A Doubt about Gödel's Theorem

LEO LIBERTI

Centre for Process Systems Engineering
Imperial College of Science, Technology and Medicine
London, UK SW7 2BY

`l.liberti@ic.ac.uk`

28th December 2000

Abstract

It is relatively easy to have doubts when first confronted with Göedel's First Incompleteness Theorem. In particular, it is easy to mistake "truth" for "provability". In this short note I try to explain how to dispel this common doubt.

A Doubt about Gödel's Theorem

The idea for this short account was born when I stumbled upon an apparent "contradiction in terms", or syntactical contradiction, while reading Gödel's First Incompleteness Theorem. This is not to say that I found the contradiction in Gödel's proof. Rather, while trying to write down the proof to see if I had correctly understood it, I followed a train of thought that brought me to a reasoning which had to be correct for Gödel's proof to work, but was apparently faulty. I was sure I had made a mistake somewhere, and of course I had. Nonetheless I think my understanding of Gödel's theorem is better now than it was before, and as I heard this doubt formulated by many people when they first try to understand Göedel's First Incompleteness Theorem, I thought a short note might help clarify the point.

I understand sources are a good point to start from. The following excerpts come from [vHe67]. In particular I am quoting from *Kurt Gödel, Over formally undecidable propositions of Principia Mathematica and other similar systems*, which is the original article that Gödel wrote about undecidability of self-referential propositions.

The goal Gödel aims at is to prove that the theories described in Russell and Whitehead's *Principia Mathematica* are incomplete. The same can be said of Zermelo and Fränkel's ZF axiomatic system and generally all in formal systems that embed the natural numbers and are consistent (that is, the system doesn't prove any contradiction). A precise definition of "incompleteness" that can be expressed *inside* the theory is the following.

1.1 Definition
The theory T is incomplete if and only if there exists a sentence ϕ in T such that T does not prove ϕ and T does not prove $\neg\phi$.

This definition is sound only as long as the concepts of "theory", "proof" and "existence of a sentence in a theory" are precisely defined, but this would make this short note a lengthy article. For precise definitions of these terms, see .... For the purposes of this note, it suffices to say that the language which is used to express mathematical manipulations is composed of a finite number of constants (i.e. $\neg$ = "not", $\vee$ = "or", $\exists$ = "exists", $\in$= "belongs to", $\rightarrow$= "implies", = which means "is equal to", and brackets which indicate precedence) and a countable number of symbols indicating variables (v_i). Other symbols can be derived from these basic ones. For example, $\wedge$, $\forall$, $\leftrightarrow$ are derived symbols. There are some rules on combining the symbols in the language, but these are mostly pretty obvious (e.g. "$\neg \vee v_1(\exists$" is not a valid sentence, whereas "$\exists v_1(v_1 \in v_1)$" is valid). From now on, whenever I speak of a "sentence", I shall mean a *valid* sentence. A proof is a finite sequence of sentences such that the $i+1$-th sentence is obtained from the i-th sentence by an application of logic inference rules like *modus ponens*. A sentence is provable if there is a proof such that the first sentence of the proof is provable. An *axiom* is a sentence which we decide to be provable *a priori*. For example,

$$\forall x \forall y (\forall z (z \in x \leftrightarrow z \in y) \rightarrow x = y)$$

is an axiom. It says that if two sets contain exactly the same elements, then they are equal. A theory consists of a countable set of axioms and all of the sentences that can be derived from the axiom by proofs. A sentence exists in a theory if there is a proof that derives it from the axioms.

Before quoting Gödel's proof let me add that the source of my confusion when I first read the proof stems from definition (1.1). I had read [Hof80] many years before, and I had formed the opinion that a theory is incomplete if it can express a true statement but cannot prove it. In fact, this is expressed on the very last page of part I of [Hof80].

1.2 Definition
A theory T is incomplete if and only if there exists a ϕ in T such that ϕ is true and T does not prove ϕ.

As we shall see, (1.1) implies (1.2), even though they exist in different contexts. In the first there is no reference whatsoever to the concept of "truth", while it is present in the second. Hence, the first can be expressed and used in the theory, whilst the second needs to be "seen" from outside the theory, and thus can lead to confusion if it is applied without careful thought.

1.3 Proposition
Definition (1.1) implies (1.2).

Proof. Suppose we find ϕ in T such that $T \nvdash \phi$ and $T \nvdash \neg\phi$. By dichotomy, exactly one of ϕ and $\neg\phi$ must be true. Hence there is a true sentence that T cannot prove. □

If one assumes that "a contradiction" and "a false sentence" are the same thing, it is also possible to show the full equivalence of the two definitions, for let ϕ be true and unprovable, then $\neg\phi$ is false. Since T is consistent, T can't prove $\neg\phi$, *qualis erat demonstrandum.* This is an important point. A "contradiction" is something that can be expressed *inside* T (as in $1 = 0$), whilst the concept of "falseness", like that of "truth", involves an interpretation of the symbols, and may not necessarily be expressed in T.

The proof that PM is incomplete, as given by Gödel, extends for the whole of his article, which is lengthy, so I will quote the unrigorous description of the proof found in the first pages of his article.

1.4 Theorem (Gödel's proof of incompleteness of PM)
Formulae of a formal system (take PM) are finite sequences of primitive symbols (variables, constants, brackets or separation signs), and it is easy to define rigorously which sentences are valid and which aren't. Similarly proofs are finite sequences of formulae. From a metamathematical point of view the choice of the symbols to be used is irrelevant, so we shall take the natural numbers. Consequently, a formula is a finite sequence of natural numbers, and a formal proof is a finite sequence of finite sequences of natural numbers. [...] One can show that PM can express the concepts of "formula", "formal proof" and "provable sentence". In particular it is possible to find a formula $F(v)$ in PM (with exactly one free variable v) which, when interpreted with the meaning of the terms in PM, asserts: "v is a provable sentence". This allows us to construct an undecidable proposition in PM, that is, a sentence A for which neither A nor $\neg A$ are provable from PM.

Proof. Suppose we have ordered all formulae of PM with exactly one free variable in a sequence $R(n)$. We observe in passing that both concepts "formula with exactly one free variable" and the well-ordering R can be defined in PM. We shall express with $[R(n), m]$ the sentence obtained when substituting m in the free variable of the formula $R(n)$. $[R(n), m]$ can also be expressed in PM. We now define a class K such that

$$\forall n \in \mathbb{N}(n \in K \leftrightarrow \neg\mathrm{Prv}([\mathrm{R(n),n}])) \tag{1}$$

where $\mathrm{Prv}(\phi)$ means: ϕ is a provable sentence. Since everything we have written can be expressed in PM, so can K; in other words, there exists a formula with exactly one free variable S such that $[S, n]$, when interpreted with the meaning of the terms of PM, says "$n \in K$". Since S has exactly one free variable, there must be a natural number q such that $S = R(q)$. Let ϕ be the sentence $[R(q), q]$, and we show that ϕ is undecidable in PM.

1. Suppose that ϕ is provable: then ϕ should be true. In this case $q \in K$ by definition of $R(q)$, and hence $\neg\mathrm{Prv}(\phi)$ by definition of K. Contradiction.

2. Suppose that $\neg\phi$ is provable: then $\neg\phi$ should be true, hence $q \notin K$ and $\mathrm{Prv}(\phi)$. That is, both ϕ and $\neg\phi$ are provable, which is impossible because PM is consistent[1].

[1] I.e. no contradiction can be proved from PM.

Hence we must conclude that ϕ is undecidable in PM, that is, PM is incomplete (as defined in (1.1)). □

It is not quite correct to say that this is a "proof". As Gödel himself puts it, this is the idea of the proof. The proof goes on at length to show that all of the concepts used in the reasoning can in fact be expressed in the terms of PM. Nonetheless, the basic idea contained in the proof is to employ a sentence ϕ that says: "ϕ cannot be proved".

If one stops for a moment to think, it seems evident that instead of the points 1. and 2. in the proof of theorem (1.4), one could have argued in the following way.

1.5 Theorem
PM is incomplete.

Proof. Let ϕ stand, as above, for $[R(q), q]$. Suppose $\mathrm{Prv}(\phi)$. Then $q \in K$ and hence $\neg\mathrm{Prv}(\phi)$, contradiction. Now, instead of saying "suppose $\mathrm{Prv}(\neg\phi)$" as in the above proof, we use the dichotomy of every logical assertion (in this case $\mathrm{Prv}(\phi)$). Since we have proved that assuming $\mathrm{Prv}(\phi)$ leads to a contradiction, we must conclude that the only possible way out is $\neg\mathrm{Prv}(\phi)$, that is, that ϕ is *not* provable. We now interpret ϕ and we notice that it says: "ϕ is not provable". Since we have just proved that ϕ is not provable, we must infer that ϕ is true. Hence we have found a sentence (namely ϕ), for which we have proved that:

1. ϕ is not provable
2. ϕ is true

Consequently we have shown that PM is incomplete (as defined in (1.2)). □

Summing up, we are in the situation described by the following phrase:

We have proved ϕ cannot be proved to be true. Then we have proved that ϕ is true.

This, however, is a contradiction in terms; and this is exactly what my first reaction to Göedel's theorem has been. I first thought the mistake was in the proof given in (1.5), but couldn't find it. Successively I stumbled upon the following passage from [Hof80]:

> TNT[2] incorporates valid methods of reasoning, and therefore TNT never has falsities for theorems. In other words, anything which is a theorem of TNT expresses a truth. So if[3] G were a theorem, it would express a truth, namely "G is not a theorem". [...] By being a theorem, G would have to be a falsity. Relying on our assumption that TNT never has falsities for theorems, we'd be forced to conclude that *G is not a theorem*. This is all right; it leaves us, however, with a lesser problem. Knowing that G is not a theorem, we'd have to concede that G expresses a truth. [...] We have found a string which expresses a true statement yet the string is not a theorem.

This confirms that the proof given in (1.5) is in fact correct. I did not go as far as doubting Hofstadter. The mistake resides in the phrase in italics:

We have proved ϕ cannot be proved to be true. Then we have proved that ϕ is true.

It is true that we have proved ϕ cannot be proved to be true. But we have *not* proved, at least not *inside* PM, that ϕ is true. We have *interpreted* the meaning of ϕ in order to decide that it is true, we

[2]By TNT Hofstadter means a system which is equivalent to the Peano Arithmetics.
[3]G is the Hofstadter's construction of the self-referential sentence $[R(q), q]$.

haven't employed any mechanically logical method. I must thank my friend Fabio Roda, a distinguished philosopher who has been working on automated reasoning for some time, for the explanation that solved my doubt. Dr. Roda says:

> Gödel doesn't prove that ϕ is unprovable and true... You have never "written down on paper" the proof that ϕ is true.

I then found this passage in [Hof80] saying exactly the same thing.

> A string of TNT has been found; it expresses, unambiguously, a statement about certain arithmetical properties of natural numbers; moreover, by reasoning outside the system we can determine not only that the statement is a true one, but also that the string fails to be a theorem of TNT.

The phrase "by reasoning outside the system" is crucial. Strangely enough, we should only speak about truth when we interpret sentences, not when we use mechanical rules of derivation. In the latter case, we should talk about "provability". When the proof of a theorem is expressed inside the theory then that theorem is *provable*. As it turns out, there is a better definition of "truth" than that of interpreting a phrase, but it seems to be less general than the concept of truth every human being possesses. It is possible to define truth inside a system by posing limitations to the truth. A sentence might be true inside a certain model[4] and false inside another model, even while using the same language and the same axioms! Alessandro Andretta, an Associate Professor at Turin University, speaks about Gödel's theorem and concludes:

> Note that I have NEVER used the word "true" but only "provable". The two concepts are different. In order to speak about "true" I have to consider a model inside which to work. This assumes a certain quantity of set theory. ω is a model[5] for PA[6] but it is not the only one. There are non-standard models of PA, some of which satisfy "PA + PA is ω-inconsistent[7]", and other curiosities.

References

[Hof80] Douglas R. Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid.* Penguin, London, 1980.

[vHe67] Jean van Heijenoort (ed.). *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931.* Harvard University Press, Cambridge, MA, 1967.

[4] A model, loosely speaking, is a class of entities to which the axioms apply. Most commonly, it is a class of sets.

[5] ω is the same as $\mathbb{N}$ union $\{0\}$.

[6] Peano Arithmetics.

[7] A system T is ω-inconsistent if there is a formula $\psi(n)$ such that, for any fixed n, T proves $\psi(n)$ but T can't prove the sentence $\forall n \psi(n)$.