How do we know that G is true?

Virgil Drăghici
Dept. of Philosophy
Babeş-Bolyai University
Cluj-Napoca, Romania
E-mail: vidra007@gmail.com

Abstract: Two philosophical arguments, e.g. that the meaning of an expression outruns its use and that the human arithmetical thinking is not entirely algorithmic (Argument Lucas/Penrose), base their theses on Gödel's (first) incompleteness theorem. But both in these arguments and in some of their criticisms the word "true" is often used ambiguous: it swings between a licit metamathematical use and an illicit transfer of it in a formal system. Our aim is to strengthen the dividing line between the licit and illicit of it by considering the similarities of the two philosophical arguments, via Gödel's theorem (section 1), some limits of provability in Peano Arithmetic (PA) (section 2), and the ways we know that Gödel's sentence G is true (section 3). As regards this last point, we show that an argument for the G's truth can be developed only in a nonconservative extension of PA (section 4), by using a Δ-uniform reflection principle (with or without the truth predicate). In all these considerations no ambiguity of truth occurs.

Keywords: Incompleteness theorems, Lucas-Penrose argument, Peano Arithmetic, provability, truth, reflection principles.

1. Preliminaries

1.1. Gödel's theorem and some of its versions

One of the greatest discoveries of the 20th century, with logical, mathematical and philosophical implications, is the incompleteness phenomenon.[1] Such a celebrated result was mainly exploited in two famous arguments, firstly, in the argument according to which our arithmetical understanding outruns any systematic description (or meaning transcends use), and secondly, in the antimechanistic view of mind (or Lucas-Penrose argument) according to which "[h]uman mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth".[2] In order to set out the aim of this paper and the related arguments let us briefly review Gödel's results and their connection to the above noted arguments. This preliminary part is essential for the considerations of the next sections.

The basic system of our discussions will be Peano Arithmetic (PA). PA is said to be *consistent* if for no closed formula (sentence) α holds: $\mathrm{PA}\vdash\alpha$ and $\mathrm{PA}\vdash\neg\alpha$. PA is said to be ω-consistent if for no formula $\alpha(\mathrm{x})$ the following holds: $\mathrm{PA}\vdash\alpha(\boldsymbol{0})$, $\mathrm{PA}\vdash\alpha(\boldsymbol{1})$, $\mathrm{PA}\vdash\alpha(\boldsymbol{2})$,... and $\mathrm{PA}\vdash\exists\mathrm{x}\neg\alpha(\mathrm{x})$. If in this definition $\alpha(\mathrm{x})$ is *primitive recursive*, then PA is said to be 1-*consistent*.

Let $Pf(x,y)$ be the following relation of intuitive arithmetic[3]: "x is (the Gödel number of) a proof in PA of the formula (with Gödel number) y". This is primitive recursive (also decidable), hence it is representable (expressible) in the formal PA by, say, a formula $\pi(\mathrm{x},\mathrm{y})$, i.e. for any natural numbers k_1, k_2 holds:

$$Pf(k_1,k_2) \text{ is true iff } \mathrm{PA}\vdash\pi(\boldsymbol{k}_1,\boldsymbol{k}_2) \tag{1}$$

$$Pf(k_1,k_2) \text{ is false iff } \mathrm{PA}\vdash\neg\pi(\boldsymbol{k}_1,\boldsymbol{k}_2) \tag{2}$$

where $\boldsymbol{k}_1$, $\boldsymbol{k}_2$ are the numerals for the corresponding natural numbers.

---

[1] Cf. K. Gödel [1].

[2] Cf. R. Penrose [1], 76.

[3] The distinction between *intuitive* and *formal* variables of arithmetic will be given by using x, y and x, y, respectively.

If for a relation R only (1) is the case, then it is called *weak* representable by the respective formula in PA. As is well-known, recursive enumerable sets (or Σ_1-sets) are weakly representable, in contrast with decidable sets which are representable.

If a formula α is *provable* in PA, then we say that *there is* a proof *x* of the formula *y*. This is an example of a recursively enumerable relation, weakly representable in PA by the open formula $\exists x\pi(x,y)$.[4] This formula is the provability predicate for PA: Bew(y).

There are many ways to construct Gödel's sentence G. The simplest one is by the application of diagonal lemma.

Diagonal lemma (DL). If $\alpha(y)$ is a formula of PA containing y free, then there is a sentence G such that

$PA \vdash S \equiv \alpha \ulcorner S \urcorner$

where $\ulcorner S \urcorner$ is the formal name of the sentence S (or the numeral for the Gödel number of S).

Let us now consider the formula $\forall x \neg \pi(x,y)$, where $\pi(x,y)$ is the formula of PA representing the proof relation *Pf(x,y)* in PA. Then the intuitive meaning of this universally quantified formula is: there is no proof in PA of the formula *y*. By one application of DL we have

$$PA \vdash G \equiv \forall x \neg \pi(x, \ulcorner G \urcorner) \tag{3}$$

that is, the sentence G is provably *equivalent* to a formula asserting, via arithmetization, that G is not provable in PA. The Gödel's first incompleteness theorem can now be stated.

Th G1. *a) If* PA *is consistent, then* $PA \nvdash G$.

---

[4] We say that $\exists x\pi(x,y)$ weakly represents the *theoremhood* in PA, or the set *Th* of theorems of PA.

b) If PA *is* ω*-consistent, then* PA$\nvdash\neg$G.

The Gödel's second incompletess theorem is stated in following terms: if PA is consistent, then the formula Con(PA) expressing in the language L_{PA} of PA the consistency of PA is not provable in PA, i.e.

Th G2. *If* PA *is consistent, then* PA$\nvdash$Con(PA).

As is well-known, Th G2 is just Th G1a) *formalized*.

Actually, in the proof of Th G1b) only the 1-consistency of PA is needed.[5]

Penrose uses Th G1 in order to argue against the mechanistic view of the mind. His version of Gödel's result is stated in terms of Turing machines[6] as follows:

Gödel's theorem (Turing version). *If* $A(q,n)$ *is a sound procedure for ascertaining the non-halting of the computation* $C_q(n)$, *then it is incomplete*.

More detailed, $A(q,n)$ is sound if whenever A stops on the input (q,n), the computation $C_q(n)$ does not stop. By using Cantor's diagonal method[7] we can argue that for some k the computation $C_k(k)$ does *not* stop and $A(k,k)$ cannot stop either. So, if the halting of $A(q,n)$ is a sufficient condition for the non-halting of a Turing machine, then it is not a necessary one. So if $A(q,n)$ is sound, then it is incomplete.

Is not hard to see that Th G1 a) and Turing version are connected, given the fact that every formal system can be recast as a theorem-proving machine and vice-versa. Indeed, Turing version is similar to Kleene's

---

[5] Cf. the variant of *Gödel's theorem* below, via the equivalence (5).
[6] Cf. R. Penrose [1], 74-75.
[7] Turing's proof of nondecidability of the halting problem for Turing machines is based on such diagonal arguments; cf. A. Turing [1].

generalization[8] of Gödel's theorem, according to which there is no correct and complete formal system for the intuitive predicate $(y)\overline{T_1}(x,x,y)$. The "translation" of one form in the other can be given in this way. If S is a formal system and $\alpha(x,y)$ is a formula of S expressing the intuitive predicate $P(x,y)$, then S is *sound* (correct) for P if this conditional holds: if $\vdash \alpha(\boldsymbol{q},\boldsymbol{n})$, the $P(q,n)$, for any pairs of numbers (q,n). And S is *complete* for P if the converse is the case. By putting $P(q,n)$ iff $C_q(n)$ does not stop, and $A(q,n)$ stops iff $S \vdash \alpha(\boldsymbol{q},\boldsymbol{n})$, the "translation" is accomplished. With $P(x,x)$ for Kleene's predicate $(y)\overline{T_1}(x,x,y)$, the Kleene's form of Gödel's theorem follows: *if* S *is sound for the intuitive predicate* $P(x,y)$ *then it is not complete for it*, i.e. there is a k such that $P(k,k)$ is true and $S \nvdash \alpha(\boldsymbol{k},\boldsymbol{k})$.[9]

But the Gödelian sentence G is equivalent to the π_1-sentence $\forall x \neg\pi(x,\boldsymbol{g})$. By some fundamental results in proof theory[10] the following holds:

$$\mathrm{Con(S)} \equiv \pi_1\text{-soundness(S)} \tag{4}$$

$$1\text{-Con(S)} \equiv \Sigma_1\text{-soundness(S)}, \tag{5}$$

where Con(S) is the formula of S expressing its consistency, and $\pi_1(\Sigma_1)$-soundness(S) is the sentence expressing the soundness of S restricted to π_1 and Σ_1-sentenes, respectively. By these equivalences, Gödel's theorem for PA can be restated in this form:

Gödel's Theorem. *a) If* PA *is* π_1*-sound, then* $\mathrm{PA} \nvdash \mathrm{G}$.

b) If PA *is* Σ_1*-sound, then* $\mathrm{PA} \nvdash \neg\mathrm{G}$.

[8] Cf. S.C. Kleene [1], Th. XIII (Part I), p. 302.
[9] Cf. S.C. Kleene [1], Th. XIII (Part II), p. 303.
[10] Cf. C. Smorynski [1]. Th. 4.1.4, Th. 4.2.2.

1.2. Two arguments based on Gödel's theorem and some criticisms

This correspondence between Gödel's theorem and its Turing version reveals us the connection between the two mentioned arguments based on Gödel's result: our intuitive arithmetical understanding cannot be encapsulated neither in a formal system (i.e. the meaning of an expression transcends its use[11]), nor, *equivalently*, by using of some set of algorithmic procedures like Turing machines. Referring to the second case, Penrose says[12] that

> [...] if we *know* that A is sound, then we *know* that $C_k(k)$ does not stop. Thus, we know something that A is unable to ascertain. It follows that A *cannot* encapsulate our understanding.

Both arguments have raised many objections. But many times even such objections use the word "true" in a careless way, obscuring by that the critical analysis they intend to do. Let us consider two notable examples, directly related to the above arguments.

In referring to a line of argumentation as that contained in the quotation from Penrose given above Hilary Putnam considers that the use of Gödel's result in such an argument

> [...] is a misapplication of Gödel's theorem, pure and simple. Given an arbitrary machine T, all I can do is find a proposition [G] such that I can prove:

[11] E.g. the meaning of "true of all natural numbers" outruns any attempt to be characterized by a formal system.
[12] Cf. R. Penrose [1], 75.

> (3) If T is consistent, [G] is true,
>
> where [G] is undecidable by T if T is in fact consistent. However, T can perfectly well prove (3) too! And the statement [G], which T *cannot* prove (assuming consistency), I cannot prove either (unless I can prove that T is consistent, which is unlikely if T is very complicated).[13]

In referring to the question whether Gödel's result can be an argument against the idea that the meaning of an arithmetical concept outruns its use in a formal system, Michael Dummett wrote[14]:

> Considered as an argument to a hypothetical conclusion – that *if* the system is consistent, then $\forall xA(x)$ is true – this reasoning can of course be formalized in the system.

Both quotations contain a misuse of the word "true", a use of it beyond the limits admitted by the argument they give in order to criticize the antimechanist view or the thesis that meaning transcends use. So, can a Turing machine "perfectly well prove (3) too", as Putnam thinks? And can a formal system S prove the conditional "*if* the system is consistent, then $\forall xAx$ is true", as Dummett considers?

The answer to these questions is of course "no". Since, for example, if we want to *accurately* formalize Putnam's saying "(3) If T is consistent, [G] is true", from the quotation above, what we obtain is just the implication $Con(T) \supset Tr(\ulcorner G \urcorner)$, containing the semantical predicate $Tr(x)$. And, as is

[13] Cf. H. Putnam [1], 366.

[14] Cf. M. Dummett [1], 192.

well-known, by Tarski's theorem such an insertion in a theory of its own truth predicate does make it inconsistent. So, Putnam's assertion that "T can perfectly well prove (3) too!" is pure and simple false. The same is the case with Dummett's assertion "[...] *if* the system is consistent, then $\forall x A(x)$ is true".

We do not intend neither to review the above mentioned argument, not to review their criticisms. Our aim is rather to sharpen the dividing line between the licit and illicit uses of true, by considering what can and what cannot be proved in PA and the ways the truth of G can be argued.

2. What can and what cannot be proved in PA

Let us begin by note the correspondence between PA and the modal system GL, known as provability logic. This connection is given by the following equivalence[15]

$GL \vdash \alpha$ if and only if, for every realization *, $PA \vdash \alpha^*$ (Eq)

That is, every theorem of GL is, interpreted, a theorem of PA and conversely. As is well-known, by arithmetization our metamathematical assertions can be converted in arithmetical functions and relations, representable (under some assumptions) in PA. The assertion "S is provable in PA" is the sentence $Bew(\ulcorner S \urcorner)$, "S is not provable in PA" is the sentence $\neg Bew(\ulcorner S \urcorner)$, "PA is consistent" is the sentence $\neg Bew(\ulcorner \bot \urcorner)$, where $\bot$ is the logical falsity, and "S is undecidable in PA" is the sentence $\neg Bew(\ulcorner S \urcorner) \wedge \neg Bew(\ulcorner \neg S \urcorner)$. Evidently, an assertion is provable in PA if its arithmetization is provable in PA.

[15] Cf. R. Solovay [1], G. Boolos [1], Ch. 3 and Ch. 9.

A remarkable result concerning the provability in PA is the following theorem.[16]

Löb's theorem. *If* $\mathrm{PA}\vdash \mathrm{Bew}(\ulcorner S\urcorner)\supset S$, *then* $\mathrm{PA}\vdash S$, *whose formalization in* PA *is the following expression*:

$\mathrm{PA}\vdash \mathrm{Bew}(\ulcorner(\mathrm{Bew}(\ulcorner S\urcorner)\supset S\urcorner)\supset \mathrm{Bew}(\ulcorner S\urcorner)$.

By (Eq), $\mathrm{GL}\vdash \Box(\Box p\supset p)\supset\Box p$.

An expression of the form $\mathrm{Bew}(\ulcorner S\urcorner)\supset S$ is called the *reflection principle for* S. The converse of Löb's theorem is also provable in PA, hence

$\mathrm{PA}\vdash \mathrm{Bew}(\ulcorner S\urcorner)\supset S$ if and only if $\mathrm{PA}\vdash S$,

and this equivalence characterizes the provable cases of reflections, i.e. a reflection for a sentence S is provable if and only if S is provable.

As is well-known, Th G1a) can be formalized in PA, and the corresponding formula *is* provable in PA,[17] i.e.

$$\mathrm{PA}\vdash \mathrm{Con(PA)}\supset G, \tag{1}$$

where Con(PA) is a formula of PA expressing the consistency of PA. It can be $\neg\mathrm{Bew}(\ulcorner\bot\urcorner)$, respectively $\forall x\neg\pi(x,\ulcorner \boldsymbol{0{=}1}\urcorner)$. And the converse of the implication in (1) is also provable, since if there is a sentence not provable in PA, then PA is consistent, i.e.

$$\mathrm{PA}\vdash G\supset \mathrm{Con(PA)} \tag{2}$$

whence

$$\mathrm{PA}\vdash \mathrm{Con(PA)}\equiv G \tag{3}$$

By DL, G is the fixed point of the formula $\forall x\neg\pi(x,y)$ or $\neg\mathrm{Bew}(y)$, respectively, so

[16] Cf. M.H. Löb [1].

[17] Gödel gave only some hints of how this proof can be carried out, the full proof as such was given latter by Bernays, in D. Hilbert, P. Bernays [1], 283-340. The implication in (1) is just the formalization of Gödel's second incompleteness theorem Th G2.

$$\mathrm{PA}\vdash G\equiv\neg\mathrm{Bew}(\ulcorner G\urcorner) \quad (4)$$

and therefore by (3) and (4)

$$\mathrm{PA}\vdash \mathrm{Con(PA)}\equiv\neg\mathrm{Bew}(\ulcorner G\urcorner) \quad (5)$$

Based on the fact that $\mathrm{GL}\vdash\Box(p\equiv\neg\Box p)\supset\Box(p\equiv\neg\Box\bot)$, its arithmetical counterpart is provable in PA, i.e. $\mathrm{PA}\vdash\mathrm{Bew}(\ulcorner G\equiv\neg\mathrm{Bew}(\ulcorner G\urcorner)\urcorner)\supset \mathrm{Bew}(\ulcorner G\equiv\neg\mathrm{Bew}(\ulcorner\bot\urcorner)\urcorner)$. It follows, via (4), that

$$\mathrm{PA}\vdash G\equiv\neg\mathrm{Bew}(\ulcorner\bot\urcorner) \quad (6)$$

The conditional b) of Th G1 can also be formalized in PA. Let us suppose that 1-Con(PA) is a formula of the language of PA expressing "PA is 1-consistent". Then b) is the formula $\text{1-Con(PA)}\supset\mathrm{Bew}(\ulcorner\neg G\urcorner)$ or, equivalently, by (4), $\text{1-Con(PA)}\supset\neg\mathrm{Bew}(\ulcorner\mathrm{Bew}(\ulcorner G\urcorner)\urcorner)$. And, moreover, this implication *is* a theorem of PA, i.e.

$$\mathrm{PA}\vdash \text{1-Con(PA)}\supset\neg\mathrm{Bew}(\ulcorner\mathrm{Bew}(\ulcorner G\urcorner)\urcorner). \quad (7)$$

An informal argument for the truth of this implication is this. By Th G1a) G is not a theorem of PA. That is, $Pf(k,g)$, where g is the Gödel number of G, is false for any k. Hence $\mathrm{PA}\vdash\neg\pi(\boldsymbol{k},\boldsymbol{g})$ for any k, by representability.[18] Now, if PA is 1-consistent, then $\mathrm{PA}\nvdash\exists x\pi(x,\boldsymbol{g})$, equivalently $\mathrm{PA}\nvdash\mathrm{Bew}(\ulcorner G\urcorner)$. And therefore, by (4) and (6), if PA is 1-consistent, then $\mathrm{PA}\nvdash\mathrm{Bew}(\ulcorner\bot\urcorner)$ and, consequently, $\mathrm{PA}\nvdash\mathrm{Bew}(\ulcorner\mathrm{Bew}(\ulcorner\bot\urcorner)\urcorner)$, $\mathrm{PA}\nvdash\mathrm{Bew}(\ulcorner\mathrm{Bew}(\ulcorner\mathrm{Bew}(\ulcorner\bot\urcorner)\urcorner)\urcorner)$ and so on.

Therefore, by (7) the second part of first incompleteness theorem, Th G1b), is also formalizable in PA and provable in PA.

Finally, let us note an important result concerning the undecidability of G (and therefore of Con(PA)) in PA.

[18] Or by Σ_1-completeness of PA.

The modal system GL proves the following implication: $\neg\Box\Box\bot\supset(\neg\Box\neg\Box\bot\wedge\neg\Box\neg\neg\Box\bot)$. Hence its arithmetical counterpart is provable in PA, that is

$$PA\vdash(\neg\Box\Box\bot\supset(\neg\Box\neg\Box\bot\wedge\neg\Box\neg\neg\Box\bot)^{*}. \tag{8}$$

This means that PA proves the following metamathematical assertion (formalized): if the inconsistency of PA is not provable in PA, then G (and hence Con(PA) is *undecidable* in PA.

Therefore, (1)-(8) are some of the most remarkable assertions *provable* in PA. Let us see, concisely, what cannot be proved in PA.

Of course, if PA is 1-consistent, then PA is consistent, since if $\alpha(x)$ is a formula such that $PA\vdash\neg\alpha(\boldsymbol{n})$ for any n, then by 1-consistency the formula $\exists x\alpha(x)$ is not provable in PA. So PA is consistent, otherwise any formula would be provable in PA. The converse of this conditional does not hold, so PA does not prove it, i.e.

$$PA\nvdash Con(PA)\supset 1\text{-}Con(PA) \tag{1*}$$

The proof is simple, for by (4), (6) and (7), if it were provable, then in PA would be provable $Con(PA)\supset\neg Bew(\ulcorner Bew(\ulcorner\bot\urcorner)\urcorner)$, equivalent to $PA\vdash\neg Bew(\ulcorner\bot\urcorner)\supset\neg Bew(\ulcorner Bew(\ulcorner\bot\urcorner)\urcorner)$, by (3) and (6), and to $PA\vdash Bew(\ulcorner Bew(\ulcorner\bot\urcorner)\urcorner)\supset Bew(\ulcorner\bot\urcorner)$, respectively. From where, by Löb's theorem $PA\vdash Bew(\ulcorner\bot\urcorner)$, i.e. PA would be 1-inconsistent.

Nor the converse of the implication in (7) is provable in PA, i.e.

$$PA\nvdash\neg Bew(\ulcorner Bew(\ulcorner G\urcorner)\urcorner)\supset 1\text{-}Con(PA) \tag{2*.1}$$

respectively,

$$PA\nvdash\neg Bew(\ulcorner Bew(\ulcorner\bot\urcorner)\urcorner)\supset 1\text{-}Con(PA) \tag{2*.2}$$

The proof can be given by *reductio*, based on the fact that $\text{PA}\vdash 1\text{-Con}\supset\neg\text{Bew}(\ulcorner\text{Bew}(\ulcorner\text{Bew}(\ulcorner\bot\urcorner)\urcorner)\urcorner)$ and one application of Löb's Theorem

$$\text{PA}\nvdash\neg\text{Bew}(\ulcorner\bot\urcorner)\supset\neg\text{Bew}(\ulcorner\text{Bew}(\ulcorner G\urcorner)\urcorner) \tag{3*.1}$$

$$\text{PA}\nvdash\neg\text{Bew}(\ulcorner\bot\urcorner)\supset\neg\text{Bew}(\ulcorner\text{Bew}(\ulcorner\bot\urcorner)\urcorner) \tag{3*.2}$$

We note a remarkable difference between (1*) and (2*), on the one hand, and (3*), on the other hand. Both implications in (1*) and (2*) are *metamathematically false* statements of PA. The formulas in (3*), though not provable in PA, are *metamathematically true* statements of PA. I.e. by an *informal* argument, (3*) can be proved as being true. The proof is simple, let us show this for (3*.1).

Proof (reductio). The conditional formalized in (3*.1) is therefore: if PA is *consistent* then the sentence G is not refutable in PA. Let us suppose that $\text{PA}\vdash\neg G$, then $\text{PA}\vdash\text{Bew}(\ulcorner\neg G\urcorner)$, by derivability conditions for Bew(y). But we also have $\text{PA}\vdash\text{Bew}(\ulcorner G\urcorner)$, by (4) above. So, if $\text{PA}\vdash\neg G$, then $\text{PA}\vdash\text{Bew}(\ulcorner G\urcorner)\wedge\text{Bew}(\ulcorner\neg G\urcorner)$. But $\text{GL}\vdash(\Box p\wedge\Box\neg p)\supset\Box\bot$. So $((\Box p\wedge\Box\neg p)\supset\Box\bot)^*$ is a theorem of PA. Hence, for $p^*=G$ does follow $\text{PA}\vdash\text{Bew}(\ulcorner\bot\urcorner)$, i.e. PA proves its own inconsistency. So, if PA is *consistent*, then $\neg G$ is not provable in PA.[19]

But by Th G1a), if PA is *consistent*, then the sentence G is also not provable in PA. So, informally, the simple assumption of *consistency* is sufficient for the undecidability of G.

By (3*.1) the nonprovability of $\neg G$ does *not* follow *formally* from the simple consistency. Hence, *formally* the undecidability of G does not follow from the simple consistency of PA, or from a single reflection (since

[19] In S. McColl's view the implication in (3*.1) is a notable example of a truth not knowable by a Turing machine; cf. S. McCall [1].

$\neg$Bew(⌜⊥⌝) is just the reflection Bew(⌜⊥⌝)$\supset\perp$). So the undecidability of G in PA needs a stronger hypothesis, as (8) above shows. This hypothesis, $(\neg\Box\Box\perp)^*$, equivalent to $(\Box\Box\perp\supset\perp)^*$ is just the conjunction of two reflections: $(\Box\Box\perp\supset\Box\perp)^*$ and $(\Box\perp\supset\perp)^*$, the last expression being just the expression $\neg$Bew(⌜⊥⌝), i.e. Con(PA).

Let us (temporarily) summarize. What really shows the *metamathematical* proof of Gödel's theorem? By Th G1a), if PA is consistent, then G is not provable in PA, and by Th G1b), if PA is 1-consistent, then $\neg$G is not provable in PA. Both conditionals are formalizable in the language of PA, and *mathematically* provable in PA (cf. (1) and (7)). By (3), (5) and (6) the following equivalences are also provable in PA: Con(PA)≡G≡$\neg$Bew(⌜G⌝)≡$\neg$Bew(⌜⊥⌝). And the implication in (1) is just the formalization of Gödel's second incompleteness theorem, Th G2.

So, formalizable or not, Th G1 proves *conditional* assertions, of the form "If... then___", and *nothing more*.

But, as we noted above, both conditionals of Th G1 can be metamathematically provable only under the assumption of *simple* consistency. However, in this case the second part, b), of Th G1, formalized, though a true sentence of PA, is no more provable in PA (by (3*)).

Hence, by (3*) and (8), the undecidability of G (and Con(PA)) in PA does not *formally* follow only from the assumption of simple consistency.

If, after all, what is proved, formally or not, by Th G1 is a conditional assertion, *without* any reference to the notion of *truth* of the undecidable sentence G, then which are the reasons of sayings of the form: *by Gödel's theorem we know that* G *is true*? Of course, since the language L_{PA} has no semantical terminology, talking about G's truth is carried out outside PA.

As often as not the *truth* of the sentence G is associated to Gödel's theorem Th G1 a). But, how do we conclude that the sentence G is *true*? By saying "G is true" we essentially take into account anyone of the following two things:

a) G is *metamathematically* (informally) provable as being *true*.

b) G is *formally* provable as being *true*.

Let us question both ways.

3. How do we know that G is true?

The most explanatory form of the semantical argument for the truth of G is given in Dummett's paper on this topic.[20] But some arguments, in a more poor fashion, have been given previously.[21] On Lucas view, for example, the truth of G follows directly from Gödel's theorem Th G1a), for G is a self-referential sentence, asserting "This formula is unprovable-in-the-system", and therefore

> if the formula "This formula is unprovable-in-the-system" is unprovable-in-the system, then it is true that the formula is unprovable-in-the-system, that is, "This formula is unprovable-in-the-system" is true.[22]

This line of argumentation, based on Gödel's theorem, but more analytical, is given later by Mendelson.[23] On his view, G is a sentence of the form $\forall x \neg \pi(x,\boldsymbol{g})$, where g is the Gödel number of G and $\pi(x,y)$ is the formula

---

[20] Cf. M. Dummett [1].
[21] E.g. E. Nagel and J. Newman [1], and J.R. Lucas [1].
[22] Cf. J.R. Lucas [1], 121.
[23] Cf. E. Mendelson [1], 144.

expressing the intuitive relation $Pf(x,y)$: "x is a proof of y". So G states that the relation $Pf(x,g)$ is false for any numerical value of x, that is, there is no proof in PA of the sentence G. But G states exactly this thing. So, by Gödel's theorem, if PA is consistent, then G is *not* provable in PA. So G is unprovable in S, but a *true* sentence of S.

Both these examples base their argument on Gödel's result.

Let us consider Dummett's semantical argument.

On his view, the argument for the truth of G is given in the following terms:

> The statement [G] is of the form $\forall x A(x)$, where each one of the statements A(0), A(1), A(2),... is true: since A(x) is recursive, the notion of truth for these statements is unproblematic. Since each of the statements A(0), A(1), A(2),... is true in every model of the formal system, any model of the system in which [G] is false must be a non-standard model. [...] whenever, for some predicate B(x), we can recognize all of the statements B(0), B(1), B(2),... as true in the standard model; then we can recognize that $\forall x B(x)$ is true in that model.[24]

Furthermore,

> The argument for the truth of [G] proceeds under the hypothesis that the formal system in question is consistent. The system is assumed, further, to be such that, for any decidable predicate B(x) and any numeral ***n***, B(***n***) is provable if it is true, –B(***n***) is

[24] Cf. M. Dummett [1], 191; the argument given in Mendelson [1], 144, is similar.

provable if B($\boldsymbol{n}$) is false (the notions of truth and falsity, for such statements being, of course, unproblematic). The particular predicate A(x) is such that, if A($\boldsymbol{n}$) is false for some numeral $\boldsymbol{n}$, then we can construct a proof in the system of $\forall$xA(x). From this it follows – on the hypothesis that the system is consistent – that each of A(0), A(1), A(2),... is true.[25]

Let us briefly analyse Dummett's argument for the truth of G. As we noted above, G is the sentence $\forall x \neg\pi(x,\boldsymbol{g})$, where $\pi(x,y)$ is the formula expressing in PA the primitive recursive relation *Pf*(*x,y*): "*x* is a proof of *y*". It is a primitive recursive formula and so is the formula, $\neg\pi(x,\boldsymbol{g})$. Being decidable, for every value of x, the corresponding instances are provable in PA, i.e. $PA \vdash \neg\pi(\boldsymbol{0},\boldsymbol{g})$, $PA \vdash \neg\pi(\boldsymbol{1},\boldsymbol{g})$, $PA \vdash \neg\pi(\boldsymbol{2},\boldsymbol{g})$,... On Dummett's view the truth of G follows by these two steps:

1. Being provable, all sentences $\neg\pi(\boldsymbol{0},\boldsymbol{g})$, $\neg\pi(\boldsymbol{1},\boldsymbol{g})$, $\neg\pi(\boldsymbol{2},\boldsymbol{g})$,... are *true* (in the standard model M).

2. All these sentences being true (in M), the universally quantified sentence $\forall x \neg\pi(x,\boldsymbol{g})$ is also *true* (in M).

The second step is "trivial" (p. 192) indeed, for it follows by semantics of the universally quantified sentences, i.e. $\forall$xA(x) is true in M if and only if A(x) is true in any assignment μ in M, if and only if A($\boldsymbol{n}$) is true for any natural number *n*.

The key step is, of course, 1, for it expresses the idea that *if a formula is provable in* PA*, then it is true* (*in* M). It expresses the *soundness*

[25] Cf. M. Dummett [1], 192.

of PA, or Δ-reflection principle for PA (Δ-Refl), and cannot be unproblematically inserted *in* PA.[26]

Of course, Dummett is right in asserting that under the assumption of consistency of PA, all the instances of $\neg\pi(x,\boldsymbol{g})$ are true, for if one of them would be false, then it can be argued that G is provable in PA. Indeed, let us suppose that for a natural number k, the sentence $\neg\pi(\boldsymbol{k},\boldsymbol{g})$ is false. Then $\pi(\boldsymbol{k},\boldsymbol{g})$ is true, so it is provable *in* PA by Σ_1-completeness of PA. Hence $Pf(k,g)$ is true, by representability, that is the sentence with Gödel number g, i.e. G itself, is provable in PA, i.e. PA$\vdash$G, contrary to what Gödel's theorem Th G1 a) asserts.

So, to sum up: *metamathematically*, the argument for the truth of G depends essentially on something not co-optable in PA, i.e. on the idea: *every* (*closed*) *formula, provable in* PA*, is true in* M. Let us detail this idea below.

4. Conservative and nonconservative extensions of PA

The truth of the Gödelian sentence G can therefore be *metamathematically* argued, in the ways noted above. But what about a *proof* of the fact that G is true? Of course, two mentions have to be made *ab initio*. Firstly, by Th G1a), such a proof cannot be carried out *within* PA, and secondly, it depends on what is meant by a proof of the "truth of G". If what is to be proved is Tr(⌜G⌝), where "Tr(x)" is the truth predicate for a specified language, then a semantical apparatus must be added to the given theory. But if what is required is only a derivation of G, then the things are different. Let us see!

[26] Cf. sect. 5 below.

Let L_{PA} be any standard formalized first-order language of Peano Arithmetic. Let $L^*_{PA} = L_{PA} \cup \{Tr(x), \ulcorner ... \urcorner\}$ be an extension of L_{PA} (i.e. $L_{PA} \subset L^*_{PA}$, obtained from L_{PA} by adding a truth predicate for the initial language L_{PA} and a quotation operator. Now, if we consider all equivalences of the form $Tr(\ulcorner\alpha\urcorner)\equiv\alpha$, for any closed α, belonging to L_{PA}, as axioms of a theory *Defl*,[27] then it is just a *truth* theory for the language L_{PA}.

An important difference between various extensions of a theory concerns the nature of such an extension, i.e. whether the respective extension is *conservative* or not. Essentially, the theory T^* is a conservative extension of T if the following holds: *if* α *is a sentence of* L_T *and* $T^* \vdash \alpha$, *then* $T \vdash \alpha$, i.e. any sentence of the language of T, provable in T^* is also provable in T.

Let now $PA^*=PA\cup$*Defl* be an extension of PA. Is it sufficient to prove that G is true by proving $Tr(\ulcorner G\urcorner)\equiv G$? Not at all! For, as is well-known, such an extension of PA is conservative over PA.[28] So for a proof of G's truth a stronger theory is needed. Which one? Tarski saves us again.[29] Let us sketch Tarkian theory of truth (via satisfaction).

Let *Sat* be a theory of satisfaction, formulated in a language $L^*_{PA} = L_{PA} \cup Sat(x, y)$, an extension of the language of PA.[30] The theory *Sat* contains the axioms for the semantical predicate Sat(x,y), and allows the

---

[27] *Deflationism*, a theory of truth in Tarski's sense, i.e. it satisfies Tarski's truth Schema T (or Convention T): $Tr(\ulcorner\alpha\urcorner)\equiv\alpha$, for any sentence α.
[28] An interesting proof of its conservativeness, is given in J. Ketland [1], Theorem 1.
[29] Cf. A. Tarski [1].
[30] This metalanguage is much more powerful than the metalanguage of *Defl*, since it contains a variety of syntactical predicates, operators, the satisfaction predicate Sat(x,y) and many other entitities.

construction of the so-called inductive definition of truth and an explicit definition of truth (given in L^{*}_{PA}, in terms of the predicate Sat(x,y)).[31]

Let $PA^{*}=PA\cup Sat$ be an extension of PA. This theory proves Tarski's schema T (it is therefore an extension of *Defl*), but proves much more. As Tarski's showed,[32] in PA^{*} we can prove that the axioms of PA are true and its rules of deduction preserve truth, hence PA^{*} proves that *any formula of* PA, *provable in* PA *is true*. But this is just a reflection for PA: $\forall x(Bew(x)\supset Tr(x))$. Therefore, a remarkable fact about PA^{*}:

$PA^{*}\vdash Refl(PA){:}\forall x(Bew(x)\supset Tr(x))$

It is clear that such an extension of PA is *not* generally conservative over PA, i.e. PA^{*} proves some sentences of L_{PA} not provable in PA. A simple example shows this. Let us take for x in Refl(PA) the "value" ***0=1***. Then $PA^{*}\vdash Bew(\ulcorner \boldsymbol{0=1}\urcorner) \supset Tr(\ulcorner \boldsymbol{0=1}\urcorner)$. Hence $PA^{*}\vdash Bew(\ulcorner \boldsymbol{0=1}\urcorner)\supset(\boldsymbol{0=1})$ (by Schema T). But $PA^{*}\vdash\neg(\ulcorner \boldsymbol{0=1}\urcorner)$. Hence $PA^{*}\vdash\neg Bew(\ulcorner \boldsymbol{0=1}\urcorner)$, i.e. $PA^{*}\vdash Con(PA)$. But Con(PA) is *not* a theorem of PA, by Th G2.

In a similar fashion, a proof of Gödel's sentence G of PA in PA^{*} is an easy task. By (4), section 2, $PA^{*}\vdash G\equiv\neg Bew(\ulcorner G\urcorner)$, and by Refl(PA), $PA^{*}\vdash Bew(\ulcorner G\urcorner)\supset Tr(\ulcorner G\urcorner)$ and therefore $PA^{*}\vdash Bew(\ulcorner G\urcorner)\supset G$, by Schema T. Hence $PA^{*}\vdash Bew(\ulcorner G\urcorner) \supset\neg Bew(\ulcorner G\urcorner)$ and thus, by propositional calculus, $PA^{*}\vdash\neg Bew(\ulcorner G\urcorner)$, i.e. $PA^{*}\vdash G$, and finally $PA^{*}\vdash Tr(\ulcorner G\urcorner)$ by Schema T.

Actually, the provability of G's truth in PA^{*} could be obtained directly from the above proof of Con(PA) in PA^{*}, via (3) of section 2: $PA\vdash Con(PA)\equiv G$ and Schema T.

---

[31] *Sat* is the theory of satisfaction (truth) for the language L_{PA}.
[32] Cf. A. Tarski [1], Theorem 5; comp. and S. Feferman [2], 16, Theorem 2.5.3.

To sum up, the formal proof of "G is true" can be given only in a non-conservative extension PA^* of PA, a theory in which the reflection Refl(PA), i.e. the sentence $\forall x(Bew(x)\supset Tr(x))$, is provable.

In other words, since both sentences, G for PA and Con(PA), are sentences of PA, i.e. all the symbols occurring in them are symbols of L_{PA}, and since G and Con(PA) are not provable in PA, but are provable in PA^*, this means that PA^* does not extend PA conservatively. Or, the proof of "G is true" (essentially based on the provavility in PA^* of Refl((PA)) and the conservativeness of PA^* over PA are *not* compatible.[33]

A question arises: is this incompatibility, due to the use of the truth predicate Tr(x) in the above proof, beyond the deflationary licit use of it in Schema T?

As our proof above suggests, the answer to this question is "no". But a convincing argument for this answer requires some considerations on reflection principles.

5. Reflection principles (more about them)

So, what do we really need in order *to prove* the Gödel sentence G, *without* any *formal* use of the truth predicate Tr(x)? A hint is given by the Dummett's metamathematical argument (in sect. 3). If we succeed in inserting both steps of this argument in a formal system of arithmetic, without using the truth predicate, then G can be derived. As we saw, his argument is essentially this: if 1. all provable instances of the formula $\neg\pi(x,\boldsymbol{g})$ are true, then 2. the universal quantified sentence $\forall x\neg\pi(x,\boldsymbol{g})$ is also

[33] S. Shapiro [1] uses this kind of incompatibility in order to argue against the deflationism as a complete account of truth. Briefly, the argument is: since the "conservativeness is essential to deflationism" (497) and no conservative extension of PA proves that G is true, it follows that deflationism is not an adequate theory of truth.

true. How can they be inserted? As we know, the sentence G is the fixed point of the formula $\forall x \neg \pi(x,y)$, where $\pi(x,y)$ is the formula of PA expressing formally in PA the primitive recursive relation *Pf*(*x,y*): "*x* is a proof of *y*". $\pi(x,y)$ is a Δ-formula and so is $\neg\pi(x,y)$, hence all instances of the formula $\neg\pi(x,\boldsymbol{g})$, where G is the Gödel number of G, are provable in PA, i.e.:

PA$\vdash \neg\pi(\boldsymbol{0},\boldsymbol{g})$, PA$\vdash \neg\pi(\boldsymbol{1},\boldsymbol{g})$, PA$\vdash \neg\pi(\boldsymbol{2},\boldsymbol{g})$,...

In order to derive $\forall x \neg\pi(x,\boldsymbol{g})$, we need a rule of the following form:

$$\text{R-}\omega \quad \frac{\text{PA}\vdash \neg\pi(\boldsymbol{0},\boldsymbol{g}),\ \text{PA}\vdash \neg\pi(\boldsymbol{1},\boldsymbol{g}),\ \text{PA}\vdash \neg\pi(\boldsymbol{2},\boldsymbol{g}),\ldots}{\forall x \neg\pi(x,\boldsymbol{g})}$$

But such a rule, does not mimic, by itself, Dummett's reasoning, for in his argument the truth of conclusion (the sentence G) follows from the truth of the infinite sequence of sentences $\neg\pi(\boldsymbol{0},\boldsymbol{g})$, $\neg\pi(\boldsymbol{1},\boldsymbol{g})$, $\neg\pi(\boldsymbol{2},\boldsymbol{g})$,... So the missing step 1 can be explicitly introduced in one of the following forms: a) by supplementing the above rule with a *reflection principle* of the form: all provable sentences in PA are assertable, or by b) a reformulation of the premise of this rule. In the last case that infinite sequence of sentences have to be replaced by a formula expressing the following fact: PA proves that all instances of $\neg\pi(x,\boldsymbol{g})$ are provable. Indeed, this premise of R-ω can be formally expressed by the formula $\forall x \mathrm{Bew}(\ulcorner\neg\pi(\boldsymbol{x},\boldsymbol{g})\urcorner)$.[34] Moreover, this formula *is provable* in PA, i.e. PA$\vdash \forall x \mathrm{Bew}(\ulcorner\neg\pi(\boldsymbol{x},\boldsymbol{g})\urcorner)$. So, in order to derive the Gödel sentence $\forall x \neg\pi(x,\boldsymbol{g})$ one application of the following rule will do the job:

[34] This is possible, since PA is an axiomatic theory (i.e. the set of its theorems is recursively enumerable), so its proof relation can be expressed, via arithmetization, by a primitive recursive predicate $\pi(x,y)$.

$$\text{R-}\omega\text{ (weak)}[35] \quad \frac{\mathrm{PA}\vdash \forall x\mathrm{Bew}(\ulcorner\neg\pi(\boldsymbol{x},\boldsymbol{g})\urcorner)}{\forall x\neg\pi(x,\boldsymbol{g})}$$

But a remark has to be added. The provability in PA of the formula $\forall x\mathrm{Bew}(\ulcorner\neg\pi(\boldsymbol{x},\boldsymbol{g})\urcorner)$ does *not* entail that this form of ω-rule is a *part* of PA. So, what really means "the provability of G by using this rule" is just the following fact: G is provable in a *proper* extension of PA, i.e. G is provable in PA^*=PA+R-ω (weak).[36]

In a celebrated paper[37] S. Feferman proved that R-ω (weak) is equivalent to the following two uniform reflection principles:[38]

URefl_1: $\forall x\mathrm{Bew}(\ulcorner\alpha(\boldsymbol{x})\urcorner)\supset\forall x\alpha(x)$; α has only x free

URefl_2: $\forall x[\mathrm{Bew}(\ulcorner\alpha(\boldsymbol{x})\urcorner)\supset\alpha(x)]$; α has only x free

In the above reflection principles the only restriction is that the formula α does contain only x free. Otherwise α can have an arbitrary complexity. But in order to derive G or Con(PA) in an nonconservative extension of PA this strengthening of reflections is not necessary. As we saw, in the premise of R-ω (weak), the formula $\neg\pi(x,\boldsymbol{g})$ is a Δ-formula (i.e. decidable). So what is really needed in order to derive $\forall x\neg\pi(x,\boldsymbol{g})$ as a conclusion is a R-ω (weak) or, equivalently, URefl_1 or URefl_2, restricted to formulas of this kind. Indeed such a reflection principle is available.[39] A result of C. Smorynski[40] helps us to establish its validity. By his Theorem

---

35 "R-ω weak", i.e. weaker than the preceding form R-ω, since its premise is stronger than the provability of each of the sentences in the premise of R-ω.
36 Of course, PA^* is a *proper* extension of PA only if PA is consistent.
37 Cf. S. Feferman [1].
38 Cf. S. Feferman [1], Theorem 2.19(i), 276, via Definition 2.16, 274.
39 It is usually called "uniform primitive recursive reflection".
40 Cf. C. Smorynski [1].

4.1.4 (p. 846), if $\alpha(x)$ in all these principles is a π_1-formula, then the following equivalences can be established over PA:

$Con(PA) \equiv URefl_1 \equiv URefl_2$.

But these equivalences are preserved by restricting $\alpha(x)$ to Δ-formulas and, by Feferman's considerations, R-ω (weak) is equivalent to anyone of them. Hence any extension of PA with anyone of the above mentioned reflections is equivalent to an extension of PA with Con(PA). And by Th G1a) this is just what is needed in order to prove the sentence G.

Let us show how a proof of G in $PA^* = PA \cup R\text{-}\omega$ (weak) (or in any equivalent extension of PA) can be carried out.

We show, firstly, that the premise of R-ω (weak) holds. For this task let us suppose that PA is consistent and G is provable in PA, i.e. $PA \vdash G$. This means that there is a *k* the Gödel number of a proof of G in PA. Hence *Pf*(*k,g*) is true. By representability of *Pf* in PA, this implies that $PA \vdash \pi(\boldsymbol{k},\boldsymbol{g})$. But G is the fixed point of the formula $\forall x \neg\pi(x,y)$, so the provability of G in PA entails that $PA \vdash \forall x \neg\pi(x,\boldsymbol{g})$, and hence $PA \vdash \neg\pi(\boldsymbol{k},\boldsymbol{g})$. Therefore, PA proves $\pi(\boldsymbol{k},\boldsymbol{g})$ and $\neg\pi(\boldsymbol{k},\boldsymbol{g})$, contradicting the assumed consistency of PA.

So there is no *k* the Gödel number of a proof in PA of the sentence G. This means that for any natural number *k*, *Pf*(*k,g*) is false. So, by representability, for any *k*, $PA \vdash \neg\pi(\boldsymbol{k},\boldsymbol{g})$. And, by the above considerations, $PA \vdash \forall x Bew(\ulcorner \neg Bew(\boldsymbol{x},\boldsymbol{g}) \urcorner)$. By one application of R-ω (weak) or, equivalently, of $URefl_1$, the Gödel sentence $\forall x \neg\pi(x,\boldsymbol{g})$ follows.

6. Conclusions

Gödel's famous discovery of the incompleteness phenomenon is still the base of some philosophical controversies. Probably the disputes around the thesis "meaning is use" and the algorithmic nature of arithmetical thinking are the most striking. Both the supporters of the idea that our arithmetical understanding outruns any attempt to formalize it and the supporters of the idea that our arithmetical understanding is not entirely algorithmic base their considerations on Gödel's first incompleteness theorem. As we saw in section 1, the logical mechanism connecting both views is given by the correspondence between the standard form of Gödel's theorem and its Turing version. By its standard form, the Gödelian sentence G is not provable in PA (assuming the consistency of PA) but it is metamathematically argued as being true, and by its Turing form no sound procedure for ascertaining the nonhalting a computation is complete. In both cases there are true sentences transcending their algorithmic justification.

But in some criticisms of these views, focused on the difference between truth and provability, some careless use of "true" tends to blur just the distinction they want to make explicit. After all, the key point in these arguments and in their criticisms is just the capital distinction between what can and cannot be proved in PA.

Of course, by Tarski's theorem the formal system of PA cannot contain a formula Tr(x) expressing the truth predicate for PA. So although many metamathematical assertions contain the word "true", their *formalization* cannot be inserted *within* PA, even though their usual reading re-establishes the use of "true" (as is the case in the two quotations above). Then what is really proved within PA is just an implication of the form Con(PA)$\supset$G and nothing more. Neither the conditional "if PA is consistent, then G is true" nor the fact that G is true is provable in PA. Moreover, the fact that a metamathematical assertion is formalizable in the language of PA does not imply that it is also provable within PA. As we showed in section 2, the assertion "if PA is consistent, then G is undecidable in PA" is true, formalizable in the language of PA but not provable within PA. The same is the case with the reflection principles, expressed in the language of PA, i.e. without using the word "true".

The sections 3 and 4 analyse two licit ways of showing that Gödelian sentence G is *true*. The first is just the Dummett's semantical argument, i.e. a *metamathematical* argument that G is *true* in the standard model of PA. The second is a proof of G in an nonconservative extension of PA. This can be given either by using reflection principles containing the truth predicate for the language of PA (PA with a Tarskian semantic can prove such a reflection), or by using reflection principles without the truth predicate. This second route allows a formal proof of G in an extension of PA, that mimics Dummett's (or Mendelson's) semantical argument but without using "true". Given the fact that the open formula $\neg\pi(x,\boldsymbol{g})$, part of the Gödelian sentence $\forall x\neg\pi(x,\boldsymbol{g})$, is a Δ-formula, the minimal extension of PA, required by the proof of G, will be the extension of PA with primitive recursive R-ω (weak), or equivalently, with Feferman's uniform reflection principles restricted to Δ-formulas.

References

Boolos, G., [1] *The Logic of Provability*, Cambridge University Press, 1993.

Dummett, M., [1] "The philosophical significance of Gödel's theorem (1963)", in *Truth and Other Enigmas*, London, Duckworth, 1978, 186-201.

Feferman, S., [1] "Transfinite recursive progressions of axiomatic theories", *Journal of Symbolic Logic*, 27, 1962, 259-316.

[2] "Reflecting on incomplenetess", *Journal of Symbolic Logic,* 56, 1991, 1-49.

Gödel, K., [1] "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatsh. Math. Phys.*, 38, 1931, 173-198.

Hilbert, D., P. Bernays, [1] *Grundlagen der Mathematik*, vol. II, Berlin: Springer, 1939.

Ketland, J., [1] "Deflationism and Tarski's Paradise", *Mind*, 108, 1999, 69-94.

Kleene, S. C., [1] *Introduction to Matemathematics*, fourth edition, North-Holland Publishing Co., Amsterdam, 1964.

Löb, M. H., [1]" Solution of a problem of Leon Henkin", *Journal of Symbolic Logic*, 20, 1955, 115-118.

Lucas, J. R., [1] "Minds, Machines and Gödel", *Philosophy*, 36, 1961, 120-124.

McCall, S., [1]"Can a Turing machine know that the Gödel sentence is true?", *The Journal of Philosophy*, 96, 1999, 525-532.

Mendelson, E., [1] *Introduction to Mathematical Logic*, D. Van Nostrand Company, Inc. 1964.

Nagel, E., J. Newman, [1] *Gödel's Proof*, New York: New York University Press, 1958.

Penrose, R., [1] *Shadows of the Mind*, Oxford University Press, 1994.

Putnam, H., [1]"Minds and machines", in *Philosophical Papers* 2, Cambridge University Press, 1975, 362-385.

Shapiro, S., [1] "Proof and Truth: Through Thick and Thin", *The Journal of Philosophy*, XCV, 1998, 493-521.

Smorynski, C., [1] "The incompleteness theorems", in *Handbook of Mathematical Logic*, J. Barwise ed., North-Holland Publishing Company, Amsterdam, 1977, 821-865.

Solovay, R., [1] "Provability interpretations of modal logic", *Israel Journal of Mathematics*, 25, 1976, 287-304.

Tarski, A., [1] "The concept of truth in formalized languages", in *Logic, Semantics, Metamathematics: Papers by Alfred Tarski from* 1922-1938, Oxford: Clarendon Press, 1956, 152-278. Originally published as "Der Wahrheitsbegriff in den formalisierten Sprachen", *Studia Philosophia* I, 1936.

Turing, A., [1] "Systems of logic based on ordinals", *Proc. London Math. Soc.* (2) 45, 1939, 161-228.