THEORIA

THEORIA, 2019, **85**, 8–17
doi:10.1111/theo.12169

On the Arithmetical Truth of Self-Referential Sentences

by

KAAVE LAJEVARDI
La Société des Philosophes Chômeurs

and

SAEED SALEHI
University of Tabriz

Abstract: We take an argument of Gödel's from his ground-breaking 1931 paper, generalize it, and examine its validity. The argument in question is this: *the sentence G says about itself that it is not provable, and G is indeed not provable; therefore, G is true.*

Keywords: Gödel's first incompleteness theorem, the Gödel sentence, self-reference, truth, arithmetic, soundness, ω-consistency

1. Introduction

As is well known, Gödel begins his 1931 masterpiece with an introductory section, Section 1, wherein he explains the main ideas behind the (first) incompleteness theorem in an informal and intuitive way, with an explicit caveat that those remarks are being made "without any claim to complete precision". What he does in that section is, among other things, to introduce, in a very lucid way, the idea of encoding the syntax and that of diagonalization.[1]

Our primary purpose in this article is not to philosophize about the notion of truth or to do exegetical work on its role in Gödel's classic paper. Rather, we wish to draw attention to an informal argument, to the effect that the Gödel sentence of the system *Principia Mathematica* is true (that is, true in $\mathbb{N}$), which is presented by Gödel in his introductory section. The argument in question goes as follows (Gödel, 1931, p. 151; original emphasis):

1 Gödel says that his method of proof is applicable to any formal system that has two conditions, the second of which being "every provable formula is *true* in the interpretation considered" (Gödel, 1931, p. 151, emphasis added). He then writes, "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*" (Gödel, 1931, emphasis added). This "much weaker" condition, called *ω-consistency* by Gödel, is one of the points of focus in our discussion. Note that *ω*-consistency is stronger than (simple) consistency: see, e.g., (Isaacson, 2011, Theorems 14,15).

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).

Here [$R(q)$;q] is what is now called the *Gödel sentence* of a theory (or "system") which is subject to the first incompleteness theorem (e.g., Peano Arithmetic PA, or any of its recursively axiomatizable ω-consistent extensions). This is nowadays denoted by G, which is defined to be any sentence P which is (provably) equivalent to $\neg\mathsf{Pr}(\#P)$, where Pr is the provability predicate of the theory and $\#P$ is the standard term for the Gödel number of P (see Section 3 below for more on the definition).

As we understand the above passage, its logical form is the following, where A is an arbitrary sentence expressible in the language of the theory:

(1) A says about itself that it has a property F.
(2) A indeed has the property F.
(3) Therefore: A is true.

That Gödel says that it follows *at once*[2] that G is true suggests that, in Gödel's view, we are not dealing with an enthymeme–it seems to us that, for Gödel, the argument scheme has no missing premises. It is our task in the next section to argue that the (1)–(3) argument scheme is *invalid*, that is to say, there are situations wherein the premises are true while the conclusion false. Naturally enough, the validity of the argument hinges on its terms, in particular on what it is for a sentence to be "true", and what is meant by a sentence "saying about itself" that it has a certain property. As for the first term, it is almost obvious from Gödel's introductory section that, like many modern writers in mathematical logic, when he writes "true" simpliciter, he means *true in the standard model*, $\mathbb{N}$.[3] Regarding the notion of saying something about oneself, we shall consider two interpretations that might be ascribed to Gödel.

2. The Invalidity of the Argument Scheme

How are we to understand the expression "A says about itself that it has the property F"? Modulo an agreement over the meaning of "holding", to which we shall return shortly, we find it quite plausible to think that if the conditional $A \rightarrow F(\#A)$ holds, then A says, *inter alia*, that A has property F. Now, to stop here would provide us an easy–perhaps too easy–invalidation of Gödel's informal argument, for if we take, as we should, "holding" to mean being true in

2 Gödel's adverb "sofort" could also be translated as *immediately.*

3 Thus Gödel's (1931) footnote 4 on page 145: "no other notions occur but + (addition) and (multiplication), both for natural numbers, and in which the quantifiers (x), too, apply to natural numbers only."

the standard model of natural numbers $\mathbb{N}$, and take "saying" to mean implying, then, taking γ to be the sentence $\mathsf{G} \wedge (0 = 1)$, the sentences $\gamma \rightarrow \neg\mathsf{Pr}(\#\gamma)$ and $\neg\mathsf{Pr}(\#\gamma)$ both hold but γ does not. Let us then assume that Gödel intended each sentence in question in premise (1) to say *exactly* about itself that it has property F; to wit: the biconditional $A \leftrightarrow F(\#A)$ holds. And this, in fact, is what is taken by some authors to be the meaning of "A says about A that it has property F" (see Halbach & Visser, 2014; Milne, 2007; and references therein).

As for the meaning of "holding", we already presented evidence for the claim that by this, Gödel meant satisfaction in the standard model, $\mathbb{N}$. However, let us recognize another reading of it–which is also suggested in the literature (see Halbach & Visser, 2014; Milne, 2007 and references therein)–namely, *being provable in a given theory.* Having fixed a theory T (which we suppose to be an ω-consistent, recursively axiomatizable extension of PA), we then have eight possible ways of interpreting the (1)–(3) argument scheme. Here is the complete list, of which we find (VII), (V) and (III) the most interesting.[4]

$$\text{(I)}\ \frac{\mathbb{N} \vDash A \leftrightarrow F(\#A), \quad \mathbb{N} \vDash F(\#A)}{\mathbb{N} \vDash A} \qquad \text{(II)}\ \frac{\mathbb{N} \vDash A \leftrightarrow F(\#A), \quad \mathbb{N} \vDash F(\#A)}{T \vdash A}$$

$$\text{(III)}\ \frac{\mathbb{N} \vDash A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{\mathbb{N} \vDash A} \qquad \text{(IV)}\ \frac{\mathbb{N} \vDash A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{T \vdash A}$$

$$\text{(V)}\ \frac{T \vdash A \leftrightarrow F(\#A), \quad \mathbb{N} \vDash F(\#A)}{\mathbb{N} \vDash A} \qquad \text{(VI)}\ \frac{T \vdash A \leftrightarrow F(\#A), \quad \mathbb{N} \vDash F(\#A)}{T \vdash A}$$

$$\text{(VII)}\ \frac{T \vdash A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{\mathbb{N} \vDash A} \qquad \text{(VIII)}\ \frac{T \vdash A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{T \vdash A}$$

Of these, (I) and (VIII) are of course valid because of the truth-condition of the material conditional and Modus Ponens, respectively. For all other cases, we will present triples (A, F, T) which invalidate them.

Theorem 2.1. *The argument* (IV) *is invalid for* $A = \mathsf{G}$, $F(x) \equiv (x = \#\mathsf{G})$, *and* $T = \mathsf{PA}$.

4 With the wisdom of hindsight, we now know that some of these schemata unduly overgeneralize Gödel's informal argument, for according to the second incompleteness theorem, which Gödel intended to present in a never written sequel to his 1931 work, no (consistent) theory can prove its own consistency. Thus, no consistent theory can prove the unprovability of any sentence, but each of (III), (IV), (VII) and (VIII) has a premise to the effect that the F of (the Gödel number of) A is provable in the theory. It is for the sake of comprehensiveness that we list all the eight possible readings.

Proof. Obviously, $\mathbb{N} \vDash F(\#G)$ and $T \vdash F(\#G)$. By Gödel's proof, $\mathbb{N} \vDash G$ holds and so $\mathbb{N} \vDash G \leftrightarrow F(\#G)$. On the other hand, by Gödel's theorem, $T \nvdash G$. □

Theorem 2.2. *The arguments* (II) and (VI) *are invalid for* $A = G$, $F(x) \equiv \neg\mathsf{Pr}(x)$, *and* $T = \mathsf{PA}$.

Proof. We already have $\mathbb{N} \vDash G \leftrightarrow \neg\mathsf{Pr}(\#G)$. Since by Gödel's theorem we have $\mathsf{PA} \nvdash G$, it follows that $\neg\mathsf{Pr}(\#G)$ is true, i.e., $\mathbb{N} \vDash \neg\mathsf{Pr}(\#G)$. □

Of course the arguments (III), (V) and (VII) are all valid if T is a *sound* theory (i.e., when $\mathbb{N} \vDash T$). As we mentioned in the Introduction, Gödel replaces the soundness condition with the weaker condition of ω-consistency. As Isaacson (2011) mentions, Gödel states in (1931), without giving any argument, that ω-consistency is "much weaker" than soundness. It is shown in Isaacson (2011, Proposition 19), (with a proof attributed to Kreisel in the 1950s) that there exists a false sentence K such that the theory $\mathsf{PA} + K$ is ω-consistent. Moreover, the sentence K can be taken to be a diagonal sentence of a formula $H(x)$; i.e., $K \leftrightarrow H(\#K)$ holds (is PA-provable and true in $\mathbb{N}$).

Theorem 2.3. *The argument* (V) *does not hold for* $A = K$, $F(x) \equiv (x = \#K)$, *and* $T = \mathsf{PA} + K$.

Proof. By $\mathsf{PA} \vdash F(\#K)$, we have $\mathsf{PA} + K \vdash K \leftrightarrow F(\#K)$. Now, $\mathbb{N} \vDash F(\#K)$ holds trivially. By Isaacson (2011, Proposition 19), $\mathbb{N} \nvDash K$. □

Theorem 2.4. *Neither* (III) *nor* (VII) *holds for* $A = K$, $F(x) \equiv H(x)$, *and* $T = \mathsf{PA} + K$.

Proof. We already have $\mathsf{PA} \vdash K \leftrightarrow H(\#K)$ and $\mathbb{N} \vDash K \leftrightarrow H(\#K)$ by definition, and so $\mathsf{PA} + K \vdash K \leftrightarrow H(\#K)$ holds too. The latter also implies that $\mathsf{PA} + K \vdash H(\#K)$. Finally, by Isaacson (2011, Proposition 19), we have $\mathbb{N} \nvDash K$. □

Let us acknowledge the fact that perhaps it is only by overgeneralizing Gödel's informal argument that we are making it invalid. Had we not abstracted from the specific properties of A, F and T, Gödel's informal argument would be valid, even in the interesting cases of (III), (V) and (VII) (though it would lose much of its appeal). This is substantiated in the following:

Proposition 2.5. *If* A, F *are both* Π_1 *and* T *is an* ω*-consistent extension of* PA, *then the arguments* (III), (V) *and* (VII) *are valid.*

Proof. If $A, F(x) \in \Pi_1$, then $F(\#A)$ and $A \leftrightarrow F(\#A)$ are both Σ_2. By Isaacson (2011, Theorem 17), all the T-provable, Σ_2-sentences are true. So, the provability

of $F(\#A)$ or $A \leftrightarrow F(\#A)$ in T implies their truth. Whence (III), (V) and (VII) all reduce to (I). □

3. Back to the Truth of the Gödel Sentence

Traditionally, the sentence G is called *the* Gödel sentence of the theory in consideration (say, of PA). One probable reason for this is that if G' is any other sentence which is equivalent to *its* unprovability, then G and G' are equivalent (see, e.g., Lindström, 1996).

But this is not a convincing argument when the theory is not sound. To see this, let S be the theory $\mathsf{PA} + \neg\mathsf{Con}(\mathsf{PA})$, where $\mathsf{Con}(\mathsf{PA})$ is the consistency statement of PA, to wit: $\neg\mathsf{Pr}(\#[0 = 1])$. Then S is not sound, but it is consistent by Gödel's second incompleteness theorem. We show that for every true sentence τ, the sentence $\tau' = \mathsf{Con}(S) \wedge \tau$ is equivalent to its unprovability in S; later we show that this holds for every false S-refutable sentence ρ too. Now, take τ to be an arbitrary sentence (which need not be true). We note that $S \vdash \neg\mathsf{Con}(S)$ (because $S \vdash \neg\mathsf{Con}(\mathsf{PA})$ and $\mathsf{PA} \subset S$) and so:

$$S \vdash \mathsf{Con}(S) \wedge \tau \rightarrow \neg\mathsf{Pr}_S(\#[\mathsf{Con}(S) \wedge \tau]).$$

Also, $S \vdash \mathsf{Pr}_S(\#\theta)$ for any θ; in particular, $S \vdash \mathsf{Pr}_S(\#[\mathsf{Con}(S) \wedge \tau])$, whence:

$$S \vdash \neg\mathsf{Pr}_S(\#[\mathsf{Con}(S) \wedge \tau]) \rightarrow \mathsf{Con}(S) \wedge \tau,$$

which shows that $\tau' = \mathsf{Con}(S) \wedge \tau$ is equivalent to its unprovability in S:

$$S \vdash \tau' \leftrightarrow \neg\mathsf{Pr}_S(\#\tau').$$

Now, take ρ to be a false S-refutable sentence (i.e., any ρ with $\mathbb{N} \nvDash \rho$ and $S \vdash \neg\rho$); for example, any false Π_1-sentence is S-refutable by the Σ_1-completeness of S. Then by $S \vdash \mathsf{Pr}_S(\#\rho)$ we have:

$$S \vdash \rho \leftrightarrow \neg\mathsf{Pr}_S(\#\rho),$$

which shows that ρ is equivalent to its unprovability in S.

It should be noted that no true sentence of the form $\mathsf{Con}(S) \wedge \tau$ is equivalent to any false S-refutable sentence, even though they are all provably equivalent inside S.

The moral is that mere equivalence (inside the theory) to its own unprovability does not make a sentence worthy of the title "the Gödel sentence", simply because there could be *more than one* such sentences, even up to equivalence. We submit the following as an improvement upon the usual convention:

Definition 3.1. By the *Gödel sentence* of a recursively enumerable theory T we mean any sentence P satisfying the following properties:

$$\text{(i) } T \vdash P \leftrightarrow \neg \mathsf{Pr}_T(\#P) \text{ and (ii) } \mathbb{N} \vDash P \leftrightarrow \neg \mathsf{Pr}_T(\#P).$$

(As mentioned above, all such sentences are equivalent–in $\mathbb{N}$, and provably in T whenever $T \supseteq \mathsf{PA}$.)

Following the advice of an anonymous referee for *Theoria*, in the Appendix we present an alternative definition which works for theories extending a sound 'base' theory strong enough to prove the uniqueness of the Gödel sentence of those theories.

Remark 3.2. Apart from worries about unsound theories considered at the beginning of this section, we admit that there is some infelicity in calling each member of a class of sentences "the Gödel sentence" of a theory: after all, each sentence is a specific syntactic object. For want of a better term, one may think of calling the totality of that class of mutually equivalent sentences the *Gödel proposition* of the theory, where "proposition" is understood to denote what is said by each and every one of those sentences.[5]

We can now show that the Gödel sentence/sentences of a theory, as defined by 3.1, is/are true if and only if the theory is consistent (cf. Isaacson, 2011, Theorems 10, 11]):

Theorem 3.3. *Let P be a Gödel sentence of a recursively enumerable, Σ_1-complete theory T. Then P is true if and only if T is consistent.*

Proof. If P is true then, by Definition 3.1(ii), so is $\neg \mathsf{Pr}_T(\#P)$, hence T is consistent. Conversely, if T is consistent then it cannot prove P (for if P were T-provable, then on the one hand by Definition 3.1(i) T would prove $\neg \mathsf{Pr}_T(\#P)$, and on the other hand $\mathsf{Pr}_T(\#P)$ would be a true Σ_1-sentence, hence provable in T, which would contradict the consistency of T). Thus $\neg \mathsf{Pr}_T(\#P)$ is true, whence, by Definition 3.1(ii), P is true. □

5 As the same anonymous referee puts it, in the main part of this article we are using the term "the Gödel sentence" *par abus de langage* (such abuse of the language is of course abundant in the literature on the axiomatic incompleteness phenomenon). However, the connoisseur of the philosophy of language will testify that the very concept of *proposition* has its own problems, too.

It has been argued in the literature (see Boolos, 1990, or Raatikainen, 2005, and references therein) that since the consistency of the theory implies (in fact: *is equivalent to*) its Gödel sentence(s), even provably so inside the theory, for the theory to be able to 'see' the truth of its Gödel sentence(s) it is required that the consistency of the theory should be seen by the theory. Let us note, *en passant*, that a sentence which *is equivalent to its own provability inside the theory* is not thereby (equivalent to) the Gödel sentence of the theory: at the beginning of Section 3 we had $S \vdash \mathsf{Con}(S) \leftrightarrow \rho$ for every false Π_1-sentence ρ (noting that since false Π_1-sentences are refutable in Σ_1-complete theories, we have $S \vdash \neg\rho$ and so by $S \vdash \mathsf{Pr}_S(\#\rho)$ we also have that $S \vdash \rho \leftrightarrow \neg\mathsf{Pr}_S(\#\rho)$), but $\mathsf{Con}(S)$ is a true Π_1-sentence while ρ is not.

Our last result provides a necessary and sufficient condition for the truth of all the Π_1-sentences that are equivalent to their unprovability inside the theory.

Theorem 3.4. *For a recursively axiomatizable extension T of* PA*, all of the Π_1-sentences θ which satisfy $T \vdash \theta \leftrightarrow \neg\mathsf{Pr}_T(\#\theta)$ are true if and only if $T + \mathsf{Con}(T)$ is consistent.*

Proof. If $T + \mathsf{Con}(T)$ is not consistent then $T \vdash \neg\mathsf{Con}(T)$, and so $T \vdash \mathsf{Pr}(\#\psi)$ for every ψ. Also, for every false Π_1-sentence ρ we have $T \vdash \neg\rho$. Whence, $T \vdash \rho \leftrightarrow \neg\mathsf{Pr}_T(\#\rho)$ holds, which shows that every false Π_1-sentence is equivalent to its unprovability in T. Suppose now that a false Π_1-sentence θ satisfies $T \vdash \theta \leftrightarrow \neg\mathsf{Pr}_T(\#\theta)$. Then $\neg\theta$ is a true Σ_1-sentence, whence $T \vdash \neg\theta$. This, by $T \vdash \theta \leftrightarrow \mathsf{Con}(T)$, implies that $T \vdash \neg\mathsf{Con}(T)$, and so the theory $T + \mathsf{Con}(T)$ is not consistent. □

The consistency of $T + \mathsf{Con}(T)$ is a strictly stronger condition than the simple consistency of T (see Isaacson, 2011, Corollary 37), a condition satisfied by all ω-consistent (or even Σ_1-sound) theories (see Isaacson, 2011, Theorem 36). Thus, if a Π_1-sentence ρ says, inside a consistent theory, that it is not provable in that theory, this, in itself, is no reason for believing that ρ is true, unless the theory is also consistent with its own consistency statement. Finally, let us note that the consistency of $T + \mathsf{Con}(T)$ is a necessary and sufficient condition for the independence (unprovability and unrefutability) of the Gödel sentences from the theory (see Isaacson, 2011, Theorem 35).

4. Conclusion

Gödel's original argument was this: the sentence G says about itself that it is not provable in T, and G is indeed not provable in T; therefore G is true. A necessary condition for the validity of this argument is the consistency of T (see Theorem 3.3; note that here G need not be a Π_1-sentence). If it is only inside the

theory that G says that it is not provable in T (i.e., if $T \vdash G \leftrightarrow \neg \mathsf{Pr}_T(\#G)$ but $\mathbb{N} \not\models G \leftrightarrow \neg \mathsf{Pr}_T(\#G)$), then G need not be true (see the remark before Definition 3.1), for T might not be sound, so that T might tell lies about G or might be mistaken about what G says. So, Gödel's argument is invalid in this case, even if G is a Π_1-sentence. But it is valid if G is a Π_1-sentence and if T is ω-consistent (as Gödel wanted to be–see Theorem 3.4 and the explanations after its proof). Over-generalizing Gödel's informal argument makes it invalid, whether we take "saying" to be inside the theory or 'in the real world' ($\mathbb{N}$), and whether we take "holding" to mean being true (in $\mathbb{N}$) or to mean provable (in an ω-consistent, recursively axiomatizable extension of Peano Arithmetic)–see Theorems 2.1, 2.2, 2.3, and 2.4. However, the argument remains valid for Π_1-sentences and Π_1-properties, as they were in Gödel's case (Gödel, 1931)–see Proposition 2.5 above.

Acknowledgements

The authors are grateful to Alasdair Urquhart for discussion and encouragement, and they warmly thank two anonymous referees of this journal for their comments and suggestions which made the article more elegant and much smoother. Saeed Salehi was supported by the office of the vice chancellor for research and technology, University of Tabriz, Iran.

References

Boolos, G. (1990) "On "Seeing" the Truth of the Gödel Sentence." *Behavioral and Brain Sciences* 13(4): 655–656 Also repr. in R. Jeffrey (ed.), *Logic, Logic and Logic* (1999), pp. 389–391. Cambridge, MA: Harvard University Press.

Gödel, K. (1931) "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I." *Monatshefte für Mathematik und Physik* 38(1): 173–198. Translated as "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems, I.", in S. Feferman et al. (eds), *Kurt Gödel Collected Works, Volume I: Publications 1929–1936* (1986), pp. 135–152. Oxford: Oxford University Press.

Halbach, V. and Visser, A. (2014) "Self-reference in Arithmetic, I." *The Review of Symbolic Logic* 7(4): 671–691. "Self-reference in Arithmetic, II." ibid.: 692–712.

Isaacson, D. (2011) "Necessary and Sufficient Conditions for Undecidability of the Gödel Sentence and its Truth." In D. DeVidi, M. Hallett and P. Clarke (eds), *Logic, Mathematics, Philosophy: Vintage Enthusiasms—Essays in Honour of John L. Bell*, pp. 135–152. Dordrecht: Springer.

Lindström, P. (1996) "Provability Logic—A Short Introduction." *Theoria* 62(1–2): 19–61.

Milne, P. (2007) "On Gödel Sentences and What They Say." *Philosophia Mathematica* 15(2): 193–226.

Raatikainen, P. (2005) "On the Philosophical Relevance of Gödel's Incompleteness Theorems." *Revue Internationale de Philosophie* 59(4): 513–534.

Appendix

Definition 3.1 quenches our thirst for a pretty precise talk about the validity of Gödel's informal argument and our quest for the truth of the Gödel sentences of consistent theories (Theorem 3.3). Yet, arguably, it may not be considered a fully satisfactory definition, for the (Π_1) Rosser sentence R_S of $S = \mathsf{PA} + \neg\mathsf{Con}(\mathsf{PA})$ satisfies both requirements of Definition 3.1 for $U = S + \neg R_S$ (which is consistent by Rosser's theorem). As a matter of fact, this Rosser sentence is U-equivalent to what our suggested definition would announce to be the Gödel sentence of the theory (i.e., the Gödel sentence of the theory U).[6]

Depending on one's philosophical views (or lack thereof) on what the Gödel sentence and Rosser sentence of theories really say, this fact may be considered either as an observation about the equivalence of the Gödel sentence of a theory to the Rosser sentence of a super-theory, or else as showing the inadequacy of Definition 3.1. In this short article we chose not to get involved in such issues; neither do we want to become over-technical here–the more "purely logical" aspect of our topic will be further investigated in a work with a more logico-mathematical flavour. However, *if* the above-mentioned fact turns out to be a defect of our Definition 3.1, here is our attempt at a remedy.

We isolate a 'base theory' **B** such that:

($\boldsymbol{\alpha}$) the theory **B** is sound (i.e., $\mathbb{N} \vDash \mathbf{B}$);
($\boldsymbol{\beta}$) for every recursively enumerable theory T there exists some sentence φ such that $\mathbf{B} \vdash \varphi \leftrightarrow \neg\mathsf{Pr}_T(\#\varphi)$; and
($\boldsymbol{\gamma}$) it has the uniqueness property, in the sense that for every sentences φ and ψ and for every recursively enumerable theory $T \supseteq \mathbf{B}$, if $\mathbf{B} \vdash \varphi \leftrightarrow \neg\mathsf{Pr}_T(\#\varphi)$ and $\mathbf{B} \vdash \psi \leftrightarrow \neg\mathsf{Pr}_T(\#\psi)$ then $\mathbf{B} \vdash \varphi \leftrightarrow \psi$.

Certainly PA can be such a base theory.[7] We can now have:

Definition 4.1. Let T be a recursively enumerable extension of **B**. The sentence P is called a *Gödel proposition* of T when $\mathbf{B} \vdash P \leftrightarrow \neg\mathsf{Pr}_T(\#P)$ holds.

---

6 For this point we are indebted to an anonymous referee of *Theoria*.

7 The weakest theory that is known to satisfy all the three conditions is S^1_2 (or equivalently, $\mathrm{I}\Delta_0 + \Omega_1$); it is currently not known whether the condition ($\boldsymbol{\gamma}$) holds for weaker theories like $\mathrm{I}\Delta_0$, IE_1, Q or R, though it is a proven fact that they all satisfy ($\boldsymbol{\beta}$). Let us also note that while the satisfaction of ($\boldsymbol{\beta}$) and ($\boldsymbol{\gamma}$) by PA are mathematically proven theorems, the satisfaction of ($\boldsymbol{\alpha}$) by it, or by any other theory, is something like an article of faith (or, as the same anonymous referee suggested, a *basic insight*). The structure of $\mathbb{N}$ almost forces us, intuitively, to believe that PA is a sound theory (and so are its sub-theories).

Where the theory under consideration is an extension of the base theory, Definition 4.1 is more general than Definition 3.1, and Theorem 3.3 holds for it too. In most of the textbooks that deal with Gödel's incompleteness theorems, this is how "the Gödel sentence" is constructed for recursively enumerable theories that extend a base theory.