Defending Gödel against Floyd-Putnam's Wittgenstein

In a paper published in the very last year of the last century, Juliet Floyd and Hilary Putnam argued that Wittgenstein's so-called "notorious paragraph" may be understood as containing an idea of great philosophical interest,[1] namely

> (**FPW**) Let *P* be the Gödel sentence of the system *PM*[2]. Then if one assumes that ¬*P* is provable in *PM,* one should give up the translation of *P* by the English sentence '*P* is not provable'.

I think Floyd and Putnam are in fact right in saying that what they attribute to Wittgenstein is of philosophical interest. In this note, however, I will make two specific points against them: (A) with a minimally charitable reading of Gödel's introduction to his 1931 paper,[3] **FPW** does not serve as a successful attack on what might be called the *popular* interpretation of the first incompleteness theorem, an interpretation according to which Gödel has shown that there are

[1] "A Note on Wittgenstein's 'Notorious Paragraph' about the Gödel Theorem" [hereafter **FP**], *The Journal of Philosophy*, 97 (2000): 624-632. See also Timothy Bays's critical note, "On Floyd and Putnam on Wittgenstein on Gödel", ibid., 101 (2004): 197-210, and their reply to Bays, "Bays, Steiner, and Wittgenstein's 'Notorious' Paragraph about the Gödel Theorem" [hereafter **FP2**], ibid., 103 (2006): 101-110.

[2] That is, the system of Whitehead and Russell's *Principia Mathematica*.

One question in the recent debate over Wittgenstein's comments on Gödel is whether or not one may consider systems like Peano Arithmetic on a par with *PM* (see the articles referred to in the previous footnote, as well as Bays's reply to **FP2**, available at http://www3.nd.edu/~tbays/papers/wnp2.pdf). My concerns here are independent of this issue.

[3] "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I", in Solomon Feferman et al., eds., *Kurt Gödel Collected Works, Volume I: Publications 1929-1936* (New York: Oxford University Press, 1986), pp. 145-195.

true but unprovable sentences; (B) since something perfectly analogous to **FPW** can be said concerning the second incompleteness theorem, the fact that Wittgenstein does not express any qualms about the second theorem gives us some evidence that Floyd and Putnam might have been overly charitable to Wittgenstein in ascribing to him a deep insight into Gödel's theorem and a solid knowledge of ω-consistency and related notions. What gives additional force to (B) is that, unlike the case of the first theorem, the second theorem is *not*—or not so obviously—immune to an **FPW**-based attack. Theses (A) and (B) will be defended in sections II and III respectively.

I will start with a clarificatory note (section I); before embarking on that, let me, in conformity with the tradition, quote the notorious paragraph in full.

> I imagine someone asking my advice; he says: "I have constructed a proposition (I will use '*P*' to designate it) in Russell's symbolism, and by means of certain definitions and transformations it can be so interpreted that it says: '*P* is not provable in Russell's system'. Must I not say that this proposition on the one hand is true, and on the other hand is unprovable? For suppose it were false; then it is true that it is provable. And that surely cannot be! And if it is proved, then it is proved that it is not provable. Thus it can only be true, but unprovable."
>
> Just as we ask, " 'provable' in what system?", so we must also ask, " 'true' in what system?" 'True in Russell's system' means, as was said: proved in Russell's system; and 'false in Russell's system' means: the opposite has been proved in Russell's system.—Now what does your "suppose it is false" mean? *In the Russell sense* it means 'suppose the opposite is proved in Russell's system'; *if that is your assumption,* you will now presumably give up the interpretation that it is unprovable. And by 'this interpretation' I understand the translation into this English sentence.—If you assume that the proposition is provable in Russell's system,

> that means it is true *in the Russell sense*, and the interpretation "*P* is not provable" again has to be given up. If you assume that the proposition is true in the Russell sense, *the same* thing follows. Further: if the proposition is supposed to be false in some other than the Russell sense, then it does not contradict this for it to be proved in Russell's system. (What is called "losing" in chess may constitute winning in another game.)[4]

I.

While in his formal and official statement of the incompleteness theorems Gödel does not mention truth, it is very popular indeed—and Gödel's own informal discussion in Section 1 of his paper is at least partly responsible for the popularity of this practice—to interpret the first incompleteness theorem as saying that for every consistent axiomatizable theory *T* which is capable of incorporating a certain small part of arithmetic, there is a sentence *P* which is *true* (i.e., true in the standard model **N**) and is independent of *T*. Unlike many other commentators, Floyd and Putnam read the notorious paragraph not as an attempt at refuting Gödel's *theorem*, but as an argument against this *interpretation* of it. For the sake of discussion, let us grant Floyd and Putnam that Wittgenstein is in fact just criticizing the following Target Argument for the popular reading of the first incompleteness theorem (extracted, with obvious little changes, from the notorious paragraph quoted above):

> (**TA**) Suppose *P* were false; then it is true that it is provable. And that surely cannot be! And if it is proved, then it is proved that it is not provable. Thus it can only be true, but unprovable.

[4] Ludwig Wittgenstein, *Remarks on the Foundations of Mathematics*, G.H. von Wright et al., eds. (Cambridge, Mass.: The MIT Press, revised ed., 1978), I, Appendix III, §8, pp. 118 f.

How is Wittgenstein's criticism of **TA** supposed to go through? Neither in their original article nor in its sequel, Floyd and Putnam are ideally clear here—it is not crystal clear how exactly is **FPW** related to the popular reading of the first incompleteness theorem. It seems to me that what **FP**'s Wittgenstein had in mind is something along with the following argument. (If you already find **FP(2)**'s presentation clear, you may skip to the next section.)

> (1) When talking about *P* being *true* or *false*, your use of 'true' and 'false' should be qualified by means of your answer to the question, 'True/false in what system?'
>
> (2) The only relevant answer is: 'In *T*.'[5] *True* means *T*-provable, *false* means *T*-refutable.
>
> (3) **FPW**. That is to say, if you assume that $\neg P$ is provable in *T*, then you have to give up the interpretation of *P* as saying that *P* is not provable.
>
> (4) But then you cannot say that the sentence *P* is unprovable *and true*. The reason is that, although *P* is in fact unprovable if *T* is consistent, *P* can no longer be understood as expressing its own unprovability. In other words, the reductio assumption (of **TA**) alters the meaning of *P,* so that when you assume that *P* is false [i.e., refutable in *T*], *P* can no longer be interpreted as talking about its own *provability*.
>
> Therefore,
>
> (5) The popular reading of Gödel's theorem is untenable (insofar as it is supposed to be supported by **TA**).

[5] Here *T* is the theory (or "system") which is subject to the first incompleteness theorem, *e.g.,* the system of *PM*. Once again, I do not intend to enquire into the following question: So far as Wittgensteinian considerations are concerned, are we allowed to replace *PM* with Peano Arithmetic (or with Robinson's *Q*, etc.)?

As for (1) and (2), I take **FP**'s word for it that they are justified if we adopt Wittgenstein's way of looking at the project of *Principia Mathematica* (see **FP**, pp. 629-631). Concerning (3), Floyd and Putnam do a superb job in justifying it. Their argument for (3) goes like this. If *P* is refutable in *T*, then, as Gödel proves in his paper, *T* is ω-inconsistent, an easy consequence of which being that no model of *T* is isomorphic to the standard model **N**. By Gödel's construction, the sentence *P* is, in the eye of *T*, equivalent to $\neg\exists x\text{PROOF}(x, \#P)$, where PROOF represents, in *T*, the effectively computable relation Proof, a relation which holds between two natural numbers *a* and *b* iff *a* is the Gödel number of a *T*-proof of a sentence whose Gödel number is *b*. Now if it turns out that *T* has no standard model (*e.g.,* because of its ω-inconsistency), we can no longer interpret PROOF in standard natural numbers, hence it is not clear at all that PROOF is saying anything about the relation Proof. More specifically, if *m* is a non-standard number in the universe of a model **M** of *T*, we have no reason to think of $\mathbf{M} \vDash \text{PROOF}[m,\#P]$ as saying anything about the existence of a *T*-proof of *P*.

I think (1)-(5) is a lovely argument. However, and despite my dislike of Gödel's remarks in his introduction, I think the (1)-(5) argument does not work against **TA**, at least not if we augment **TA** with its obvious missing premise. This is the subject matter of my next section.

II.

FP's Wittgenstein holds that **TA** is not a good argument, for if one assumes that ¬*P* is provable (and, for **FP**'s Wittgenstein, this is what it means for *P* to be *false*), one will then "presumably give up the interpretation" that *P* is expressing its own unprovability. The reason given by **FP**'s Wittgenstein, to recapitulate, is founded on the well-known fact that if ¬*P* is provable then *T* is ω-inconsistent. Now, informative as these observations are, I do not think that they defuse **TA**. Or, to be more precise, I think they do not defuse **TA** *in the context of Gödel's paper*. Let me explain.

As made explicit by Gödel (op. cit., p. 147), the argument sketched in the introduction to his 1931 paper was "without any claim to complete precision". It is, therefore, no surprise that there is no mention of the rather unintuitive (and perhaps ad-hoc looking) notion of ω-consistency throughout his introductory Section 1 (pp. 145-151). Surely Gödel knew it as well as anyone that the simple consistency of *T* does not imply the irrefutability of *P*,[6] and in the technical part of his paper he defined the notion of ω-consistency and employed it to prove the first incompleteness theorem.[7] That being so, charity requires us to believe that, even in the informal and introductory part of his paper, the technical assumption of ω-consistency is at work. But then the (1)-(5) argument is blocked, for the assumption of ω-consistency makes **FPW**, as it occurs in that argument, vacuous.

The assumption of ω-consistency, which is explicitly included in the statement of the first incompleteness theorem (Theorem VI in Gödel op. cit, p. 181), guarantees that ¬*P* is *not* provable; insofar as the arguments adduced by Floyd and Putnam are concerned, it also

[6] Shortly after the publication of Gödel's paper, J.B. Rosser presented *another* self-referential sentence whose irrefutability depends on the simple consistency of *T*. It remains true, however, that the simple consistency of *T* does not guarantee the unprovability of ¬*P*. See also the next footnote.

[7] Gödel op. cit., pp. 173 ff. At least since the 1950s it is known that the full power of ω-consistency is not needed to prove the first incompleteness theorem. Kreisel weakened the assumption of ω-consistency to 1-consistency, and even this is more than what is needed. For a comprehensive survey of this and relevant notions, see Daniel Isaacson, "Necessary and Sufficient Conditions for Undecidability of the Gödel Sentence and Its Truth", in David DeVidi et al., eds., *Logic, Mathematics, Philosophy, Vintage Enthusiasms: Essays in Honour of John L. Bell* (New York: Springer, 2011), pp. 135-152.

deprives Wittgenstein of any reason to think that “the ‘translation’ of *P* as ‘*P* is not provable in PM’ is untenable” (**FP**, p. 625). As it stands, the (1)-(5) argument is simply irrelevant to Gödel’s informal argument, given that Gödel was fully aware of the requirement of ω-consistency.

I find it unfortunate that, in his introduction, Gödel talks about the *truth* of the Gödel sentence of a system, and I am in agreement with Alister Watson when he says that Gödel’s informal discussion of the theorem “obscures, rather than illuminates the point”.[8] However, if what I have said here is correct, Gödel’s remarks are not susceptible to the worries that Floyd and Putnam put in Wittgenstein’s mouth.

III.

Some years ago, Georg Kreisel reported that in the early 1940s Wittgenstein had told him that, “having been put off by the introduction,” he had never read Gödel’s proof.[9] This would be of some interest if one’s question were whether Wittgenstein, the man, had understood Gödel’s theorems, given that the notorious paragraph was written in 1937. That, however, is not my concern here. The point I am going to make is the following hypothetical: *if* someone of Wittgenstein’s calibre reads Gödel’s paper and makes the observations **FP** claims he made, *then* it is rather odd that he does not make similar observations concerning the second incompleteness theorem. In fact, he could have made a stronger case against (the prose of) the second theorem. Here is why.

[8] “Mathematics and Its Foundations,” *Mind*, 47 (1938): 440-451, at p. 446, quoted by **FP**, op. cit, at p. 628n10.

[9] “Second Thoughts Around Some of Gödel’s Writings: A Non-academic Option”, *Synthese*, 114 (1998): 99-160. For Floyd’s scepticism concerning this report, see her “Proof *versus* Prose: Wittgenstein on Gödel, Tarski and Truth”, *Philosophia Mathematica* (3), 9 (2001): 280-307, at p. 285n14.

Already in his introductory Section 1 we hear Gödel saying that the precise analysis of the informal argument he just gave "leads to surprising results concerning consistency proofs for formal systems", and that these results "will be discussed in more detail in Section 4 (Theorem XI)". Theorem XI is what is now called *Gödel's second incompleteness theorem*, formulated thus in Gödel's paper:[10]

> *Let* κ *be any recursive consistent class of* FORMULAS; *then the* SENTENTIAL FORMULA *stating that* κ *is consistent is not* κ-PROVABLE; in particular, the consistency of *P* is not provable in *P*, provided *P* is consistent (in the opposite case, of course, every proposition is provable [in *P*]).

Gödel then goes on to sketch the proof.

Now the important thing for Wittgenstein (or for **FP**'s Wittgenstein anyway) would be: What is exactly the sentential formula which states that κ is consistent? In Gödel's notation, this is the formula Wid(κ), which is defined in Gödel's footnote 63 to be (E*x*)(Form(*x*) & $\overline{\mathrm{Bew}}_{\kappa}(x)$). The required definitions are given in items 23 and 46 of Section 2, respectively, and Bew is the exact same predicate about which **FP**'s Wittgenstein said its interpretation as *provable* has to be given up if the system (here κ) turns out to be ω-inconsistent. Now if the translation of Bew as *provable* is doubtful, then of course the translation of Wid as *consistent*

[10] Footnotes omitted, italics and the square brackets are present in the version appeared in the *Collected Works*, op. cit., p. 193. The attentive reader will surely not mistake the reference of Gödel's '*P*' with Wittgenstein's: in Gödel, *P* is the system under question, while Wittgenstein uses '*P*' to denote the Gödel sentence of the system. Also, note that what Gödel calls *recursive* [rekursive] is nowadays called *primitive recursive*; the scope of the first incompleteness theorem as Gödel states it, then, is wider than what a modern-day reader might understand from the sentence.

is no less doubtful. One would then expect to see the same kind of attack on the translation of Wid(κ) as stating that 'κ is consistent'. Wittgenstein, however, disappoints this expectation. And that is not the end.

That is not the end because, although Gödel's assumption of ω-consistency in the official statement of the first theorem makes **FPW** vacuous, no such assumption is being made in the case of the second theorem: ω-consistent or not, κ does not prove Wid(κ) provided that κ is simply consistent. Unlike the first incompleteness theorem, the official formulation of the second incompleteness theorem does admit of ω-inconsistent systems, and **FP**'s Wittgenstein could have said that in those systems one should give up the interpretation of Wid(κ) as the *consistency* of the system.[11]

To summarize, although I am sympathetic to the insightful claim of **FPW**, I have argued for two theses. First, I think **FPW** is not applicable to the first incompleteness theorem because the assumption of ω-consistency is present there. Secondly, I have argued that **FPW** *is* applicable to the second incompleteness theorem, so that one must have reservations in

[11] Throughout my argument, I overlooked Timothy Bays's case against **FPW** (see footnotes 1 and 3 above), to the effect that in the apocalyptic aftermath of a discovery that Peano Arithmetic (or another axiomatizable extension of Robinson's *Q*) is ω-inconsistent, it is more probable that mathematicians will give up the ω-inconsistent system than giving up the familiar translation of *P*. For the purposes of my argument here, I need not make my mind about Bays's thesis. The second incompleteness theorem is supposed to say something quite general about a wide array of systems, viz. they cannot prove their own consistency if they are in fact consistent. Even if we intentionally make up an ω-*in*consistent theory—presumably for studying something other than the standard model **N**—we still want to say that *it* cannot prove its consistency if it is consistent. But now the **FPW** insight challenges the interpretation of Wid(κ) (or, Con(κ) in the modern jargon) as the *consistency* of κ.

saying that the theorem says something about the unprovability of the *consistency*, and this I find of genuine philosophical importance—though the fact that Wittgenstein did not comment on it diminishes our confidence in his understanding of Gödel's proof.