A comprehensive, more advanced book on these themes is:


Another useful book, including also some more advanced topics is:
The more philosophical aspects around the incompleteness theorems are surveyed in the following two sources (Franzén is an accessible, informal, and yet reliable, explanation of the incompleteness theorems):


The following two papers survey various issues around the first incompleteness theorem:


Finally, there is an open-source e-book that contains a presentation of the incompleteness theorems:


**Bibliography**


Academic Tools
- How to cite this entry.
- Preview the PDF version of this entry at the Friends of the SEP Society.
- Look up this entry topic at the Internet Philosophy Ontology Project (InPhO).
- Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources
- Gödel on the Net, by Torkel Franzén (Luleå University of Technology).
- Papers and Reviews by Sol Feferman on Gödel's incompleteness theorem:

Related Entries

Acknowledgments
The author would like to thank Richard Zach for his careful and valuable comments on the earlier drafts of this entry.
The author and SEP editors would like to thank Richard O’Callaghan for bringing to our attention that, in an earlier version of the supplement on the Diagonalization Lemma, the definition of the substitution function, although standard, didn’t suffice for the purposes of the proof.