JPhil CI.4 (April 2004): 197–210

On Floyd and Putnam on Wittgenstein on Gödel

Timothy Bays

In a recent discussion piece,[1] Juliet Floyd and Hilary Putnam present a new analysis of Wittgenstein's "notorious paragraph" on Gödel's first incompleteness theorem. Textually, they claim that Wittgenstein's remarks have been widely misunderstood, and they argue that Wittgenstein had a better understanding of Gödel's theorem than he has often been credited with. Substantively, they find in Wittgenstein's remarks "a philosophical claim of great interest," and they argue that, when this claim is properly assessed, it helps to vindicate some of Wittgenstein's broader views on Gödel's theorem.

Here, I address the second of these two arguments (while disclaiming the scholarly credentials required to assess the first, the purely textual, one).[2] I begin by examining the central claim which Floyd and Putnam attribute to Wittgenstein, showing that their argument for this claim is inadequate and that the claim itself is almost certainly false. I then argue that, even if Wittgenstein's central claim *were* true, it would not lead to the conclusions Floyd and Putnam think it does. At the end of the day, I conclude that Floyd and Putnam provide no new insights into Gödel's theorem by way of their reading of Wittgenstein.

1 On Floyd and Putnam

Let's begin with the relevant passage from Wittgenstein.[3]

> 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 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 can 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 been proved in Russell's system"; *if that is your assumption* you will now presumably give up the interpretation that

[1] Juliet Floyd and Hilary Putnam, "A Note on Wittgenstein's "Notorious Paragraph" about the Gödel theorem," *The Journal of Philosophy* 97 (2000): 624–632. Hereafter WNP.

[2] For more on the textual questions, see Juliet Floyd, "On Saying What You Really Want To Say: Wittgenstein, Gödel, and the Trisection of the Angle," in *From Dedekind to Gödel: Essays on the Development of the Foundations of Mathematics*, ed. Jaakko Hintikka (Boston: Kluwer, 1995); Juliet Floyd, "Prose versus Proof: Wittgenstein on Gödel, Tarski and Truth," *Philosophia Mathematica* 9 (2001): 280–307; and Mark Steiner, "Wittgenstein as his Own Worst Enemy: The Case of Gödel's Theorem," *Philosophia Mathematica* 9 (2001): 257–279.

[3] Ludwig Wittgenstein, *Remarks on the Foundations of Mathematics*, ed. G.E. von Wright, R. Rees and G.E.M. Anscombe and trans. G.E.M. Anscombe (Cambridge: MIT, 1956): I, Appendix III, §8.

> 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 willing in another game.)

In this passage, we find Wittgenstein criticizing a relatively common interpretation of Gödel's first theorem: that the theorem shows—or helps to show—that there are true but unprovable sentences of ordinary number theory.[4] Wittgenstein objects to this interpretation, partially because he is skeptical concerning the notion of "Truth" in play here ("'True' in what system?"), and partially because he is opposed in principle to the derivation of "philosophical" claims from "mathematical" arguments.[5]

In their analysis of this passage, Floyd and Putnam focus on the following key claim (paraphrased from the second paragraph of Wittgenstein's discussion):

> **KC:** If one assumes that $\neg P$ is provable, then one should give up the "translation" of P by the English sentence 'P is not provable.'

In what follows, I'll start by sketching the basic mathematics behind this claim. I'll then examine Floyd and Putnam's argument for the claim and what they take to follow from it. Finally, in sections 2 and 3, I'll explain how and why their analysis goes wrong.

I begin with the relevant mathematics. In his original paper on incompleteness,[6] Gödel defines two formulas in the language of formal number theory. These formulas—which I will call "**Proof**(x, y)" and "**Subst**(x, y, z)"—have the following nice property:[7]

> **Numeralwise Representability:** Let m be the code of a sentence, ϕ, and let n be a natural number:
>
> 1. If n is the code of a proof of ϕ, then PA $\vdash$ **Proof**$[\hat{n}, \hat{m}]$.
> 2. If n is not the code of a proof of ϕ, then PA $\vdash \neg$**Proof**$[\hat{n}, \hat{m}]$.
>
> Similarly, let $\phi(v_0)$ be a formula with one free variable, let p be the code of ϕ, and let n and m be natural numbers:

[4] Wittgenstein focuses on number theory as formulated in "Russell's system"—i.e., the system of *Principia Mathematica*. There is, however, nothing in his argument which depends on this particular choice of background logic. For expository convenience, I will recast the argument in terms of ordinary, first-order Peano Arithmetic. Later—in section 4—I will discuss the possible philosophical significance of formulating the argument in "Russell's system."

[5] See, e.g., *Remarks* VII §19. See also *Philosophical Investigations* §124.

[6] Kurt Gödel, "On formally undecidable propositions of *Principia Mathematica* and related systems," in *Collected Works* (New York: Oxford University Press, 1986): 145–199. Hereafter FUP.

[7] To unpack the terminology/notation here, I note that "coding" is simply a way of associating natural numbers to syntactic objects like formulas and proofs; done properly, it allows syntactic properties like "being well-formed" or "being a valid proof" to be treated as number-theoretic properties of the associated codes. Similarly, the notation $\hat{n}$ simply represents the expression for the number n in the language of formal number theory: specifically, $0 + 1 + \ldots + 1$, with n 1's.

1′. If m is the code of $\phi(\hat{n})$, then PA $\vdash$ **Subst**$[\hat{p}, \hat{n}, \hat{m}]$.

2′. If m is not the code of $\phi(\hat{n})$, then PA $\vdash \neg$**Subst**$[\hat{p}, \hat{n}, \hat{m}]$.

Using these formulas, Gödel defines the sentence which Wittgenstein later calls "P" (although Gödel himself calls it something different). Initially, he defines the formula:

$$\psi(v_0) = \exists y\,[\mathbf{Subst}(v_0, v_0, y)\ \&\ \neg\exists z\,\mathbf{Proof}(z, y)]$$

Letting e_0 be the code for $\psi(v_0)$, Gödel sets:

$$P = \psi(\hat{e}_0) = \exists y\,[\mathbf{Subst}(\hat{e}_0, \hat{e}_0, y)\ \&\ \neg\exists z\,\mathbf{Proof}(z, y)].$$

This, then, gives the basic definition underlying Wittgenstein's (and later Floyd and Putnam's) discussion. There are two things we should notice about this definition.

First, the property of numeralwise representability explains why someone might think that P should be interpreted as saying "P is unprovable." On the assumption that PA is sound—i.e., that $\mathbb{N} \models$ PA—the numeralwise representability of **Proof**(x, y) and **Subst**(x, y, z) entails the following, somewhat more semantic, property:[8]

> **Arithmetic Expressibility.** For any two numbers n and m, $\mathbb{N} \models$ **Proof**$(\hat{n}, \hat{m})$ if and only if m is the code of a sentence, ϕ, and n is the code of a proof of ϕ.
>
> Similarly, for any three numbers n, m and r, $\mathbb{N} \models$ **Subst**$(\hat{n}, \hat{m}, \hat{r})$ if and only if n is the code of a formula, $\phi(v_0)$, and r is the code of the sentence $\phi(\hat{m})$.

Using arithmetic expressibility, therefore, we can make perfect sense of the claims that $\neg\exists z\,\mathbf{Proof}(z, y)$ "says that" the sentence coded by y is unprovable and that **Subst**$(\hat{e}_0, \hat{e}_0, y)$ "says that" y is the code of P. Together, these two claims provide the intuitive basis for interpreting P as "P is unprovable."

Second, our definition explains why someone might think that P is a true but unprovable sentence of number theory. If we filter P though arithmetic expressibility, we get the following result:

$\mathbb{N} \models P$	if and only if	there is an n, such that $\mathbb{N} \models$ **Subst**$(\hat{e}_0, \hat{e}_0, \hat{n})$ and $\mathbb{N} \models \neg\exists z\,\mathbf{Proof}(z, \hat{n})$.
	if and only if	there is an n, such that n is the code of P and $\mathbb{N} \models \neg\exists z\,\mathbf{Proof}(z, \hat{n})$.
	if and only if	there is no m, such that m is the code of a proof of P.
	if and only if	PA $\nvdash P$.

This equivalence leaves us with two options: **either** $\mathbb{N} \models P$ and PA $\nvdash P$ **or** $\mathbb{N} \not\models P$ and PA $\vdash P$. Notice, however, that the second of these two options contradicts the soundness of PA (since it trivially entails that $\mathbb{N} \not\models$ PA). Hence, we are forced to accept the first option. And this first option looks an awful lot like the claim that P is true but unprovable.

[8] It is important to note that these facts about arithmetic expressibility can be proved directly, without passing through numeralwise representability. (Gödel himself approaches arithmetic expressibility this way in his 1931 paper.) Nonetheless, once we have numeralwise representability, it provides a quick route to arithmetic expressibility.

This, therefore, gives us a sketch of the claims Floyd and Putnam want to argue against (that P says "P is unprovable" and that P itself is "true but unprovable"). To understand their objections to these claims, we can begin by looking at their argument for KC.[9] So suppose, just for the sake of argument, that PA is consistent but that we have discovered that PA $\vdash \neg P$. The first thing to notice is that this entails that $\mathbb{N} \not\models$ PA (i.e., since $\mathbb{N} \models \text{PA} \Longrightarrow \mathbb{N} \models \neg P \Longrightarrow \mathbb{N} \not\models P \Longrightarrow \text{PA} \vdash P \Longrightarrow$ PA is inconsistent). It follows, therefore, that PA is satisfied *only* by non-standard models of number theory—i.e., by models which are not isomorphic to the natural numbers. To use the technical jargon, it follows that PA is *ω-inconsistent.*[10]

Consider, then, a specific model of PA—call it $\mathbb{M}$. Since PA $\vdash \neg P$, we know that $\mathbb{M} \models \exists x\, \mathbf{Proof}(x, \widehat{\ulcorner P \urcorner})$. Therefore, there is some element $m \in \mathbb{M}$ such that $\mathbb{M} \models \mathbf{Proof}(m, \widehat{\ulcorner P \urcorner})$. But, by numeralwise expressibility, $\mathbb{M} \models \neg\mathbf{Proof}(\hat{n}, \widehat{\ulcorner P \urcorner})$ for each natural number n. Hence, the relevant m is not one of the ordinary natural numbers—it is, of necessity, one of the "non-standard" elements of $\mathbb{M}$.[11] Further, and this is the second thing to notice here, there is no interesting sense in which this non-standard m "codes up" a proof of P—or of any other formula, for that matter. Given this, we have no reason to think that the formula $\mathbf{Proof}(x, y)$, *as it gets interpreted by the model* $\mathbb{M}$, still captures the notion "y is the code of a sentence and x is the code of a proof of that sentence."

In this situation, therefore, Floyd and Putnam argue that it's hard to justify interpreting P as meaning "P is not provable." After all, the intuitive motivation for interpreting P this way involved interpreting P on the natural numbers—i.e., interpreting "+" as plus, "×" as times, and letting "$\exists x$" range over the natural numbers. But now it turns out that this interpretation is incompatible with PA itself (since PA has *only* non-standard models). Similarly, the interpretation of P as "P is not provable" owed something to the interpretation of $\mathbf{Proof}(x, y)$ suggested by arithmetic expressibility. But, as we have just seen, this interpretation of $\mathbf{Proof}(x, y)$ breaks down when we interpret $\mathbf{Proof}(x, y)$ on non-standard models. Hence, once we make the assumption that PA $\vdash \neg P$—and thus that *there are no standard models for PA*—we have every reason to give up our initial interpretations of both $\mathbf{Proof}(x, y)$ and of P.[12]

This, then, is the core of Floyd and Putnam's argument for KC. What a given formula "expresses" depends on the model at which we interpret it. If we interpret $\mathbf{Proof}(x, y)$ and P on the natural numbers, then it is plausible to think that they express the facts that "x codes a proof of y" and that "P is unprovable." If we interpret them on non-standard models, then it is no longer plausible to think that they express such things.

[9] The argument occurs on pp. 625–627 of WNP.

[10] Formally, ω-inconsistency is a bit stronger than the claim that a system has only "non-standard" models. To say that PA is ω-inconsistent means that there is some particular formula, $\phi(x)$, such that for each n, PA $\vdash \neg\phi(\hat{n})$, but it is also the case that PA $\vdash \exists x\, \phi(x)$. In the specific case where PA $\vdash \neg P$, we can actually identify the relevant $\phi(x)$: for each n, PA $\vdash \neg\mathbf{Proof}(\hat{n}, \widehat{\ulcorner P \urcorner})$, but PA $\vdash \exists x\ \mathbf{Proof}(x, \widehat{\ulcorner P \urcorner})$. So, in this case, it is the very formula used in the construction of P which serves to witness the ω-inconsistency of PA.

[11] Cf. the discussion of ω-inconsistency in the preceding footnote.

[12] In general, there may also be problems concerning the interpretation of **Subst**(x, y, z). In particular, there will be non-standard elements $m_1, m_2, m_3 \in \mathbb{M}$ such that $\mathbb{M} \models \mathbf{Subst}(m_1, m_2, m_3)$, even though it's clear that m_2 does not code an ordinary natural number and that neither m_1 nor m_3 code formulas. Fortunately, this doesn't have to effect the interpretation of P, since we *can* show that $\mathbb{M}$ interprets $\mathbf{Subst}(\hat{e}_0, \hat{e}_0, z)$ correctly (i.e., $\mathbb{M} \models \mathbf{Subst}(\hat{e}_0, \hat{e}_0, m) \Longleftrightarrow \mathbb{M} \models m = \widehat{\ulcorner P \urcorner}$).

Therefore, once we assume that PA is ω-inconsistent—and hence that *all* "admissible interpretations" take place on non-standard models—we are left with *no* good reasons for interpreting P as "P is not provable."

So much for KC itself. With KC in hand, I turn to the consequences Floyd and Putnam draw from KC. To begin, KC shows that the *formal structure* of P does not force us to interpret P as "P is not provable" (since this is an interpretation we would *give up* under certain circumstances—e.g., if we discovered that PA $\vdash \neg P$). Nor, Floyd and Putnam argue, do the other details of Gödel's *mathematics* force us to interpret P this way. As formulated in his original paper, Gödel's proof is purely syntactic (using numeralwise representability to show that if PA is ω-consistent, then PA $\nvdash P$).[13] Hence, as far as the *mathematics* goes, we can dispense with *interpretation* altogether.[14] Given all this, it's unclear where the claim that P should be interpreted as meaning "P is unprovable" is supposed to *come from*. At best, this seems to be a "metaphysical claim" (WNP, 632) which gets grafted onto Gödel's mathematics; it isn't, in any interesting sense, something which *follows from* that mathematics.

At the end of the day, Floyd and Putnam take themselves to have vindicated Wittgenstein's skepticism about the common insistence that Gödel's incompleteness theorem shows that there are true but unprovable sentences of ordinary number theory. To make this claim plausible, we seem to need an interpretation of P which makes P *say* "P is not provable." But KC—along with an analysis of the actual *mathematics* of Gödel's proof—undercuts the idea that Gödel's work provides such an interpretation. As a result, Floyd and Putnam conclude that the common insistence that P is "true but unprovable" is more a "metaphysical claim" than a "mathematical result" (WNP, 632). And, while this doesn't entail that the claim is *false,* it does show "how little sense we have succeeded in giving it" (WNP, 632).

2 On KC

In this section, I step back to examine Floyd and Putnam's argument for KC in more detail. (In section 3, I'll look at their more general conclusions concerning the incompleteness theorem.) To see what's wrong with Floyd and Putnam's argument for KC, we need to back up a bit and recall the hypothetical situation KC

[13] As Mic Detlefsen has pointed out to me, the claim that Gödel's proof is "purely syntactic" needs to be treated with some care. On the one hand, Gödel's proof does not require specifying a full, formal interpretation of the symbols of his language. (Gödel himself emphasizes this fact on pp. 171, 177 and 181 of FUP.) Nor does it involve a notion of *truth* for that language. So, it isn't semantic in the way that a model-theoretic argument is semantic (or one which made use of arithmetic expressibility).

On the other hand, the argument does involve a systematic *association* between natural numbers and terms in our language (e.g., between n and $\hat{n}$). This association is crucial for proving the numeralwise representability results on which the overall proof depends. So, even Gödel's "syntactic proof" involves general correlations between natural numbers and the formal expressions which, in Gödel's own terms, "denote" those numbers.

[14] At best, Floyd and Putnam suggest, Gödel's paper gives rise to a proof-theoretic conception of truth—one under which "P is true" *means* PA $\vdash P$ and "P is false" *means* PA $\vdash \neg P$. Clearly, however, this is an interpretation which leaves no room for "true but unprovable" sentences of arithmetic.

Now, for the reasons mentioned in the last footnote, I find the claim that Gödel's mathematics doesn't involve "interpretation" somewhat problematic (though there's clearly *a* sense in which it's true). For the sake of argument, though, I will evade these complexities and simply grant the claim that Gödel's proof is "purely syntactic" and that it involves no "interpretation."

envisages. Initially, we come to Gödel's theorem with an interpretation of the language of formal arithmetic on the natural numbers—an interpretation which reads "+" as plus, "×" as times, and which lets "$\exists x$" range over the natural numbers.[15] Given this interpretation, we proceed to assume two things. First, we assume that our background axiomatization of arithmetic is *sound*—i.e., that all of the axioms of PA come out *true* on our interpretation. (If we use standard model-theoretic machinery, this amounts to the claim that $\mathbb{N} \models \text{PA}$.[16]) Second, on the basis of an argument like that given on page 3, we assume that we can interpret P as "P is not provable."

Now at this point, KC suggests the following hypothetical: suppose $\text{PA} \vdash \neg P$. As we saw in section 1, this entails that $\mathbb{N} \not\models PA$. Hence, we face a choice. On the one hand, we could modify our background interpretation of arithmetic, giving up $\mathbb{N}$ as an appropriate model for our language and limiting ourselves to those (non-standard) models which happen to satisfy PA. If we make this choice, then we will be forced to give up the interpretation of P as "P is not provable" (for the reasons highlighted by Floyd and Putnam). On the other hand, we could keep our background *interpretation* of arithmetic language and give up the assumption that PA provides a satisfactory *axiomatization* of arithmetic.

It is clear from their paper, that Floyd and Putnam think we should take the first option (as witnessed, for instance, by their insistence that only models which satisfy PA should count as "admissible interpretations" for our language).[17] But it's equally clear that they provide no real *argument* for preferring this option. Instead, they simply ignore the possibility of *keeping* $\mathbb{N}$ as the canonical interpretation for our language while *abandoning* (or, at the very least, *modifying*) the ω-inconsistent axiomatization which stands in conflict with this interpretation.[18]

So far, this is simply an objection to Floyd and Putnam's *argument* (pointing out a lacuna which could, in principle, be filled in). But there's a deeper problem here: Floyd and Putnam's assumptions about the (hypothetical) response of the mathematical community to the discovery that $\text{PA} \vdash \neg P$ are almost certainly false. Although such a discovery would cause a great deal of consternation, I think the vast majority of mathematicians would look for ways of revising PA in order to block the proof in question—i.e., would try to isolate the specific axioms of PA which are essential to the proof and to eliminate (some of) those principles from our axiomatization.[19] There are three points to make about this.

[15] After Tarski, this interpretation would probably be fleshed out using model-theoretic machinery. But this isn't necessary. Gödel himself sketches a non-Tarskian version of the interpretation, and it's this non-Tarskian interpretation which Gödel uses in formulating the notion of arithmetic expressibility (and in proving things about that notion). See, e.g., pp. 181–191 of FUP.

[16] For the remainder of the paper, I will assume that we *have* used standard model-theoretic machinery in fleshing out our interpretation. Hence, I will use $\mathbb{N} \models \text{PA}$ to express the soundness of our axioms.

[17] The insistence that we limit ourselves to models of PA when we interpret arithmetic runs rather deep in Floyd and Putnam's paper. At one point they even suggest that, were we to find PA *inconsistent,* we should conclude that there are *no* admissible interpretations of arithmetic (see WNP: 626).

[18] This is not, perhaps, so surprising in Putnam's case, as a similar insistence on the priority of axioms over interpretations lies at the heart of his so-called "model-theoretic argument" against realism. See sections 3–4 of Timothy Bays, "On Putnam and his Models," *The Journal of Philosophy* XCVIII (2001): 331–50.

[19] My guess is that mathematicians would initially focus on the uses of induction in the proof. The hope would be that some

First, it is almost *unimaginable* that mathematicians would adopt recognizably non-standard models of arithmetic as canonical for interpreting the language of number theory. Neither would they accept a provably ω-inconsistent axiomatization of arithmetic as adequate (since it wouldn't, after all, describe the natural numbers!). Given this, the straightforward claim that mathematicians would reject the interpretation of P as "P is not provable" *because* they accept non-standard models as the basis for interpreting arithmetic is surely mistaken (though it's equally surely what Floyd and Putnam's argument requires).

Second, although mathematicians wouldn't abandon the interpretation of P as "P is not provable" for the reasons Floyd and Putnam suggest, there is one reason they might abandon it. Suppose it turns out that the specific mathematics used in proving arithmetic expressibility is somehow implicated by the discovery that PA $\vdash \neg P$. (Say, because there is some $\phi \in$ PA such that 1.) ϕ is crucial to the proof of $\neg P$ and 2.) there are numbers n and m such that m codes a formula, and n codes a proof of that formula, but PA $\setminus \{\phi\} \nvdash \mathbf{Proof}(\hat{n}, \hat{m})$.[20]) In such a case, the modifications to PA which are needed to *avoid* the acceptance of non-standard models might undercut various arithmetic expressibility results. In doing so, they might also undercut the traditional interpretation of P.

This, then, looks like a circumstance where something like KC might really be true. Still, there are some obvious worries. First, the supposition in the last paragraph is extremely implausible. The numeralwise representability results on page 2 can be proved in *very* weak fragments of arithmetic, and it's hard to imagine these fragments being problematic.[21] Second, any discovery of problems in weak arithmetic—say, in Q_0—would require such deep revisions of present mathematics that it's virtually impossible to adjudicate questions concerning "what we would/should do" in such circumstances. In particular, neither I nor Floyd and Putnam are in a position to intelligently evaluate KC under this kind of hypothetical. Finally, whatever we would do under this hypothetical, we would do it for reasons other than those Floyd and Putnam suggest. If, for instance, we would abandon the interpretation of P as "P is not provable," then we would do so as part of a larger revision of PA aimed at *keeping* the standard model of arithmetic; we wouldn't accept non-standard models in order to keep the ω-inconsistent PA.

This brings me to my third point: nothing I've just said implies that mathematicians would *lose interest* in PA and its models (i.e., under the circumstances envisioned by KC). Just as logicians now study non-standard models of arithmetic (while acknowledging that these models *are* non-standard), so mathematicians would continue to think about non-standard models after the discovery that PA $\vdash \neg P$ (indeed, they would probably study them a lot more than they do now). Nevertheless, the majority of mathematical work—and

well-motivated restriction of the induction scheme would enable us both to restore ω-consistency and to understand *why* our initial scheme went wrong (e.g., perhaps we allowed induction on some subtly-paradoxical predicate/formula). This seems *far* more likely than the abandonment of $\mathbb{N}$ which Floyd and Putnam urge upon us.

[20] Of course, this particular result would simply block the proof of arithmetic expressibility *via* numeralwise representability. The crucial case—and the case we're really interested in—is where the elimination of ϕ would undercut *every* proof of arithmetical expressibility (perhaps by undercutting the recursive definition of satisfaction). Nevertheless, the case above illustrates the *kind* of problem at issue here.

[21] So, for instance, numeralwise representability holds for systems as weak as Robinson's Q_0 (a fragment of arithmetic which involves *no* induction); it's very hard to imagine Q_0 being ω-inconsistent.

the work that would be universally recognized as *arithmetic*—would continue to involve the study of $\mathbb{N}$: it would involve interpreting the language of arithmetic *on* $\mathbb{N}$, and it would look for axioms satisfied *by* $\mathbb{N}$.[22]

These, then, are some reasons for thinking that Floyd and Putnam's argument for KC is flawed and that KC itself is mistaken. Before moving on, I think it's useful to digress a moment and examine a claim similar to KC outside the mathematical context. Suppose we have a formal axiomatization of some part of physics—call it T and suppose that it's formulated in the language $\mathcal{L}$. Next, suppose we discover that T proves things which conflict with the basic physical phenomena T was supposed to describe (e.g., T makes a series of glaringly false predictions). Then, just as in Floyd and Putnam's case, we face a choice: we can abandon T and start looking for new axioms which better describe the phenomena we're interested in, or we can abandon the physical phenomena and start studying arbitrary models of T (numerical models, perhaps).

Here, I think it's obvious that the physics community would take the first option. T was interesting only (or, at least, primarily) because it seemed to describe a specific class of physical phenomena (i.e., described it via the standard interpretation of $\mathcal{L}$). Once T goes wrong about such phenomena, then T has to be modified. In the scientific case, therefore, preserving the original interpretation of our language turns out to be far more important than preserving our original axiomatization. I see no reason to think that things are different for arithmetic; hence, I see no reason to accept Floyd and Putnam's argument for KC. At the very least, the structural similarities between their argument and the (obviously faulty) physical analog should lead us to be highly suspicious of the former.

3 On Gödel's Theorem

In this section, I turn from Floyd and Putnam's defense of KC to make some brief comments concerning the conclusions they draw from this claim. First, I concede a point: there is nothing in the formal structure of P—i.e., in P's very syntax—which forces us to interpret P as "P is not provable." Nor, as Floyd and Putnam notice, does Gödel's original proof require such an interpretation. So, if these were the only ways of providing a mathematically significant interpretation of P, then Floyd and Putnam would be right in challenging the claim that P means "P is not provable."

Fortunately, these are *not* the only respectable ways of interpreting P. The natural interpretation of the language of arithmetic—the interpretation under which "$+$" means plus, "$\times$" means times, and "$\exists x$" ranges over the natural numbers—can be made perfectly rigorous. At present, it would be most natural to do this using model-theoretic machinery or Tarski's original apparatus of language, metalanguage, meta-metalanguage, etc. But even without this machinery, we can formulate the interpretation in a mathematically

[22] A qualification is in order here. A lot of current work in arithmetic focuses on the purely algebraic properties of the natural numbers (and associated structures like $\mathbb{Z}$ and $\mathbb{Q}$). To the extent that non-standard models of $\mathbb{N}$ (and the associated non-standard versions of $\mathbb{Z}$ and $\mathbb{Q}$) continue to satisfy the relevant algebraic axioms, people working on algebraic number theory may not care about non-standardness. Nevertheless, no matter how interesting these non-standard models are (to algebrists and logicians), the canonical core of arithmetic will continue to be the study of $\mathbb{N}$, and the axioms for arithmetic will have to be axioms satisfied by $\mathbb{N}$.

perspicuous manner. Indeed, Gödel himself uses this interpretation in his 1931 paper, both to formulate the notion of arithmetic expressibility and to prove theorems involving this notion.[23]

Further, this interpretation has closer ties to the incompleteness theorem than Floyd and Putnam's argument would lead one to believe. I've already noticed that Gödel *uses* the interpretation in several parts of his paper, and he seems to *presuppose* it in many of his informal asides.[24] More substantially, the interpretation helps to explain why Gödel's theorem is *interesting* in the first place. The theorem is not interesting because it shows that a randomly selected axiom system happens to be incomplete: there are *lots* of incomplete axiom systems, and most of them are rather boring.[25] Instead, the theorem shows that a standard axiomatization of *arithmetic*—of the theory of $\mathbb{N}$—is incomplete. More significantly, the theorem doesn't just show that *this particular* axiomatization is incomplete. Rather, it shows that arithmetic is *intrinsically* incomplete: no recursive extension of PA (or even of Q_0) provides a complete axiomatization of arithmetic.[26] It is this intrinsic incompleteness of *arithmetic* which makes Gödel's first theorem so interesting.

Finally, and most importantly, although it's certainly true that Gödel's original proof of incompleteness was syntactic, there is an alternate, semantic proof which makes essential use of the interpretation at issue here. This proof starts with the arithmetical expressibility results from page 3 and then uses the argument which immediately follows those results (again on page 3) to show that P is a true but unprovable sentence (and, hence, that P is undecidable from PA).[27] So, even though Gödel didn't use this interpretation in his official proof of incompleteness, the interpretation is still very closely related to the underlying *mathematics* of Gödel's paper.[28]

When all is said and done, then, I think that Floyd and Putnam are mistaken in the conclusions they draw from KC. There is a perfectly good—and a perfectly *mathematically* respectable—interpretation of the language of arithmetic under which P expresses the fact that P is not provable. On this interpretation, Gödel's theorem really does show that P is "true but not provable." Further, this interpretation is in no way foreign to Gödel's work. Gödel uses the interpretation in the paper where he proves his incompleteness

[23] See FUP p. 181 for the definition of arithmetic expressibility. See pp. 183–187 for some theorems which use this notion.

[24] *Ibid.* 149–151.

[25] For example, the systems $\emptyset$, $\{P(c)\}$, and $\{c \neq d\}$ are all trivially, and uninterestingly, incomplete.

[26] So, there is no way to "fix" the incompleteness of PA by adding a few new axioms.

[27] It's worth noting that this semantic proof amounts to a formalization of the informal proof given in the first paragraph of Wittgenstein's remarks (see page 1). Wittgenstein, however, goes on to criticize this proof in his second paragraph. Surprisingly, it's just this criticism which Floyd and Putnam mean to be defending. See Steiner, "Wittgenstein as his Own Worst Enemy" for more on Wittgenstein's rejection of the semantic proof.

[28] Two historical comments are in order. First, Gödel himself was clearly aware of the semantic proof of incompleteness: he sketches it at the beginning of his 1931 paper (p. 149-151), and he outlines it again in a series of lectures given at Institute for Advanced Study in 1934. Second, Gödel had several reasons for avoiding this semantic proof in his 1931 paper. For one thing, the syntactic proof provides crucial "ingredients" for proving the second incompleteness theorem (while the semantic proof does not). For another, Gödel intended his theorem to have relevance to Hilbert's program, but only the syntactic proof of the theorem could have such relevance. Hence, Gödel had a number of purely mathematical reasons for using a syntactic proof in 1931. Given this, Gödel's use of the syntactic proof should not be interpreted lack of *awareness* of the semantic proof (and still less as a *rejection* of that proof).

theorem; the interpretation helps to explain the *significance* of that theorem; and the interpretation can be used—as Gödel well knew—to provide a rigorous proof of that theorem. All of this is plain, simple mathematics; there are no "metaphysical claims" anywhere on the horizon.

4 "Russell's System" and Foundations

Before concluding, I want make two remarks concerning an issue I bypassed near the beginning of this paper: the fact that I have formulated my arguments in terms of PA, while Wittgenstein, Floyd and Putnam formulated theirs in terms of "Russell's system." Now on one level, any reversion back to Russell's system would only make my argument stronger. PA is a far more widely accepted formal system than Russell's system is (or ever was). Hence, just as it's clear that we would modify PA to deal with a discovery that PA is ω-inconsistent (as argued in section 2), it's *even more clear* that we would modify Russell's system to deal with an ω-inconsistency in that context.[29] As a result, the arguments I gave in sections 2 and 3 would be even stronger if KC were reformulated in terms of Russell's system. So, there's nothing "slippery" in moving the arguments to PA for reasons of perspicuousness.

On another level, though, there may seem to be a problem here. Russell wanted his system to provide a *foundation* for mathematics: to provide the framework in which other parts of mathematics are formulated, and to set the standards for mathematical rigor. Given this, it might seem odd to think that we can "step outside" this framework to discuss its semantics (and then *reject* the framework if it doesn't live up to our expectations!). To put the point in Floyd and Putnam's terms: "to confess that this is what one has to do would be to abandon the claim for the *foundational* status of a system such as *Principia Mathematica* entirely" (WNP: 630). On this view, then, it may seem that my entire argument in sections 2 and 3 depends on adopting the wrong *attitude* toward our background axiomatization (on treating it as "just one piece of mathematics among others").[30]

---

[29]There are two technical issues which deserve comment here. First, the notion of ω-inconsistency depends on the version of Russell's system with which we are working. Gödel himself worked with a version of Russell's logic which was "superimposed" on ordinary Peano Arithmetic: in particular, Gödel's formulation proved that "every object is a number." For this formulation of Russell's system, the definition of ω-inconsistency given in footnote 10 works just fine.

If we work with the full system of PM—or, for that matter, with a system like ZFC—then our system will prove that there are objects other than numbers (and we will have to use a predicate or formula to pick out the particular objects which we want to count as "natural numbers"). For such systems, we say that T is ω-inconsistent if there is a formula, $\phi(x)$, such that for each n, $T \vdash \neg\phi(\hat{n})$, but it is also the case that $T \vdash \exists x\,[\mathbf{Number}(x)\ \&\ \phi(x)]$. This is a/the standard way of thinking about ω-inconsistency for systems which talk about more than number theory. (Note that if we simply used the definition from footnote 10, then systems like T would be *trivially* ω-inconsistent.)

Second, the discovery that PM was ω-inconsistent would leave us with several *different* options. We could modify the underlying *logic* of PM; we could change the way we formulate arithmetic *in* PM; or we could leave everything as it is and cease to regard our system as a formulation *of arithmetic* (see the comments in the main text vis-a-vis this last option). The only thing we can't do, I contend, is to leave things as they are *and* to insist that we still have a formulation of *arithmetic.*

[30]For what it's worth, Gödel himself seems to share my "bad attitude" toward Russell's system. For Gödel, Russell's system was simply one of several "related systems" to which his results applied, and he felt free to step back and discuss both the

Although this line may initially be tempting, it should be rejected for two reasons. First, there *can't* be a problem simply with stepping back to study the meta-theory of our axiomatization. For, if that were the problem, then it would indict the study of *proof theory* as much as the study of semantics (and, hence, indict the syntactic argument in Gödel's original paper). Any version of Gödel's theorem will involve some amount of "stepping back" from our foundational system (that's why it's a result in *metamathematics*), and I see no principled reasons for permitting us to step back to make syntactic generalizations, while forbidding us to step back to make semantic generalizations.[31] Therefore, if the foundationalist worry is to be cogent, it has to focus on the claim that meta-theoretic analysis can make us *give up* a foundational system (and not on the very idea of meta-theoretic analysis).

Second, there are clear cases where we *can* be forced to give up a foundational system (e.g., when it proves a contradiction).[32] And even when we don't give up a foundational system—because it's too clear, precise and elegant to give up—we can still reject it *as a foundation for a specific discipline.* If a "foundation for arithmetic" winds up proving that $2 + 2 = 5$, then it isn't a foundation for arithmetic (though it may well be a foundation for something else). Similarly, if a "foundation for analysis" proves that the derivative of $x^3 + 1$ is x^{27} then it isn't a foundation for analysis (though, again, it may be a foundation for something else).[33] It's just a *crazy* view of foundational research which holds that, once a foundational system has been proposed, we can no longer legitimately discuss whether that system is *effective* at founding the subjects we wanted it to found in the first place.

In the end, then, I don't see any problems either with my move from Russell's system to PA or with my willingness to consider modifications of "foundational" axiom systems. In particular, I don't see any "foundationalist" reasons for changing the conclusions I reached in sections 2 and 3. There's a perfectly good—and, indeed, a perfectly canonical—interpretation of arithmetic under which Wittgenstein's P really does say "P is not provable." Given this interpretation, Gödel's theorem helps to show that there are "true but unprovable" sentences of ordinary number theory. Nothing in Wittgenstein remarks—or in Floyd and Putnam's analysis of those remarks—should lead us to think otherwise.

---

syntax and the semantics of each of these systems.

[31] As a point *tu quo*, I would note that Floyd and Putnam themselves engage in purely semantic reasoning about Russell's system. So, for instance, they argue that it would be acceptable to define "truth in PM" as "holding in all models of PM" (WNP, 631). Similarly, they are willing to prove that, if PM $\vdash \neg P$, then *all models of PM* are non-standard (WNP, 625). Given this, it would be unreasonable for them to object *in principle* to the semantic analysis of foundational systems.

[32] The important point, here, is that we can recognize from the *outside* that contradictions are unacceptable. We don't just accept contradictions because our "foundational" system tells us to. Hence, we certainly have *some* ability to step back and engage in critical reflection on (purported) foundations.

[33] And, I would argue, if a "foundation for arithmetic" winds up being ω-inconsistent, then it isn't really a foundation for arithmetic (whatever other virtues it might happen to have).

References

Bays, Timothy. "On Putnam and his Models." *The Journal of Philosophy* XCVIII (2001): 331–50.

Floyd, Juliet. "On Saying What You Really Want To Say: Wittgenstein, Gödel, and the Trisection of the Angle." In *From Dedekind to Gödel: Essays on the Development of the Foundations of Mathematics*, edited by Jaakko Hintikka. Boston: Kluwer, 1995.

Floyd, Juliet. "Prose versus Proof: Wittgenstein on Gödel, Tarski and Truth." *Philosophia Mathematica* 9 (2001): 280–307.

Floyd, Juliet and Hilary Putnam. "A Note on Wittgenstein's "Notorious Paragraph" about the Gödel theorem." *The Journal of Philosophy* 97 (2000): 624–632.

Gödel, Kurt. "On formally undecidable propositions of *Principia Mathematica* and related systems." In *Collected Works*. New York: Oxford University Press, 1986.

Steiner, Mark. "Wittgenstein as his Own Worst Enemy: The Case of Gödel's Theorem." *Philosophia Mathematica* 9 (2001): 257–279.

Wittgenstein, Ludwig. *Remarks on the Foundations of Mathematics*. Edited by G.E. von Wright, R. Rees and G.E.M. Anscombe and translated by G.E.M. Anscombe. Cambridge: MIT, 1956.