On a certain fallacy concerning I-am-unprovable sentences.

KAAVE LAJEVARDI (Kaave.Lajevardi@Gmail.com)
La Société des Philosophes Chômeurs, P.O. Box 13197-73587, Téhéran, IRAN.

SAEED SALEHI (root@SaeedSalehi.ir)
Department of Mathematical Sciences, University of Tabriz, P.O. Box 51666-16471, Tabriz, IRAN.

12 February 2023

Abstract

We demonstrate that, in itself and in the absence of extra premises, the following argument scheme is fallacious: *The sentence A says about itself that it has a property F, and A does in fact have the property F; therefore A is true.* We then examine an argument of this form in the informal introduction of Gödel's classic (1931) and examine some auxiliary premises which might have been at work in that context. Philosophically significant as it may be, that particular informal argument plays no rôle in Gödel's technical results.

Going deeper into the issue and investigating truth conditions of Gödelian sentences (i.e., those sentences which are provably equivalent to their own unprovability) will provide us with insights regarding the philosophical debate on the truth of Gödelian sentences of systems—a debate which goes back to Dummett (1963).

1. Introduction.

Consider the following argument scheme, wherein boldface italic letters are placeholders for sentences and predicates of a sufficiently rich language:

(1) ***A*** says about itself that it is an ***F***

(2) ***A*** is indeed an ***F***

therefore

(3) ***A*** is true.

Harmless and perhaps trifling as the scheme (1)-(2)-(3) may look, we will argue that it is in fact invalid: there are instances of ***A*** and ***F*** which make premises (1) and (2) true while the conclusion (3) false. We will demonstrate the invalidity of the scheme and draw some morals hopefully attractive to the philosophical community.[1]

What is likely to make this fallacy of more interest is that it seems to have been committed by a logician no less than Gödel, and then somewhere in the introduction to his ground-breaking 1931 paper—but let us add immediately that we are perfectly aware that the fallacy is of no consequence to Gödel's technical results. That this *is* a fallacy is demonstrated in Section 3 below; whether or not *Gödel* has committed it is somehow a matter of judgement and depends on what one is willing to read into Gödel's text (see 4.3 below). At any rate, we think awareness of the fallacy provides us with new insights concerning arguments about the truth of Gödelian sentences of axiomatic systems.

[1] So far as we know, the fallacy was first exposed by Lajevardi & Salehi (2019). That work does not offer any diagnosis of the fallacy and minimizes the exegesis of Gödel (1931). While we sharpen some of the results of Lajevardi & Salehi (2019) and apply them to the philosophical debate on the truth of Gödelian sentences, we do not presume any knowledge of that paper.

In what follows, we first (§2) give a rough exposition of why the (1)-(2)-(3) scheme fails to be valid, followed by technical details in §3. Section 4 is an exegetical discussion of Gödel's 1931 paper. Among other things, we consider a number of points that could be made in defence of Gödel's use of a particular instance of the (1)-(2)-(3) scheme. In §5, we exploit our results to say something which we believe systematize a number of previous attempts by philosophers and logicians concerning the truth of self-referential sentences which assert their own unprovability. The Appendix presents more abstract versions of some of the issues presented in the paper.

Let us make it explicit that here we are *not* dealing with self-referential sentences as such (see 4.4 below). Rather, our subject matter is, first and foremost, those sentences which are provably equivalent in a suitable ambient theory to their own unprovability relative to that same theory—hence the 'I-am-unprovable' of our title.

2. Invalidation: an overview.

We should, of course, answer two questions prior to embarking on the invalidation. Our answers in the present section are not meant to be rigorous or complete—we only wish to give a sense of what is going on.

(α) What is it for a sentence to say something about itself?

(β) What is it for a sentence to be true? In particular: What is it for a sentence to have a property *indeed*?

In the context of modern mathematical logic, here are the answers: (α) a sentence says something about itself just in case a certain biconditional, containing the sentence and a predicate applied to a name of that sentence, is a theorem of the background theory of the context.[2] As for (β), a sentence is true just in case it holds (i.e., satisfies Tarski's standard recursive definition) in a fixed designated structure. These will become clearer in the paragraph below. And how could (1) and (2) fail to imply (3)? The how-is-it-possible question is relatively easy to answer. Here is an overview:

What a sentence ψ *says* needs an interpretative or translational work to tell, and all interpretation is theory-relative.[3] Any such work takes place in a theory T in a given language, a language which either contains ψ itself or has a means of referring to it or mentioning it, e.g., via having a name or a code for ψ. Now if, according to T, the sentence ψ is equivalent to $\Delta(\#\psi)$, where $\#\psi$ is a name for ψ in the language of T and Δ is a predicate of the same language, we say that ψ says about itself that it is a Δ. Thus, looking at Gödel's construction in his classic 1931 paper, we have a sentence G such that PM (the system of *Principia Mathematica* of Russell and Whitehead) proves the biconditional $G \leftrightarrow \neg\mathrm{Pr}(\#G)$, where Pr is a standard provability predicate for PM. It is in virtue of this that one thinks of G as *saying about itself that it is unprovable*. If the context makes it clear what the background theory is, or if the discussion is carried out rather informally and the theory is more or less the set of all we take for granted in that discussion, the theory is left unspecified.

Suppose that our background theory T is not "sound". That is to say, suppose that some of the theorems of T are not true, where *true* is relative to a given structure as said in (β) above.[4] It may so happen that amongst the false theorems of T, one is of the form $\psi \leftrightarrow \Delta(\#\psi)$. If so,

[2] See subsections 4.1 and 4.4 below.

[3] We use the term 'theory' in the standard logical sense: a *theory* is a set (usually a recursively enumerable one) of formulæ none of which contain any free variables.

[4] Normally, when one is talking about the sentences of the language of arithmetic, 'true' simpliciter means true in the standard model ℕ of natural numbers—see e.g., Gödel (1931:145*n*4) quoted below in 4.2.

then, according to *T,* the sentence ψ says about itself that it is a Δ. Our unsound theory is just in error about what ψ "really says".

Now if we somehow manage to make all these happen for a false ψ which is *in fact* a Δ (i.e., in such a way that Δ(#ψ) is true in ℕ), we are done with showing the invalidity of (1)-(2)-(3): ψ says about itself that it is a Δ (in the eye of *T*), and ψ is indeed a Δ; however, ψ is false. Hence the invalidity.

3. Invalidation: the details.

We show how to find a triplet of a theory *T*, a sentence ψ, and a predicate Δ as a counterexample to the (1)-(2)-(3) scheme.

Step 1. Let *T* be a consistent but unsound theory in the language of arithmetic.[5] For reasons which need not concern us here, it is a normal practice today to assume that *T* is an extension of Robinson's theory ***Q***.[6] As *T* is unsound, there is a false sentence φ which is provable in *T*; i.e., there is a φ with $T \vdash \varphi$ and $\mathbb{N} \nvDash \varphi$. Let Δ(*x*) be any formula whatsoever with exactly one variable *x*. Apply the celebrated diagonal lemma (Boolos and Jeffrey (1989:173)) to the formula $\varphi \leftrightarrow \Delta(x)$, to get a sentence ψ such that[7]

[5] A well known example of such theory is Peano Arithmetic to which a statement of the inconsistency of PA is added—that is to say, PA + not-Con(PA) is consistent (by Gödel's second incompleteness theorem) but not sound *if* PA is consistent to begin with.

For an inconsistent *T,* it is all too easy to invalidate the (1)-(2)-(3) argument scheme. In 4.3 we will invalidate (1)-(2)-(3) even for ω-consistent theories.

[6] ***Q*** is basically a first-order description of the algebraic properties of +, ×, 0, and 1 in ℕ, with no axiom scheme of induction. This theory is known to be more than enough for the purpose of proving the first incompleteness theorem.

[7] Cf. McGee (1992:238) and Cook (2006:128).

$$\boldsymbol{Q} \vdash \psi \leftrightarrow (\varphi \leftrightarrow \Delta(\#\psi)),$$

where $\#\psi$ is the canonical term for the Gödel number of ψ. By propositional logic and our assumption that $T \vdash \varphi$, we have

$$(*)\ \ T \vdash \psi \leftrightarrow \Delta(\#\psi).$$

On the other hand, by the well-known textbook proof of the diagonal lemma as presented in Boolos and Jeffrey *op. cit.*, we have

$$\mathbb{N} \vDash \psi \leftrightarrow (\varphi \leftrightarrow \Delta(\#\psi)),$$

so that, because of the falsity of φ in $\mathbb{N}$, we get

$$(**)\ \ \mathbb{N} \vDash \psi \leftrightarrow \neg\Delta(\#\psi).$$

Step 2. For every Δ, our (*) enables us to make the first premise (1) true. We need to take a bit of care to make sure that (2) becomes true and (3) false, and (**) helps us in taking care of them simultaneously. An obvious choice would be taking $\Delta(x)$ to be any tautological predicate like '$x = x$',[8] as $\Delta(\#\psi)$ will then hold up in $\mathbb{N}$ so that the invalidation will be completed because of (**). However, as tautologies are perhaps not immensely interesting, let us offer two other options, namely let $\Delta_1(x)$ be $\mathrm{Pr}_T(x)$, and let $\Delta_2(x)$ be $\neg\mathrm{Pr}_T(x)$, where Pr_T is a provability predicate for T. It is a pair of nice little exercises to show that the resulting

[8] Cf. Leitgeb (2002).

sentences ψ_1 and ψ_2 are such that both $\Delta_1(\#\psi_1)$ and $\Delta_2(\#\psi_2)$ hold in $\mathbb{N}$.[9]

The invalidation is now accomplished. We have introduced a procedure for producing instances of ***A*** and ***F*** showing the invalidity of the (1)-(2)-(3) scheme. Specifically, for each unsound theory *T* we have introduced two particular cases in point: first, a *false* sentence which says about itself that it is provable, and it is provable indeed; second, another *false* sentence which says about itself that it is unprovable and is unprovable indeed.

3.1. Digression: another invalid argument scheme.

Insofar as one is inclined to think that (3) follows from (1)&(2) on the basis that a sentence is true when it has the property it ascribes to itself, one may also be inclined to think that (2) follows from (1)&(3) on the basis that a true self-referential sentences does have the property it ascribes to itself—the (1)-(2)-(3) scheme seems to be *as valid as* the following:

(1') ***A*** says about itself that it is an ***F***

(2') ***A*** is true

therefore

(3') ***A*** is indeed an ***F***.

Now the cute thing about the two-step invalidation presented in this section is that its first

[9] *Solution to the exercises.*

For Δ_1 and ψ_1: If $T \vdash \psi_1 \leftrightarrow \mathrm{Pr}_T(\#\psi_1)$, then, by Löb's theorem (Boolos and Jeffrey (1988:187)), we have $T \vdash \psi_1$, so that, because of arithmetization, $\mathrm{Pr}_T(\#\psi_1)$ holds in $\mathbb{N}$; hence we have $\mathrm{Pr}_T(\#\psi_1)$ true and ψ_1 false, because of (**). Let us note that *T* needs to be substantially richer than ***Q*** to have Löb's rule; it suffices for *T* to contain ***PA*** or its fragment $I\Sigma_1$.

For Δ_2 and ψ_2: if $T \vdash \psi_2 \leftrightarrow \neg\mathrm{Pr}_T(\#\psi_2)$ and *T* is consistent, then, as demonstrated in Gödel's proof of his first incompleteness theorem, ψ_2 is *T*-unprovable, hence, by (**), $\neg\mathrm{Pr}_T(\#\psi_2)$ is true while ψ_2 false.

step suffices to show, in one breath, that *at least one of the two schemes is invalid*.[10] Assuming, intuitively, that deriving (3) from (2) looks as good as doing the reverse, this would show that there must be something wrong with the original (1)-(2)-(3) argument scheme even if we do not go through Step 2 above.

Let us invalidate the (1')-(2')-(3') scheme right away. Suppose T is a consistent and unsound theory that proves a false sentence φ. Take ψ to be the true sentence ¬φ, and take $\Delta(x)$ to be $\neg\mathrm{Pr}_T(\#\neg x)$. Now $\Delta(x)$ is true iff the sentence with the Gödel number x is consistent with T. Since $T \vdash \varphi$, we have $T \vdash \neg\psi$, and so $T \vdash \mathrm{Pr}_T(\#\neg\psi)$. Therefore $T \vdash \neg\psi \leftrightarrow \mathrm{Pr}_T(\#\neg\psi)$, which implies that $T \vdash \psi \leftrightarrow \Delta(\#\psi)$. Thus ψ says about itself that it is a Δ (in the eye of T), and ψ is true; but $\Delta(\#\psi)$ is not true since by the T-provability of ¬ψ, $\mathrm{Pr}_T(\#\neg\psi)$ is true.

4. Gödel's 1931 paper.

In this exegetical section, we examine what we think of as an instance of the (1)-(2)-(3) argument scheme occurring in the introductory and informal section of Gödel (1931), and make a number of comments—in particular, we explore a number of extenuating circumstances in favour of Gödel's informal argument. Once again, we wish to make it explicit that the invalidity of the (1)-(2)-(3) scheme does *not* affect the correctness of Gödel's mathematical results in the technical part of the 1931 paper—though it may, we hope, shed some light on some philosophical issues, to be discussed below in Section 5.

Gödel begins his 1931 paper with an introductory section, wherein, in a very lucid and

[10] Note that by virtue of (*), ψ says about itself that it is a Δ while (**) shows that ψ is true just in case ψ is *not* a Δ. Now if ψ is indeed a Δ, then ψ is not true and (1)-(2)-(3) is invalid; if, on the other hand, ψ is *not* indeed a Δ, then ψ is true and (1')-(2')-(3') is invalid. Therefore, at least one of (1)-(2)-(3) or (1')-(2')-(3') is invalid. Above we saw that the first scheme is invalid; in a moment we will see that so is the second one.

reader-friendly way, he informally introduces the ideas behind the first incompleteness theorem. By the end of the antepenultimate paragraph of that section, he has argued for the independence of what is now called ‘the Gödel sentence’ of PM. His informal proof is quite short, for two reasons: first, he does not go into the details of arithmetization and simply assumes that arithmetization can be done. Secondly, he assumes that PM is sound. The use of the second assumption makes the result actually weaker than what is stated in the technical part of the paper, where he proves a stronger version by means of substituting the soundness assumption by ω-consistency (see 4.3 below).

In his next paragraph, Gödel comments on the analogy of the argument with some antimonies. He could have stopped there—we regret that he did not, for we think the next paragraph contains a fallacy. In the final paragraph of the section, he writes (1931:151, italics in the original),

> From the remark that [*R*(*q*);*q*] says about itself that it is not provable, it follows at once that [*R*(*q*);*q*] is true, for [*R*(*q*);*q*] *is* indeed unprovable (being undecidable).

The way that the proposition is named by Gödel (i.e., ‘[*R*(*q*);*q*]’) reflects its manner of construction, which we will discuss below; however, for the ease of exposition, let us call the sentence ‘*G*’. The argument then seems to be this: *G says about itself that it is not provable, and G is indeed unprovable; therefore G is true*, which is an instance of the (1)-(2)-(3) scheme introduced at the beginning of our paper. We have argued in Section 3 above that, *as it is stated here*, the argument is invalid. Yet in what follows will try to be maximally charitable to Gödel. Besides, we are aware that Gödel himself has said, both in the 1931 paper and elsewhere, that in his introduction he does not claim to be perfectly precise.[11]

4.1. The construction of *G*, whereby Gödel guarantees the truth of the relevant instance of premise (1).

What we nicknamed '*G*' is formally constructed in Gödel (1931:173*f*), where its official description there is '17 Gen *r*'. From the construction of *G* it is quite clear that if *T* is the background theory (be it PM, ZF, or what have you) with provability predicate Pr_T, then

$$(4)\ T \vdash G \leftrightarrow \neg\mathrm{Pr}_T(\#G),$$

which, for a modern reader, is a straightforward application of the diagonal lemma to the formula $\neg\mathrm{Pr}_T(x)$.[12] A quick look at a standard proof of the lemma as presented in Boolos and Jeffrey (1989:173) shows that we also have

$$(5)\ \mathbb{N} \vDash \neg\mathrm{Pr}_T(\#G);$$

that is to say, *G* is unprovable indeed. And what does Gödel mean by *indeed*?

[11] Gödel (1931: 147); cf. Dawson (1984:88*ff*) on Zermelo-Gödel correspondence. For an old critical comment on Gödel's introduction see Helmer (1937:58), who writes "Gödel made one 'mistake', namely that of writing an introduction to his paper …". It seems to us that the point we are making here has not been noted by these logicians.

[12] Though Gödel is evidently exploiting a diagonal technique in the construction of *G*, it would be somehow anachronistic to say that he is using the *diagonal lemma*: it was Rudolf Carnap, in his 1934 *Logische Syntax der Sparche,* who first isolated the lemma as such—see Gödel (1934:363*n*23).

4.2. Gödel's 'indeed'.

When Gödel says that something is true, he means that it is true in the standard model **N**. This is multiply evidenced by his explicit note that the formulæ of the language in question should be understood in a way that in those formulæ,

> no other notions occur but + (addition) and · (multiplication), both for *natural numbers*, and in which the quantifiers (x), too, apply to *natural numbers* only.
>
> [Gödel (1931:145*n*4), emphasis ours.]

Thus by saying that G is *indeed* unprovable, Gödel surely means (5)—namely, that $\neg\mathrm{Pr}_T(\#G)$ holds in ℕ. Once again, the truth of (5) is guaranteed by the way G is constructed.

All these, we submit, show that the informal argument of the passage we quoted at the beginning of this section is

(4) $T \vdash G \leftrightarrow \neg\mathrm{Pr}_T(\#G)$

(5) $\mathbb{N} \vDash \neg\mathrm{Pr}_T(\#G)$

therefore

(6) $\mathbb{N} \vDash G$,

whose logical form we recognize as (1)-(2)-(3). Gödel says that the conclusion follows 'at once' ['sofort' in the original German] which, to our ears, suggests that no extra assumptions are needed. But perhaps he did have additional assumptions in mind? Let us see.

4.3. What Gödel officially assumed: not soundness, but ω-consistency.

Gödel's informal argument is fallacious *only if* the background theory *T* is not sound. For if all the theorems of *T* are true in ℕ*,* so is the biconditional of (4), which, together with (5), makes *G* true in ℕ, as asserted by (6). The assumption of soundness may seem to be what Gödel had in mind, for he says, in the very first paragraph of his paper, that his result holds in particular for every extension of PM,

> provided no false proposition of the kind specified … become provable owing to the added axioms. [Gödel (1931:145*f*).]

In itself, this suggests that Gödel is talking about sound theories. Also, in the next-to-last paragraph of his Section 1, he says (1931:151) that his method of proof is applicable to any formal system which is, first, of sufficient expressive power and, secondly, is such that "every provable formula is true in the interpretation considered". This, too, suggests that he is talking about sound theories only.

However, soundness is definitely *not* what Gödel had in mind as a required property of the theories under consideration in his paper, for he immediately adds,

> The purpose of carrying out the above proof with full precision in what follows is, among other things, to replace *the second of the assumptions just mentioned* by a purely formal and much weaker one. [Ibid., emphasis added.]

That much weaker and purely formal assumption is ω-consistency, so baptized by Gödel himself, which is introduced just before the statement of the first incompleteness theorem and is mentioned in it (1931:173).[13] So, although Gödel talks about truth and soundness in his informal and introductory Section 1, he avoids talking about it in the technical part of the paper—in fact, he explicitly tells us that one purpose of going formal and precise was just replacing his talk about *truth* of the theorems of certain theories (say PM) with the talk about a purely syntactical property of them.

Officially speaking, Gödel's first incompleteness theorem assumes that the theories in question are ω-consistent, a condition weaker than soundness. An invaluable source here is Isaacson (2011), wherein we find a theorem—Isaacson's Proposition 19, attributed to Kreisel—telling us about a false sentence K such that PA + K is ω-consistent.[14] This allows us to invalidate the (1)-(2)-(3) scheme via presenting an ω-consistent theory: Simply take $\Delta(x)$ to be the formula '$x = \#K$', and take ψ and T to be the sentence K and the theory PA + K, respectively. Moral: the assumption of ω-consistency of the background theory is not strong enough to save the (1)-(2)-(3) scheme from invalidity.

Having said that, we acknowledge that we *might* have invalidated Gödel's argument only via over-generalizing it. Several additional premises (including some restrictions on the complexity of the relevant predicates and formulæ) are examined in Lajevardi & Salehi (2019)

[13] By definition, a theory T (in the language of arithmetic) is *ω-consistent* iff there is no formula $\xi(x)$ such that T proves the sentence $\exists x \neg \xi(x)$ *and* proves all of the following sentences: $\xi(\mathbf{1})$, $\xi(\mathbf{2})$, ..., $\xi(\mathbf{n})$, ... (for every natural number n), where boldface roman characters denote the standard terms for the corresponding numerals. [For a formal definition see Gödel (1931:173) or Smoryński (1977:851*f*).]

Simple consistency of PM suffices for showing that G is PM-unprovable. It seems that the sole purpose of introducing ω-consistency was that Gödel was unable to show the irrefutability of G by the mere assumption of consistency. The assumption of ω-consistency was weakened to simple consistency by J.B. Rosser in his (1936), who showed the undecidability of *another* sentence.

[14] This is an almost immediate consequence of formalizing the notion of ω-consistency. See also Lindström (1997:36).

which render the scheme valid. More specifically, if ***A*** and ***F*** are of Π_1 complexity, then the extra assumption of ω-consistency guarantees the validity of the scheme –see Proposition 3 in the Appendix. Both *G* and $\neg \mathrm{Pr}_T$ are Π_1 for all the theories we have mentioned in this paper.

Did Gödel actually commit the fallacy? Perhaps not: perhaps he meant to show the truth of *G* only for sound theories and did not intend to assert the truth of *G* in general for all the theories which are subject to the official version of his first incompleteness theorem. Alternatively, he might have been careless to explicitly mention that *G* and $\neg \mathrm{Pr}_T$ are Π_1 [see footnote 17 below for Gödel's terminology]. Though such considerations clear Gödel himself of the charge of talking fallaciously, they will not acquit a number of philosophers who talked about "seeing" the truth of *G*. We return to this in 5.2 below.

4.4. The irrelevance of the notion of self-reference.

There is an ongoing discussion on what it means for a formula to say something about itself—see Halbach and Visser (2014). By no means do we want to dismiss this scholarly issue; however, if the task is to evaluate *Gödel's* argument as presented in his introductory section, we take it to be quite obvious that he had thought of the very sentence *G* as a sentence saying about itself that it is unprovable—it is, in fact, a paradigm of self-referential sentences. So, whatever the correct analysis of the concept of self-attribution might turn out to be, *what Gödel, qua the author of (1931), had in mind* must be something retrievable from what he has done in that paper. Given that his apparatus for constructing *G* is a version of the diagonal lemma, we think there is no choice but to think that, for Gödel, ***A***'s saying about itself that it has property ***F*** just means that the biconditional $\boldsymbol{A} \leftrightarrow \boldsymbol{F}(\#\boldsymbol{A})$ is provable in the system. We therefore find the (4)-(5)-(6) argument the only thing that Gödel could have had in mind in this connection, the logical form whereof we recognize as the (1)-(2)-(3) scheme.

We do agree with the statement that Halbach and Visser (2014:672) quote sympathetically from logician Craig Smoryński, that the "notion of a sentence's expressing something about itself has not proven fruitful", and we think Kripke is right when, concerning the argument for the independence of *G*, says

> The argument can be carried through without noticing explicitly that *G* says something about itself (i.e. that it itself is not provable). [Kripke (2014:239).]

Here we wish to make two minor comments. First, these insightful remarks are made in the context of *formal* mathematical logic: Smoryński and Kripke justly remark that, so far as Gödel's technical results are concerned, Gödel did not have to talk about self-reference. It should be clear, however, that an evaluation of the *informal* argument presented in the introductory part of Gödel's paper requires a working notion of self-reference. Once again, we cannot see what Gödel could have had in mind other than an informal variant of what we have presented as (4)-(5)-(6).

Secondly, if, for whatever reason, Gödel felt an urge to talk about self-reference (perhaps in order to make his introduction more attractive to the general reader), with the wisdom of hindsight we know that instead of self-reference in the standard textbook manner, which for every formula $\Delta(x)$ provides us with a sentence ψ such that the biconditional $\psi \leftrightarrow \Delta(\#\psi)$ is a theorem of the background theory (and let us call this kind of self-reference *indirect*), he could have exploited a *direct* self-reference in such a way that, for every formula $\Delta(x)$ we get a sentence ψ such that ψ just *is* the formula $\Delta(\#\psi)$.[15] Now, instead of extra assumptions concerning the soundness or complexity of Δ and φ, here is another way of providing a valid

[15] So far as we know, the idea of such a direct self-reference goes back to Kripke (1975:693). See also Visser (2004:159*f*).

version of the (1)-(2)-(3) scheme: instead of a traditional indirect self-reference, one may go the direct way and replace (1) with

(1'') $\boldsymbol{A}$ is the sentence $\boldsymbol{F}(\#\boldsymbol{A})$,

which makes the scheme trivially valid.

5. Gödelian sentences and truth.

5.1. Why did Gödel talk about truth?

The official statement of Gödel's first incompleteness theorem (1931:173) does not mention truth: it is simply a theorem to the effect that every theory of a certain specified kind—which is specified totally syntactically—is *incomplete*, in the sense that there are sentences in the language of the theory which are axiomatically undecidable on the basis of that theory: they are neither provable nor refutable. If you are a realist, as surely Gödel was from the 1940s onwards (if not earlier), you are inclined to say that each undecidable sentence is "really" either true or false—hence the *popular* version of the first incompleteness theorem: for every theory of a certain specified kind, there are *true* sentences which are unprovable on the basis of that theory. Yet this is something which is not formally presented in Gödel's Theorem. And one need not be a realist with respect to the sentences expressible in the language of arithmetic.

Regardless of one's verdict on Gödel in connection with the (1)-(2)-(3) argument in his introductory section, one may observe that that section has two drawbacks in comparison with the technical parts of the paper:

I. By posing a stronger requirement on the theories in question (soundness, instead of ω-consistency), it weakens the *content* of the first incompleteness theorem, making it applicable to a narrower class of theories.

II. By introducing the notion of truth (via soundness), it makes the *proof* of the theorem acceptable to a smaller class of mathematicians.

We have already (in 4.3) talked about ω-consistency, hence substantiated (I). As for (II), note that Gödel's formal and official proof, which is a *tour de force* of giving a purely constructive and syntactical proof of the first incompleteness theorem, is acceptable to realists and anti-realists alike;[16] but by talking about truth, Gödel actually makes parts of the argument of his introduction susceptible to becoming unconvincing for those readers who are of intuitionist persuasion. In fairness to Gödel, however, we admit that (I) and (II) are reasonable prices to pay—at least in an introductory section—to attain a higher level of perspicuity of exposition, attained by Gödel's introduction.

However, a nagging question remains: why did Gödel pause to argue that *G* is true? Even if this particular instance of the (1)-(2)-(3) scheme is saved from invalidity by the presence of the assumption of soundness in the context of Gödel's introduction, it is not quite clear how he would restore the argument for the truth of *G* in the context of unsound theories without appealing to the complexity issues (referred to in 4.3 above), which are absent from his 1931 paper.[17] And it seems to us that, even for the sake of elementary lucid exposition of topics he later deals with in full precision, the question about the truth of *G* is simply irrelevant. Why,

[16] Gödel himself comments (1931:177*n*45a) that his technical proof is intuitionistically acceptable.

[17] We are *not* saying that Gödel was unaware of well-known and elementary facts about the complexity of formulæ. He surely was—thus, for example, in a short note of the same period he talks about propositions "of the type of Goldbach" (1931a:203), by which here—and surely elsewhere—he means what is now called Π_1-formulas (see also Smith (2013:154)).

then, did he care to talk about the *truth* of *G*?

By way of speculation, we submit that the reason may have something to do with what might be thought of as Gödel's target or antagonist, the school of David Hilbert. Talking about the relationship between Gödel's incompleteness theorems and Hilbert's programme goes well beyond our competence (see Feferman (1984)); we just want to make an amateurish conjecture that, after what he perhaps thought of as a fatal blow at Hilbert's programme via the mathematical content of his incompleteness theorems, Gödel wanted to add a *coup de grâce* by arguing that while for every system of the specified kind the formula *G* is axiomatically undecidable, it *is* decidable via some acceptable reasoning accessible to human beings, hence mathematics cannot be captured by any axiomatics. We do regret that he has done so.

Now let us talk more rigorously.

5.2. On the truth of Gödelian sentences.

Concerning the age-old question of truth of Gödel sentences we think some insight can be gained from what we have investigated so far. Of the first-rate and/or well-known works on this question the following references readily come to mind: Dummett (1963), Smoryński (1977 and 1985), Boolos (1990), Milne (2007), Raatikainen (2005), Serény (2011), Shapiro (1998), Tennant (2002), and the collection edited by Horsten and Welch (2016). We wish to make two specific points before presenting our analysis of what is going on in a good number of such discussions.

5.2.1. A consistent theory may have more than one Gödelian sentences—even up to truth-value. Following Lajevardi & Salehi (2021), we choose to talk about Gödelian sentence*s* of a theory *T* not *the* Gödel sentence of *T*, for an unsound *T* may have a true I-am-unprovable sentence as well as a false one (though they are equivalent in the eye of *T*), in which case it is bizarre to talk about *the* I-am-unprovable sentence. This, however, may be more a matter of propriety of speech than anything of great logico-mathematical significance.

5.2.2. *T* has all its Gödelian sentences true if and only if *T* is sound. For this fact we present a rigorous proof. In the following lemma we assume that *T* satisfies Löb's derivability conditions.

LEMMA. *If τ is a T-provable sentence and γ is a Gödelian sentence of T, then $\tau\&\gamma$ is a Gödelian sentence of T as well.*

Proof. By the *T*-provability of τ we have $T \vdash (\tau\&\gamma) \leftrightarrow \gamma$. Therefore, by derivability conditions, we have $T \vdash \mathrm{Pr}_T(\#[\tau\&\gamma]) \leftrightarrow \mathrm{Pr}_T(\#\gamma)$ hence $T \vdash \neg\mathrm{Pr}_T(\#\gamma) \leftrightarrow \neg\mathrm{Pr}_T(\#[\tau\&\gamma])$. Since γ is a Gödelian sentence of *T*, we have

$$T \vdash (\tau\&\gamma) \leftrightarrow \gamma \leftrightarrow \neg\mathrm{Pr}_T(\#\gamma) \leftrightarrow \neg\mathrm{Pr}_T(\#[\tau\&\gamma]),$$

which shows that $\tau\&\gamma$ is a Gödelian sentence of *T*. *QED*

Now we have

THEOREM. *T is sound if anf only if all Gödelian sentence of T are true.*[18]

Proof. If γ is a Gödelian sentence of a sound theory *T*, then we have $T \vdash \gamma \leftrightarrow \neg\mathrm{Pr}_T(\#\gamma)$, which implies that the biconditional $\gamma \leftrightarrow \neg\mathrm{Pr}_T(\#\gamma)$ is a true sentence. But by the *T*-

[18] See Proposition 1 in the Appendix for a more general result.

unprovability of γ (Gödel's proof), the sentence $\neg\mathrm{Pr}_T(\#\gamma)$ is true; therefore γ is true.

On the other hand, if *T* is not sound then there is a false *T*-provable sentence τ. Let γ be any Gödelian sentence of *T* (whose existence is demonstrated by the diagonal lemma). Then τ&γ is a false sentence which is, by the Lemma, a Gödelian sentence of *T*. Therefore *T* has a false Gödelian sentence. *QED*

Now, to the main business of this section.

5.2.3. Are Gödelian sentences true, after all? It seems to us that a good number of the works on the truth of Gödelian sentences are shaky or even outright fallacious. This has been observed or claimed by some philosophers such as Shapiro (1998); yet we think we can offer a brief and more systematic analysis.

i. Occasionally, one encounters an argument like this: *Each Gödelian sentence is true because it says that it is unprovable and it is indeed unprovable (if the theory is consistent).* In itself, this is a fallacy (as shown in our Section 3 above). To say the least, the italicized argument needs more assumptions for its validity. It's a pity to see this fallacious-or-gappy argument in some early editions of first-rate textbooks such as Boolos and Jeffrey (1989:186) or Mendelson (1979:159).[19]

ii. Let us not worry about the *reasoning* of those, like Dummett (1963), who argue that Gödelian sentences of consistent theories are true; let us ask if their *conclusion* is correct—that is to say: Does the consistency of a theory *T* guarantee the truth of all its Gödelian sentences? (We will assume that the theories in question are recursively enumerable

[19] Happily, the mistakes are not there anymore. Concerning Boolos and Jeffrey (1989), it is noteworthy that from the fourth edition onwards (prepared by John P. Burgess), the fallacious passage is simply omitted. Mendelson, on the other hand, now shows the truth of Gödelian sentences only for sound theories—thus on page 209 of the sixth edition, published in 2015, we have the assumption that the theory "is a true theory".

extensions of Robinson's ***Q***.) Theorem 3.3 in Lajevardi & Salehi (2021) answers this affirmatively, with a proviso. Recall that P is a Gödelian sentence of T iff the biconditional $P \leftrightarrow \neg\mathrm{Pr}_T(\#P)$ is provable in T. Now if, *moreover,* the biconditional is also true in the standard model $\mathbb{N}$, then P is true if and only if T is consistent. The condition on the truth of the biconditional $P \leftrightarrow \neg\mathrm{Pr}_T(\#P)$ is easily satisfied *if* P is constructed via the celebrated diagonal lemma.

iii. BUT, not all Gödelian sentences of a theory need satisfy this extra condition. The above argument does not show that *every* Gödelian sentence of a consistent theory is true; rather, it shows it only for those Ps asserting their own unprovability such that the biconditional $P \leftrightarrow \neg\mathrm{Pr}_T(\#P)$ is also true in $\mathbb{N}$. Enter our Theorem:

Even if the reader does not share our worry concerning the impropriety of the term 'the Gödel sentence', he or she may concede this much: the argument mentioned in 5.2.3ii does not show that every sentence which is T-equivalent to its own T-unprovability is true if T is consistent. If one's question is whether every sentence asserting its own unprovability is true, one should be noted that, because of the Theorem we proved above, the truth of all such sentences is equivalent to the soundness of T, and—you may recall—soundness is *much stronger* an assumption than ω-consistency, which is in turn stronger than simple consistency.[20]

[20] It may even be demonstrated that every false sentence is a Gödelian sentence of some unsound theory (see the Appendix).

6. Conclusion.

As a matter of logical fact, the (1)-(2)-(3) argument scheme displayed at the beginning of this paper is invalid even if we assume that the background system is ω-consistent (and *a fortiori*, even if the system is consistent). Gödel's particular instance of it—the (4)-(5)-(6) argument—is valid (even for unsound theories), but only because the involved sentence and predicate are of certain complexity—that is to say, because they are both Π_1, or "of the type of Goldbach" in Gödel's terminology. Interestingly, as we will see in the Appendix, the way Gödel constructs his self-referential sentence makes the premise (5) a consequence of his (4), and therefore redundant.

Apart from logic chopping, our observations are of some consequence concerning the debate on the truth of *Gödelian sentences*—or, to avoid verbal disputes, the truth of *each and every sentence which asserts its own unprovability*. Our Theorem proves that nothing less than the full power of soundness of the system guarantees the truth of all such sentences.

References

Boolos, G.S. (1990). On “seeing” the truth of the Gödel sentence. *Behavioral and Brain Science*, 13: 654–657. Reprinted in: R. Jeffrey (ed.), G. Boolos, *Logic, Logic, and Logic*, Harvard University Press (1999), pp. 389-391.

Boolos, G.S., and R. Jeffrey (1989). *Computability and Logic*, 3rd ed., Cambridge University Press.

Cook, R.T. (2006). There are non-circular paradoxes (but Yablo’s isn’t one of them!). *The Monist*, 89: 118-149.

Dawson, J.W. Jr. (1984). The reception of Gödel’s incompleteness theorems. *PSA 1984*, 2: 253–271. Reprinted in: T. Drucker (ed.), *Perspectives on the History of Mathematical Logic* (1991) pp. 84–100.

Dummett, M.A.E. (1963). The philosophical significance of Gödel’s theorem. *Ratio* 5: 140-155. Reprinted in: Dummett, *Truth and other Enigmas*, Duckworth, 1978, pp. 186–201.

Feferman, S. (1984). Kurt Gödel: conviction and caution. *Philosophia Naturalis,* 21:546–562. Reprinted in: Feferman, *In the Light of Logic*, Oxford University Press, 1998, pp. 150–164.

Gödel, K. (1931). On formally undecidable propositions of *Principia mathematica* and related systems I. Reprinted in: Gödel (1986), pp. 145–195.

Gödel, K. (1931a). Discussion on providing a foundation for mathematics. Reprinted in: Gödel (1986), pp. 201–205.

Gödel, K. (1934). On undecidable propositions of formal mathematical systems. Reprinted in: Gödel (1986), pp. 346–371.

Gödel, K. (1986). *Kurt Gödel Collected Works, Volume I: Publications 1929–1936*, edited by S. Feferman *et al.*, Oxford University Press.

Halbach, V. and A. Visser (2014). Self-reference in arithmetic I & II. *The Review of Symbolic Logic*, 7: 671–691, 692–712.

Helmer, O. (1937). Perelman versus Gödel. *Mind*, 46: 58–60.

Horsten, L. and P. Welch, eds. (2016). *Gödel's Disjunction: The Scope and Limit of Mathematical Knowledge*, Oxford University Press.

Issacson, D. (2011). Necessary and sufficient conditions for undecidability of the Gödel sentence and its truth. In: D. DeVidi et al. (eds.), *Logic, Mathematics, Philosophy, Vintage Enthusiasms: Essays in Honour of John L. Bell*, Springer, pp. 135–152.

Kripke, S.A. (1975). Outline of a theory of truth. *The Journal of Philosophy*, 72: 690–716.

Kripke, S.A. (2014). The road to Gödel. In: J. Berg (ed.), *Naming, Necessity, and More: Explorations in the Philosophical Work of Saul Kripke*, Palgrave MacMillan, pp. 223–241.

Lajevardi, K., and S. Salehi (2019). On the arithmetical truth of self-referential sentences. *Theoria*, 85: 8-17.

Lajevardi, K., and S. Salehi (2021). There may be many Gödel sentences. *Philosophia Mathematica*, 29: 278-287.

Leitgeb, H. (2002). What is a self-referential sentence? Critical remarks on the alleged (non-)circularity of Yablo's paradox. *Logique et Analyse*, 45: 3-14.

Lindström, P. (1997). *Aspects of Incompleteness*. Springer.

Löb, M.H. (1955). Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic*, 20: 115-118.

McGee, V. (1992). Maximal consistent sets of instances of Tarski's scheme (T). *Journal of Philosophical Logic*, 21: 235-241.

Mendelson, E. (1979). *Introduction to Mathematical Logic*. 2nd ed., D. van Nostrand Company.

Milne, P. (2007). On Gödel sentences and what they say. *Philosophia Mathematica*, 15: 193–226.

Raatikainen, P. (2005). On the philosophical relevance of Gödel's incompleteness theorems. *Revue Internationale de Philosophie,* 59:513–534.

Rosser, B. (1936). Extensions of some theorems of Gödel and Church. *The Journal of Symbolic Logic*, 1: 87–91.

Serény, G. (2011). How do we know that the Gödel sentences of a consistent theory is true?. *Philosophia Mathematica*, 19: 47–73.

Shapiro, S. (1998). Induction and indefinite extensibility: the Gödel sentence is true, but did someone change the subject?. *Mind*, 107: 597–624.

Smith, P. (2013). An Introduction to Gödel's Theorems. 2nd ed., Cambridge University Press.

Smoryński, C. (1977). The Incompleteness Theorems. In: J. Barwise (ed.), *Handbook of Mathematical Logic*, North-Holland, pp. 821–865.

Smoryński, C. (1985). *Self-Reference and Modal Logic*. Springer.

Tennant, N. (2002). Deflationism and the Gödel phenomena. *Mind*, 111: 551–582.

Visser, A. (2004). Semantics and the liar paradox. In: D. Gabby and F. Günthner (eds.), *Handbook of Philosophical Logic,* 2nd ed., Volume XI, pp 149–240.

Appendix.

We present some refinements and generalizations of some technical results presented in the paper. For the sake of brevity, most of the proofs are omitted.

A1. More self-referential sentences.

Gödel's proof made the self-referential sentence 'I am unprovable' famous; his proof of the first incompleteness theorem shows that, for certain theories, every such sentence is in fact unprovable. Next, in the early 1950s, Leon Henkin asked: What if a sentence says about itself that it is provable? Löb's answer (1955) is that such sentences are in fact provable. And one moral of our paper is that from the very fact of the (un)provability of an I-am-(un)provable sentence, its truth does not follow.

Call a predicate Δ *self-fulfilling* with respect to a given theory T iff the following is the case: every sentence ψ which, with respect to T, says about itself that it is a Δ, is indeed a Δ. Thus (see our two examples in Section 3 above) *is provable* and *is unprovable* are both self-fulfilling.

Here is a more interesting self-fulfilling predicate: *is decidable*. Suppose that, in the eye of T, a sentence ψ says about itself that it is axiomatically decidable, i.e., $T \vdash \psi \leftrightarrow \mathrm{Pr}_T(\#\psi) \vee \mathrm{Pr}_T(\#\neg\psi)$. Then we have $T \vdash \neg\psi \leftrightarrow \neg\mathrm{Pr}_T(\#\psi) \wedge \neg\mathrm{Pr}_T(\#\neg\psi)$, so that $T + \neg\psi$ proves the consistency of $T + \neg\psi$. Hence, by Gödel's second incompleteness theorem, $T + \neg\psi$ is *in*consistent and we have $T \vdash \psi$. By arithmetization, $\mathrm{Pr}_T(\#\psi)$ is true in $\mathbb{N}$, and so is $\mathrm{Pr}_T(\#\psi) \vee \mathrm{Pr}_T(\#\neg\psi)$. Therefore, every sentence which, in the eye of T, says about itself that it is T-decidable is in fact T-decidable—more specifically: any such sentence is T-provable.

Let us say that Δ is *self-falsifying* with respect to a given theory *T* iff every sentence which says in *T* about itself that it is a Δ, is indeed *not*-Δ. Examples include *is refutable* and *is consistent with* predicates.

Our argument in §3 actually proves the following more general proposition.

Proposition 1.[21] *Let T be a consistent theory (containing Robinson's arithmetic* **Q***). Let Δ be a self-fulfilling predicate and let Θ be a self-falsifying predicate with respect to T. Then the following are equivalent:*

(i) *The soundness of T.*

(ii) *The truth of all the sentences which assert inside T that they are Δ.*

(iii) *The falsehood of all the sentences asserting inside T that they are Θ.*

We remark, without offering a proof, that there are predicates that are neither self-fulfilling nor self-falsifying—examples include the predicates *is a universal sentence* and *is an existential sentence.*

A2. False Gödelian sentences.

In 5.2.2 above, we proved that a theory is sound if and only if all its Gödelian sentences are true. Perhaps more interestingly, every false sentence can be a Gödelian sentence of a sufficiently strong sound theory. This follows from the next proposition.

Proposition 2. *A sentence is T-unprovable if and only if it is a Gödelian sentence of a consistent extension of T.*

[21] The experts are invited to compare this to Theorem 24.7 of Smith (2013:182).

Corollary. *For every sound theory S and every false sentence φ, there exists a consistent extension T of S such that φ is a Gödelian sentence of T.*

A3. More on ω-consistency.

Which levels of soundness are guaranteed by ω-consistency? Which Gödelian sentences of ω-consistent theories are true? Our answer generalizes Theorem 17 of Isaacson (2011).

Proposition 3. *Let T be an ω-consistent extension of* **Q**. *Then every T-provable Π_3-sentence is true, and so is every Gödelian Π_3-sentence of T.*

This is a boundary result, since, as we have already mentioned in 4.3, ω-consistent theories may have false provable Σ_3-sentences. By the Lemma of the subsection 5.2.2 above, the conjunction of that provable false Σ_3-sentence with an arbitrary Gödelian Π_1-sentence results in a false Gödelian Σ_3-sentence. Hence ω-consistent theories may have false Gödelian Σ_3-sentences, though all of their Gödelian Π_3-sentences are true by Proposition 3.

We hope by now the reader shares our view concerning the fallacious (or enthymematic) character of the passage we quoted from Gödel (1931:151) whose logical form we recognized as the (1)-(2)-(3) scheme—he or she may now produce what Gödel should have written instead (say by adding a premise about the complexity of the relevant predicate and sentence, or using a kind of direct self-reference à la Kripke).

Let us conclude by noting that Gödel's argument does in fact contain a *redundant* premise. Recall that we formalized Gödel's argument thus:

(4) $T \vdash G \leftrightarrow \neg \mathrm{Pr}_T(\#G)$

(5) $\mathbb{N} \vDash \neg \mathrm{Pr}_T(\#G)$

therefore

(6) $\mathbb{N} \vDash G$.

Now what we said in this Appendix shows that (5) is actually redundant, for *is unprovable* is a self-fulfilling predicate for consistent theories.