Tarski's undefinability theorem

Tarski's undefinability theorem, stated and proved by Alfred Tarski in 1936, is an important limitative result in mathematical logic, the foundations of mathematics, and in formal semantics. Informally, the theorem states that *arithmetical truth cannot be defined in arithmetic*.

The theorem applies more generally to any sufficiently strong formal system, showing that truth in the standard model of the system cannot be defined within the system.

Contents

History

In 1931, Kurt Gödel published the incompleteness theorems, which he proved in part by showing how to represent the syntax of formal logic within first-order arithmetic. Each expression of the formal language of arithmetic is assigned a distinct number. This procedure is known variously as Gödel numbering, *coding* and, more generally, as arithmetization. In particular, various *sets* of expressions are coded as sets of numbers. It turns out that for various syntactic properties (such as *being a formula*, *being a sentence*, etc.), these sets are computable. Moreover, any computable set of numbers can be defined by some arithmetical formula. For example, there are formulas in the language of arithmetic defining the set of codes for arithmetic sentences, and for provable arithmetic sentences.

The undefinability theorem shows that this encoding cannot be done for semantic concepts such as truth. It shows that no sufficiently rich interpreted language can represent its own semantics. A corollary is that any metalanguage capable of expressing the semantics of some object language must have expressive power exceeding that of the object language. The metalanguage includes primitive notions, axioms, and rules absent from the object language, so that there are theorems provable in the metalanguage not provable in the object language.

The undefinability theorem is conventionally attributed to Alfred Tarski. Gödel also discovered the undefinability theorem in 1930, while proving his incompleteness theorems published in 1931, and well before the 1936 publication of Tarski's work (Murawski 1998). While Gödel never published anything bearing on his independent discovery of undefinability, he did describe it in a 1931 letter to John von Neumann. Tarski had obtained almost all results of his 1936 paper *Der Wahrheitsbegriff in den formalisierten Sprachen* between 1929 and 1931, and spoke about them to Polish audiences. However, as he emphasized in the paper, the undefinability theorem was the only result not obtained by him earlier. According to the footnote of the undefinability theorem (Satz I) of the 1936 paper, the theorem and the sketch of the proof were added to the paper only after the paper was sent to print. When he presented the paper to the Warsaw Academy of Science on March 21, 1931, he wrote only some conjectures instead of the results after his own investigations and partly after Gödel's short report on the incompleteness theorems "Einige metamathematische Resultate über Entscheidungsdefinitheit und Widerspruchsfreiheit", Akd. der Wiss. in Wien, 1930.

Statement of the theorem

We will first state a simplified version of Tarski's theorem, then state and prove in the next section the theorem Tarski actually proved in 1936. Let L be the language of first-order arithmetic, and let N be the standard structure for L. Thus (L, N) is the "interpreted first-order language of arithmetic." Each sentence x in L has a Gödel number $g(x)$. Let T denote the set of L-sentences true in N, and T^* the set of Gödel numbers of the sentences in T. The following theorem answers the question: Can T^* be defined by a formula of first-order arithmetic?

Tarski's undefinability theorem: There is no L-formula $True(n)$ that defines T^*. That is, there is no L-formula $True(n)$ such that for every L-formula A, $True(g(A)) \leftrightarrow A$ holds.

Informally, the theorem says that given some formal arithmetic, the concept of truth in that arithmetic is not definable using the expressive means that that arithmetic affords. This implies a major limitation on the scope of "self-representation." It *is* possible to define a formula $True(n)$ whose extension is T^*, but only by drawing on a metalanguage whose expressive power goes beyond that of L. For example, a truth predicate for first-order arithmetic can be defined in second-order arithmetic. However, this formula would only be able to define a truth predicate for sentences in the original language L. To define a truth predicate for the metalanguage would require a still higher "metametalanguage", and so on.

The theorem just stated is a corollary of Post's theorem about the arithmetical hierarchy, proved some years after Tarski (1936). A semantic proof of Tarski's theorem from Post's theorem is obtained by reductio ad absurdum as follows. Assuming T^* is arithmetically definable, there is a natural number n such that T^* is definable by a formula at level $\mathbf{\Sigma}^{\mathbf{0}}_{\boldsymbol{n}}$ of the arithmetical hierarchy. However, T^* is $\mathbf{\Sigma}^{\mathbf{0}}_{\boldsymbol{k}}$-hard for all k. Thus the arithmetical hierarchy collapses at level n, contradicting Post's theorem.

General form of the theorem

Tarski proved a stronger theorem than the one stated above, using an entirely syntactical method. The resulting theorem applies to any formal language with negation, and with sufficient capability for self-reference that the diagonal lemma holds. First-order arithmetic satisfies these preconditions, but the theorem applies to much more general formal systems.

Tarski's undefinability theorem (general form): Let (L,N) be any interpreted formal language which includes negation and has a Gödel numbering $g(x)$ such that for every L-formula $A(x)$ there is a formula B such that $B \leftrightarrow A(g(B))$ holds. Let T^* be the set of Gödel numbers of L-sentences true in N. Then there is no L-formula $True(n)$ which defines T^*. That is, there is no L-formula $True(n)$ such that for every L-formula A, $True(g(A)) \leftrightarrow A$ holds.

The proof of Tarski's undefinability theorem in this form is again by reductio ad absurdum. Suppose that an L- formula $True(n)$ defines T^*. In particular, if A is a sentence of arithmetic then $True(g(A))$ holds in N if and only if A is true in N. Hence for all A, the Tarski T-sentence $True(g(A)) \leftrightarrow A$ is true in N. But the diagonal lemma yields a counterexample to this equivalence, by giving a "Liar" sentence S such that $S \leftrightarrow \neg True(g(S))$ holds. Thus no L-formula $True(n)$ can define T^*. QED.

The formal machinery of this proof is wholly elementary except for the diagonalization that the diagonal lemma requires. The proof of the diagonal lemma is likewise surprisingly simple; for example, it does not invoke recursive functions in any way. The proof does assume that every L-formula has a Gödel number, but the specifics of a coding method are not required. Hence Tarski's theorem is much easier to motivate and prove than the more celebrated theorems of Gödel about the metamathematical properties of first-order arithmetic.

Discussion

Smullyan (1991, 2001) has argued forcefully that Tarski's undefinability theorem deserves much of the attention garnered by Gödel's incompleteness theorems. That the latter theorems have much to say about all of mathematics and more controversially, about a range of philosophical issues (e.g., Lucas 1961) is less than evident. Tarski's theorem, on the other hand, is not directly about mathematics but about the inherent limitations of any formal language sufficiently expressive to be of real interest. Such languages are necessarily capable of enough self-reference for the diagonal lemma to apply to them. The broader philosophical import of Tarski's theorem is more strikingly evident.

An interpreted language is *strongly-semantically-self-representational* exactly when the language contains predicates and function symbols defining all the semantic concepts specific to the language. Hence the required functions include the "semantic valuation function" mapping a formula A to its truth value $||A||$, and the "semantic denotation function" mapping a term t to the object it denotes. Tarski's theorem then generalizes as follows: *No sufficiently powerful language is strongly-semantically-self-representational.*

The undefinability theorem does not prevent truth in one theory from being defined in a stronger theory. For example, the set of (codes for) formulas of first-order Peano arithmetic that are true in N is definable by a formula in second order arithmetic. Similarly, the set of true formulas of the standard model of second order arithmetic (or n-th order arithmetic for any n) can be defined by a formula in first-order ZFC.

References

- J.L. Bell, and M. Machover, 1977. *A Course in Mathematical Logic* North-Holland.
- G. Boolos, J. Burgess, and R. Jeffrey, 2002. *Computability and Logic*, 4th ed. Cambridge University Press.
- J.R. Lucas, 1961. "Mind, Machines, and Gödel". Philosophy 36: 112–27.
- R. Murawski, 1998. Undefinability of truth. The problem of the priority: Tarski vs. Gödel. History and Philosophy of Logic 19, 153–160
- R. Smullyan, 1991. *Godel's Incompleteness Theorems*. Oxford Univ. Press.
- R. Smullyan, 2001. "Gödel's Incompleteness Theorems". In L. Goble, ed., *The Blackwell Guide to Philosophical Logic*, Blackwell, 72–89.
- A. Tarski (1936). "Der Wahrheitsbegriff in den formalisierten Sprachen" (PDF). *Studia Philosophica*. 1: 261–405. Retrieved 26 June 2013.
- A. Tarski, tr J.H. Woodger, 1983. "The Concept of Truth in Formalized Languages". English translation of Tarski's 1936 article. In A. Tarski, ed. J. Corcoran, 1983, *Logic, Semantics, Metamathematics*, Hackett.