Floyd, Putnam, Bays, Steiner, Wittgenstein, Gödel, Etc.

Timothy Bays

In their 2000 paper, "A Note on Wittgenstein's 'Notorious Paragraph' about the Gödel theorem,"[1] Juliet Floyd and Hilary Putnam present a new analysis of Wittgenstein's "notorious" remarks on Gödel's first incompleteness theorem. A substantial portion of this analysis focuses on the following key claim. Let "P" be the "I am not provable" sentence from Gödel's proof of incompleteness,[2] and let "PM" be the logical system of *Principia Mathematica.* Then the key claim is:

> **KC:** If one assumes that ¬P is provable in PM, then one should give up the "translation" of P by the English sentence "P is not provable."

In their paper, Floyd and Putnam discuss the role this claim plays in Wittgenstein's remarks; but they also argue that, whatever role it may play in Wittgenstein's own work, the claim is independently defensible, and it provides genuine insight into the philosophical significance of Gödel's first theorem.

In my 2004 paper, "On Floyd and Putnam on Wittgenstein on Gödel," I challenged Floyd and Putnam's argument for KC (while explicitly disclaiming any attempt to assess the textual issues involved in their reading of Wittgenstein). In their new paper, "Bays, Steiner and Wittgenstein's 'Notorious' Paragraph about the Gödel theorem," Floyd and Putnam respond to my challenge.[3] Here, I want to make some brief remarks concerning their response.

1 Issues of Interpretation

Let me begin by noting that a substantial portion of Floyd and Putnam's response to my paper rests on a serious, and a fairly wholesale, mischaracterization of the contents of that paper. While I don't want to sort through all the minor details here (though see footnote 16 for a bit of this), I do want to discuss one, particularly large-scale, issue of this kind. On page 104, Floyd and Putnam claim that my paper involves adopting a central strand of Mark Steiner's recent interpretation of the notorious paragraph.[4] They then

[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 an explicit construction of P along with a discussion of why P might *seem* to say "I am not provable," see pages 199–200 of Timothy Bays, "On Floyd and Putnam of Wittgenstein on Gödel," *The Journal of Philosophy* CI.4 (2004): 197–210 (hereafter FPWG).

[3] See Juliet Floyd and Hilary Putnam, "Bays, Steiner and Wittgenstein's "Notorious" Paragraph about the Gödel theorem," *The Journal of Philosophy* 103 (2006): 101–110 (hereafter BSWG).

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

spend the next 4–5 pages (almost half of their overall paper) laying out and criticizing the "Bays-Steiner interpretation" of Wittgenstein's remarks. This interpretation involves a number of tricky technical issues—e.g., concerning the exact version of Gödel's theorem at issue in Wittgenstein's remarks—and a number of broader historical questions—e.g., concerning Wittgenstein's attitude towards the soundness of systems like PA and PM. It also involves some fairly specific—and, in Floyd and Putnam's view, fairly untenable—readings of particular parts of the notorious paragraph.[5]

Now, the problem with this whole section is simple: *none* of this interpretive material actually occurs in my paper. At the beginning of the paper, I explicitly say that I am not going to engage in any Wittgenstein interpretation (see p. 197). After the first few pages of the paper—pages where I quote the notorious paragraph in order to provide the background for Floyd and Putnam's introduction of KC—Wittgenstein drops out of the discussion almost entirely. With the exception of a brief footnote on page 208, all of my later references to Wittgenstein are completely incidental, and they could easily be eliminated without affecting any of my criticisms of KC.[6]

In short, then, there simply is no "Bays-Steiner interpretation" of the notorious paragraph. This is not because Steiner and I have different interpretations of the paragraph; it is because Steiner *has* an interpretation of the paragraph and I do not.[7] I did not present such an interpretation in my previous paper, I am not presenting one in this paper, and I very much doubt that I will ever present an interpretation of this particular passage.[8] In light of this, I won't go through Floyd and Putnam's paper and individually

[5] Some details are probably in order here. On pages 105, Floyd and Putnam claim that the Bays-Steiner interpretation rests on replacing Gödel's original version of his theorem with a more sophisticated version due to J. B. Rosser (they then go on to argue that Wittgenstein was quite clearly *not* talking about Rosser's version of the theorem). On page 105, and again on pages 106–107, they claim that my discussion of Wittgenstein rests on the the assumption that systems like PA and PM are *sound,* but that Wittgenstein would have questioned this assumption (at the very least, the claim that PM is sound is too much at issue in the notorious paragraph to simply presuppose it when giving a reading of that passage). Finally, on page 106, they discuss a particular argument which they claim I "attribute" to Wittgenstein, and on page 107 they discuss the way I purportedly read Wittgenstein's puzzling remark that "what is called 'losing' in chess may constitute winning in another game."

[6] That is, these later references all occur in phrases like the following: "using these formulas, Gödel defines the sentence which Wittgenstein later calls 'P' (although Gödel himself calls it something different)" or "I have formulated my arguments in terms of PA, while Wittgenstein, Floyd and Putnam formulated theirs in terms of 'Russell's system'." Clearly, the mention of Wittgenstein in these passages is peripheral and could easily be eliminated.

[7] Let me emphasize here that I am not distancing myself from Steiner's interpretation because I *disagree* with it. I think that Steiner's paper is rich and interesting, and I have learned a lot from reading it. But I have also learned a lot from reading some of Floyd's papers on this matter and (even) from the more interpretive sections of Floyd and Putnam's original essay. I simply lack the background to take a stand on this particular issue in Wittgenstein scholarship, and so I have (explicitly and repeatedly) declined to do so.

[8] I must say that I am somewhat perplexed as to why Floyd and Putnam even *think* that my paper involves a particular interpretation of Wittgenstein. Unfortunately, their own paper doesn't help much in this regard. In the section of their paper where they develop the "Bays-Steiner interpretation" (roughly pages 104–108), there are no quotes from my paper and only two parenthetical citations (neither of which refers to anything involving Wittgenstein interpretation). To be sure, Floyd and Putnam do cite *Steiner's* paper several times in this section, but they provide no grounds for attributing Steiner's arguments to me.

disclaim each of the interpretive views which they attribute to me (I'll let this paragraph constitute an "all purpose" disclaimer). Instead, I'll focus on two specific pieces of the "Bays-Steiner interpretation" which have a direct bearing on the purely philosophical issues involved in assessing KC.

First, on page 105, Floyd and Putnam claim that my discussion depends on replacing Gödel's original formulation of the incompleteness theorem with a strengthened formulation due to J. B. Rosser. Recall, here, that Rosser's theorem involves several modifications of Gödel's original theorem. For one thing, the two theorems involve slightly different sentences (I'll use "P" for Gödel's sentence and "R" for Rosser's).[9] For another, Gödel's theorem says that if PM is consistent, then $\mathrm{PM} \nvdash \mathrm{P}$ while Rosser's theorem says that if PM is consistent, then $\mathrm{PM} \nvdash \mathrm{R}$ *and* $\mathrm{PM} \nvdash \neg\,\mathrm{R}$.[10]

Now, leaving aside any purely interpretive issues—i.e., which sentence was Wittgenstein actually talking about in the notorious paragraph—I'll note that a move to Rosser's theorem would seriously trivialize any discussion of Floyd and Putnam's KC. In thinking about KC, we want to know what would happen if 1.) PM was consistent but 2.) we discovered that $\mathrm{PM} \vdash \neg\,\mathrm{P}$. If the "P" here is Rosser's sentence, then there's nothing to discuss, for Rosser's theorem shows that the situation in question simply cannot arise. In the context of assessing Floyd and Putnam's argument, therefore, a move to Rosser's theorem would be quite unfair.[11]

Once again, though, the situation here is pretty simple. At no point in FPWG do I actually make the switch that Floyd and Putnam claim I make—i.e., the switch from Gödel's original theorem to Rosser's theorem. On page 199, I give an explicit formulation of the sentence P which I'm going to be discussing throughout the paper. Modulo a shift from PM to PA (about which I'll say more in section 3), this sentence is exactly the same as that used in Gödel's original proof, and it's the same as that which Floyd and Putnam were discussing in their original WNP paper. For the remainder of my paper, I neither change the P that I'm working with nor rely on the claim that $\mathrm{PM} \nvdash \neg\,\mathrm{P}$ (i.e., the claim which characterizes Rosser's version of the incompleteness theorem). In short: while a switch from Gödel's theorem to Rosser's theorem *would be* a big deal, it's just not a switch that occurs in the paper I actually wrote.[12]

Second, on page 105 and again on pages 106–107, Floyd and Putnam claim that my argument depends

[9] See page 105 of Floyd and Putnam's new paper for an explicit formulation of Rosser's sentence and a nice explanation of the differences between R and P.

[10] So, Gödel's original theorem doesn't generate the conclusion that $\mathrm{PM} \nvdash \neg\,\mathrm{P}$. Of course, if PM is sound—or even if it's just ω-consistent—then this conclusion does follow. But it doesn't follow from the mere consistency of PM.

[11] This is, essentially, the concern that Floyd and Putnam raise on page 105 of their paper (together with the more textual concern that moving from Gödel's theorem to Rosser's theorem is simply inaccurate as an account of what they—and perhaps Wittgenstein—were actually trying to do).

[12] As before, I'm puzzled as to why Floyd and Putnam even think that I made this particular switch. Unfortunately, this section of their paper contains *no* citations of my paper, so there's not much evidence to go on. My only hypothesis is this. On pages 259 and 262 of his paper, Steiner *does* make an explicit switch from Gödel's theorem to Rosser's theorem (as Floyd and Putnam point out on page 105 (n. 12) of their own paper). So, if you assume that Steiner and I are doing the same thing, then you might want to attribute this switch to me. I'll repeat, however, that there's nothing in my own paper which would justify this attribution.

on the assumption that we may safely assume that systems like PA (and perhaps even PM) are sound. As with the move to Rosser's theorem, this assumption would trivialize any discussion of KC. If PM is sound, then it's a straightforward *theorem* that P is true and that PM $\nvdash \neg$ P . Hence, the assumption that PM is sound is out of place in any serious discussion of KC. In Floyd and Putnam's words, the whole point of KC is "to ask whether we would hold on to our English interpretation of P as 'P is not probable' if $\neg$ P were proved *and we therefore realized that PM was not sound*" (BSWG p. 106).

Now, by this point, it should be clear what I'm going to say about this. The analysis of KC that I give in FPWG doesn't, contrary to what Floyd and Putnam assert, presuppose anything about the soundness of either PA or PM. It's true that, in the opening pages of the paper (pp. 199–200), I lay out a quick proof of Gödel's incompleteness theorem which makes limited use of the soundness of PA. I do this to provide a bit of background, to explain, as I put it there, "why someone might think that P is a true but unprovable sentence of number theory."[13] But once I turn to examining KC itself, this soundness assumption vanishes. In describing Floyd and Putnam's argument for KC, I explicitly note that the hypothesis that PA $\vdash \neg$ P entails that PA is not sound (see p. 200), and I repeat this point right at the beginning of section 2 (see p. 203).[14] In neither of these cases do I make the trivializing response "but we have a theorem showing that PA is sound, so this can't actually happen."[15] In fact, the entirety of section 2 of my paper (pp. 203–206) consists of my own analysis of what we would/should do if PA turned out not to be sound, and section 4 extends this analysis from PA to PM. So, the claim that my argument rests on the assumption that PA and PM are sound is simply false.[16]

[13] I should note that, even in this proof, the assumption that PA is sound can easily be eliminated. Once we have the arithmetic expressibility of PROOF and SUBST, we can generate the biconditional $\mathbb{N} \models$ P $\Longleftrightarrow$ PA $\nvdash$ P (see FPWG p. 200). If we also assume that PA is consistent, then Gödel's original theorem shows that PA $\nvdash$ P . So, we once again have the conclusion that $\mathbb{N} \models$ P and PA $\nvdash$ P. Nor do we need the soundness of PA to generate arithmetic expressibility. As I note on page 205 of FPWG, the soundness of Robinson's arithmetic is enough to generate the relevant numeralwise representability and arithmetic expressibility results (and there are even ways of generating arithmetic expressibility which don't go through numeralwise representability). So, while assuming that PA is sound makes our proof a bit *easier,* it is in no way essential to that proof.

[14] Since I'm using model-theoretic machinery to spell things out, the claim that PA is not sound takes the form: "$\mathbb{N} \not\models$ PA."

[15] Though this is, pretty clearly, the response Floyd and Putnam want to attribute to me. See pages 105 and 106 of BSWN.

[16] For a final time, I will note that I find it hard to see why Floyd and Putnam even *think* that my argument rests on assumptions about the soundness of PA and PM. In their discussion of these matters, the only citation of my paper is to the quick argument on page 200 (i.e., the one described above); they never mention the section of my paper where I explicitly consider the results of assuming that PA is unsound (i.e., section 2). As before, my only guess is this. On page 267 of his 2001 paper, Steiner's *does* make use of the assumption that PA is sound, and he gives an argument which looks very much like that described on p. 104 of BSWN (indeed Floyd and Putnam quote part of Steiner's argument *on* p. 104 of BSWN). Once again, then, if you simply assume that Steiner and I are doing the same thing, then you might attribute this assumption to me. But there's nothing in my own paper which would justify this attribution.

Let me make two final comments on this matter. First, I want to reiterate that nothing I've said here should be construed as a criticism of Steiner. Steiner's paper concerns a different topic than mine—in particular, he's not responding to Floyd and Putnam's WNP paper, and so he's not trying to assess their argument for KC. In Steiner's context, therefore, things like using Rosser's theorem or assuming the soundness of PA may well be legitimate. To determine whether they *are* legitimate, of course, we'd have to broach some issues in Wittgenstein scholarship, and, as I've already noted, that isn't a project which I plan to

2 My Actual Argument

So what *did* I say about KC in FPWG? To understand my argument, we need to begin by recalling what Floyd and Putnam themselves said about KC in their original WNP paper. Fortunately, their new paper provides an elegant summary of the main argument of (the relevant section of) their earlier paper. To avoid any misrepresentation, I'll simply quote Floyd and Putnam's summary:

> Inspired by Wittgenstein's notorious remarks, we argued as follows. Suppose that to our surprise we have discovered a proof of the negation of a Gödel sentence, "¬ P", in *Principia Mathematica.* Suppose too that PM is consistent. Under these suppositions it is a well-known, uncontroversial consequence of Gödel's theorem that PM will be ω-inconsistent. As may be seen from inspection of the definition of "ω-inconsistent" (see footnote 1), this means that every model of PM must contain entities that are not natural numbers. In fact, in any such model every numerical predicate with an infinite extension will "overspill", that is, contain some elements that are not natural numbers. But then, because our original rigorization of the syntactic notions applied in the English sentence came through Gödel numbering with the natural numbers alone (as in Gödel's original (1931) paper), our initial way of translating into English or German "P" (the formula of PM whose proposed translation was "P is not provable in PM") would, in this context, become more nuanced and complicated, would have to be given up in its original form, just as Wittgenstein observed. (BSWN p. 102)

undertake here. Instead, I'll just acknowledge that in *my* context—i.e., the context of evaluating Floyd and Putnam's argument for KC—these particular moves *would be* illegitimate, but I'll repeat that they are not moves that I make in my paper.

Second, the above discussion focuses on some fairly large-scale ways in which Floyd and Putnam's paper misconstrues mine. Things don't get much better when we move to the local level. On page 103, for instance, they criticize me for assuming that "inspired by Wittgenstein, we [Floyd and Putnam] wrote about Gödel in the grip of anti-realism, verificationism, and/or formalism." However, a simple search will show that none of these terms occur in the text of my paper; nor does the passage that Floyd and Putnam reference from page 202 of my paper provide the resources for justifying this kind of claim. Similarly, on page 103, they claim that I ascribe to them "a 'crazy' position which is offering 'principled reasons for permitting us to step back to make syntactic generalizations, while forbidding us to step back to make semantic generalizations'." Now, it's certainly true that I discuss this particular position in my paper (p. 210), and it's also true that a bit later I discuss the "crazy" view 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" (e.g., that once Frege has proposed the foundational system of the *Grundgesetze,* he cannot legitimately discard that system when it turns out to be inconsistent). But, it's a just a mangling of my text to mix and match these quotations the way Floyd and Putnam do, and it seriously distorts the significance of what I said. (For the record, while I disagree with the position described in Floyd and Putnam's quotation, I think there are some interesting things to be said for it, and I don't regard it as "crazy.") Finally, in footnote 20 (p. 109), Floyd and Putnam quote and criticize a passage from pages 201–202 of my paper. The careful reader will note, however, that the ellipsis in their quote spans several sentences and that it actually goes *backwards* in the text; further, the interpolated "because" doesn't capture the relation between the two clauses connected by this ellipsis (indeed, I myself don't see *any* relation between these two clauses).

Now, I could go on with this kind of example for some time, but this section (and this footnote!) are already getting quite long. I'll end, therefore, by simply urging the reader to be very cautious about accepting Floyd and Putnam's descriptions of the contents of my paper and to check (even) their explicit citations very carefully.

This, then, is the line of argument which I took myself to be challenging in FPWG. Let me say two things about how this challenge was supposed to go.

First, there are a number of ways in which my analysis of KC actually agrees with the analysis given by Floyd and Putnam. Here are three places where we can all find common ground:

1. If PM $\vdash \neg$ P, then PM is ω-inconsistent.

2. If PM is consistent but ω-inconsistent, then all of the models of PM contain non-standard natural numbers—i.e., elements which the model treats as natural numbers but which don't correspond to any of the ordinary natural numbers.

3. The translation of P as "P is not provable" depends on interpreting P at the "natural numbers alone." If we interpret P at a non-standard model—i.e., at one of the models described in 2—then there is no reason to think that this will lead to a translation of P as "P is not provable."

Here, 1 and 2 are uncontroversial results of modern mathematical logic. 3 is a somewhat more philosophical claim which Floyd and Putnam defend on page 625 of WNP. I explore this claim in some detail on pages 201–202 of FPWG, and I explicitly accept it on page 203.

Second, and this is where my analysis disagrees with that given by Floyd and Putnam, I don't think the above argument suffices to establish KC. Suppose that PM $\vdash \neg$ P. Then 1–3 show that we cannot find a model which both 1.) satisfies PM and 2.) provides a solid basis for translating P as "P is not provable." In and of itself, however, this doesn't show that we should "give up" the translation of P as "P is not provable." That would only follow if we also assumed that our translation of P should be constrained by the class of models which happen to satisfy PM. *But there's no reason to accept this assumption.* Even if PM were ω-inconsistent, it would still make perfect sense to continue interpreting P on the "natural numbers alone"; should we do so, P would continue to receive the natural translation "P is not provable." Indeed, once we discover that PM is ω-inconsistent—so its version of number theory doesn't fit the ordinary natural numbers—then we have even *less* reason to view its models as being in any way canonical for the purposes of translating number-theoretic sentences (and *more* reason to look for a revision of PM which would remove the offensive ω-inconsistency!).

In my original paper, FPWG, I chose to cash this point out in terms of PA rather than PM. I still think that this is fairly illuminating. Suppose, therefore, that we have discovered that PA $\vdash \neg$ P. By an easy analog of the above argument, this shows that $\mathbb{N} \not\models$ PA . Hence, I argued, 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. (FPWG p. 203–204)

Note that if we take this second option, then we can (and, indeed, should) continue to translate P as "P is not provable." Further, and as I argued in my last paper, there is every reason to think that the mathematical community would in fact take the second option were they to discover that PA $\vdash \neg$ P. That is, they would continue to regard $\mathbb{N}$ as the canonical model for interpreting the language of arithmetic, and they would start to look for ways of modifying PA so as to eliminate the troubling ω-inconsistency.[17]

This, then, is the main point that my discussion of KC was supposed to establish. It's not that we have an *a priori* reason for rejecting the possibility that PM $\vdash \neg$ P (e.g., a reason which comes from invoking Rosser's theorem or a meta-theoretic proof of the soundness of PM). Nor is it that P will still come out as "P is not provable" when we interpret it on non-standard models of arithmetic. Rather, it's that the mere knowledge that PM is ω-inconsistent need not—and I think should not—lead us to abandon $\mathbb{N}$ as the standard model for interpreting the language of arithmetic. Hence, it need not lead us to abandon the interpretation of P as "P is not provable." This is enough to undercut Floyd and Putnam's argument for KC; further, if I'm right about how the mathematical community would/should actually respond to the discovery that PM $\vdash \neg$ P, then we can go further and say that KC is (simply) false.

3 PA and PM

In their original paper on these matters, Floyd and Putnam formulated their discussion of KC in terms of the logical system of *Principia Mathematica*.[18] In FPWG, I chose, for reasons of expository convenience, to conduct my discussion in terms of PA. In their new paper, Floyd and Putnam make rather heavy weather over this change. On page 106, they claim that the change "dodges the issue we take Wittgenstein to have been raising,"[19] and they suggest that my analysis would fail if we limited our discussion to PM. Similarly, on page 109 (n. 20), they worry that the switch from PM to PA leads me to "misread" some of their arguments and to attribute to them some things that they "never spoke about."

Now, it seems to me that there are really three different issues here. First, there's a purely scholarly issue. Wittgenstein himself was talking about PM, so it's historically inaccurate—and perhaps even somewhat anachronistic—to conduct my discussion in terms of PA. On the more contemporary front, making the switch from PM to PA may have misled readers of FPWG. That is, when I reformulated Floyd and Putnam's central argument in terms of PA, I may have given the false impression that Floyd and Putnam themselves were writing about PA in their 2000 paper (when, in fact, it's clear that they were writing about PM).

In response, let me say that these purely scholarly objections seem misguided. As I noted on pages 2–3 above, my original paper didn't focus on Wittgenstein at all: it focused solely on the philosophical

[17] For reasons of space, I won't rehash my argument for these claims here. See section 2 of FPWG for an explanation of why I think the mathematical community would respond this way and a discussion of some of the complexities which might result.

[18] So, P was the Gödel sentence *for* PM, and KC was formulated as it is on the first page of this paper.

[19] On the Wittgenstein issue, Floyd and Putnam write: "there is no reason to think that Wittgenstein would have regarded PA in the same light as PM—and no reason to think that we, looking back all these decades later, should do so either" (p. 106).

side of Floyd and Putnam's argument for KC. In this context, therefore, issues of accuracy in Wittgenstein interpretation don't really arise. On the more contemporary front, my paper was quite explicit about the move I was making when I switched from PM to PA. On page 198 (n. 4), I noted that Wittgenstein was working in PM, but that I intended to recast the argument in terms of PA; I then directed the reader to section 4 of my paper for a more-detailed discussion this switch. At the beginning of section 4 (p. 208), I again reminded the reader that "I have formulated my arguments in terms of PA, while Wittgenstein, Floyd and Putnam formulated theirs in terms of 'Russell's system'," and I then discussed the justification for this particular reformulation. Given this, I don't think there's any real possibility that my switch from PM to PA either misled my readers or mischaracterized Floyd and Putnam's argument.

A second issue is potentially more serious. On page 106, Floyd and Putnam suggest that switching from PM to PA has some genuinely philosophical consequences for my argument. Assuming that my argument turns on the soundness of PA, they write:

> Bays's assumption is that the soundness of PA can be proved in whatever system we employ to formalize Tarski's theory of truth. But that's not the same as assuming that that system can prove the soundness of Russell's *Principia Mathematica*!... There is no reason to think that Wittgenstein would have regarded PA in the same light as PM—and no reason to think that we, looking back all these decades later, should do so either. (BSWN p. 106)

Now, if my paper had made something like the trivializing response to KC discussed back on page 4—i.e., the response which appeals to the soundness of PA to rule out the very possibility that PA $\vdash \neg$ P—then this argument would pose a serious problem. Floyd and Putnam are clearly right that the soundness of PM is a (much) stronger assumption than the soundness of PA; so, in any context where such soundness is required, a switch from PM to PA would be a pretty big deal.

Once again, though, I'll simply repeat that nothing in the argument of FPWG relied on the assumption that PA is sound (or that PM is, for that matter). Indeed, as we saw in the last section, my argument follows Floyd and Putnam in supposing that we have discovered a proof that PA is *not* sound, and it then examines the appropriate response to this discovery: do we give up PA as an axiomatization of arithmetic or do we give up "P is not provable" as the translation of P? Whatever problems a move from PM to PA may create for *this* argument, they aren't the problems Floyd and Putnam highlight. Although Floyd and Putnam *do* raise a serious concern here, it's just not a concern which engages with the actual argument of my paper.

This brings me to a final issue: how *does* the move from PM to PA effect the argument that I actually gave in FPWG? Let me sketch two possible answers to this question, starting with the answer I originally gave in section 4 of FPWG. There, I argued that if we discovered that PA $\vdash \neg$ P, then we should keep our interpretation of P as "P is not provable" and abandon the assumption that PA provides an acceptable axiomatization of arithmetic. This argument rests on my larger view that we are—and should be—more committed to preserving $\mathbb{N}$ as the canonical model for the language of arithmetic than to preserving PA as a standard axiomatization of arithmetic. So, insofar as we are even less committed to PM than we are to PA, my argument would only be strengthened by moving our whole discussion back to PM. As I put the point in FPWG:

> . . . 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. 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. (FPWG p. 208–209)

In short: Floyd and Putnam are clearly right when they claim that PA and PM should not be treated as equivalents or regarded "in the same light" (p. 106). PM is a far stronger—and a far less secure—axiom system than PA. But, while this difference may make variants of what I've called the "trivializing" argument weaker, it just makes my own argument all that much stronger.

So, there's one answer as to how the move from PM to PA might affect my overall argument. I should acknowledge, however, that there's a second answer available, and that it's an answer which would move us back in the direction of Floyd and Putnam's original argument. Suppose that someone comes to this whole discussion with three assumptions. First, they view type theory as a full-fledged foundation for arithmetic (so, they define numbers as classes of equinumerous classes of individuals, and they accept a class-theoretic definition of the successor function). Second, they've come to be quite skeptical concerning our conceptual grasp on the type-theoretic hierarchy: although they still accept type theory, they're teetering on the brink of rejecting it as incoherent. Finally, they view PM as our last and best attempt to formalize our type-theoretic intuitions. Given all this, the discovery that $\mathrm{PM} \vdash \neg \mathrm{P}$ might be the final straw which leads this philosopher (or mathematician) to give up on type theory altogether. If so, then it would also force them to give up the idea that complex type-theoretic formulas like P have *any* fixed interpretation (and so, in particular, that P can be univocally interpreted as "P is not provable"). This, therefore, might provide a genuine situation where my shift from PM to PA would really obscure some important philosophical issues.[20]

Now, although I don't want to rule this second answer entirely out of court—since I don't want to rule out the very *possibility* that someone might find themselves in the philosophical situation I just described—I myself have two problems with it. First, I find it hard to generate the requisite level of skepticism concerning the type-theoretic hierarchy. I think we have a pretty good grasp on the structure of type-theoretic versions of the natural numbers, and I think that these structures themselves usually take precedence over the axiom systems we use to describe them (e.g. over PM). So, I'm still inclined to think that we should abandon PM as an axiomatization of arithmetic, before we abandon the natural interpretation of P as "P is not provable."

[20] At least, this would happen if we make the further assumption that we're less skeptical of ordinary arithmetic then we are of type theory—i.e., that we're quite confident of our grasp on the natural numbers and that we're completely uninclined to give up our interpretation of simple arithmetical langauge on the basis of a discovery that $\mathrm{PA} \vdash \neg P$. It's only when we combine this later assumption with the three assumptions described above that we get an asymmetry between the cases of PM and PA.

Second, and more importantly, I don't think that this second answer sits comfortably with the versions of PM and P that are actually in play in the current discussion—i.e., the ones given in Gödel's original presentation of the incompleteness theorem.[21] The version of PM used in Gödel's original theorem is non-standard in two ways. First, its language includes primitive expressions for 0 and for the successor operation on the natural numbers. Second, it includes several of the Peano postulates as primitive axioms. Given this, Russell's type theory serves only to provide the background quantificational logic, and the presence of higher types in this logic gives Gödel the resources to provide recursive definitions of functions like $+$ and $\times$.[22] To describe this system in Gödel's own terms, it is "the system obtained when the logic of PM is superimposed upon the Peano axioms (with the numbers as individuals and the successor relation as primitive notation)."[23]

In this particular case, therefore, there are deep similarities between the language of first-order PA and that of the relevant version of PM—the version used in Gödel's original paper, and the one which Floyd, Putnam and Wittgenstein are presumably talking about in their own pieces. Gödel is using an explicitly arithmetical language, his axioms include a version of the Peano axioms, and his whole set-up displays a clear intent to focus on the natural numbers. Given this, there's just no real question but that $\mathbb{N}$ is supposed to provide the intended interpretation of the arithmetical primitives in Gödel's language, any more than there is in the case of ordinary, first-order PA. Nor is there any need to define $\mathbb{N}$, 0, and successor by way of complicated type-theoretic formulas. Instead, as noted above, the higher-order apparatus of Russell's type theory serves only to formalize Gödel's background logic. In this particular case, therefore, my original argument for moving between PM and PA seems especially compelling. Insofar as the logic of PM is more complicated—and so less secure—than the logic of ordinary first-order PA, it should be even more clear that we would revise this version of PM to deal with a case of ω-inconsistency than that we would revise PA. In contrast, the fact that Gödel's version of PM *doesn't* give type-theoretic definitions of $\mathbb{N}$, 0, and successor makes the "second answer" discussed above look significantly less plausible.[24]

[21] At several points, Floyd and Putnam emphasize that the P at issue in their paper—and in Wittgenstein's remarks—is the one from Gödel's original presentation of the incompleteness theorem, "the one Wittgenstein saw" (see BSWN p. 105 and 107). I will take them at their word about their own intentions on this matter, and I will defer to them concerning Wittgenstein's. At any rate, Gödel's P is the one that I was thinking of in my original paper on this issue.

[22] Let me clarify a few things here. If our language includes $+$ and $\times$ as primitives, then we don't need higher types to give—or at least to mimic—recursive definitions. Gödel's own treatment of numeralwise representability and arithmetic expressibility shows how we can do without higher types in this case. If our language only includes 0 and the successor function, however, then we *do* need higher types to generate $+$ and $\times$. In a first-order context, 0 and successor don't provide enough resources to generate—or even to mimic—more complicated functions like $+$ and $\times$. Indeed, $\langle \mathbb{N}; 0, ' \rangle$ is a so-called *minimal* structure: no non-trivial relations are first-order definable over it.

[23] See Kurt Gödel, "On formally undecidable propositions of *Principia Mathematica* and related systems," in *Collected Works* (New York: Oxford University Press, 1986): 151–155. The quoted passage is from p. 151.

[24] Although I think this historical point is telling, I don't want to overemphasize it. There are two problems here. First, despite their repeated claims to be focusing on Gödel's own formulation of P, Floyd and Putnam's explicit descriptions of P often look somewhat different from Gödel's. In particular, they tend to assume that we need a defined predicate to pick out the "natural numbers" of our system (see WNP p. 625; see also BSWN p. 102 n. 5 and p. 107), while Gödel's own formulation of PM needs—and uses—no such predicate. On Gödel's formulation, all individuals are numbers, so we can simply use first-level

4 What's the Issue?

At the end of the day, after all the misunderstandings have been cleared up and all the charges of mutual mischaracterization have been resolved, what is the underlying disagreement between myself and Floyd and Putnam really all about? In this section, I want to hazzard an answer to this question, and to tie it back to some issues from our earlier papers. Before doing so, I should note that this answer will be at least partially speculative. While I will try to be quite explicit about my own views on certain matters, some of Floyd and Putnam's views are still a bit opaque to me; hence, my comments about their views should be treated as merely conjectural. Nevertheless, I think the following remarks probably do isolate an important source of our disagreement.

As a general matter, I tend to think that mathematicians and philosophers are justified in taking a naively realistic attitude towards many—and perhaps most—of the structures of classical mathematics. I think we have a pretty good grip on the natural numbers, the real and complex numbers, the lower levels of the set-theoretic hierarchy, simple type theory over the natural numbers, etc.[25] As Floyd and Putnam observe (BSWN p. 110), this means that I'm committed to bivalence for the languages we use to describe these structures, and I think that the structures themselves usually take precedence over the axioms systems we use to characterize them.[26] In the the current controversy, this explains—or at least places into a somewhat larger context—my view that we both are and should be more committed to preserving $\mathbb{N}$ as the standard model of arithmetic than to preserving either PA or PM as standard axiomatizations.[27]

variables to pick those numbers out. Given this, I'm cautious about tying too much of my argument to the details of Gödel's own paper, since it's not entirely clear that these details correspond to the version of PM that Floyd and Putnam really want to talk about.

Second, and as I argued several paragraphs ago, I think my general argument works whether or not we use Gödel's particular formulation of type theory—i.e., I think it also works for versions of type theory in which the numbers are not taken as primitives but are defined using purely logical machinery. Although Gödel's formulation *highlights* the number-theoretic content of P quite nicely, and although it's ostensibly the formulation at issue in this discussion, it isn't really essential to my overall analysis. Again, therefore, I don't want to hang too much of my argument on the details of Gödel's original paper.

[25] For what it's worth, I would be a little more cautious about full-fledged set theory, since I think there are genuine problems concerning the intended height of the set-theoretic universe. I would also be cautious about general type theory, since I don't think there's a single intended set of individuals (or even a fixed cardinality for the set of individuals). That being said, standard formulations of arithmetic within type theory lead to isomorphic structures whatever set of individuals we use. Hence, the arithmetical portion of type theory is still quite definite.

[26] Clearly, this is too large a view to defend in a short note. Even in the case of arithmetic, exploring the view thoroughly would involve looking at things like the status of second-order logic, the relationship between our understanding of arithmetic and our understanding of the syntax of formal systems, etc. More generally, we'd have address some deep questions in the philosophical foundations of model theory—e.g., are models primarily tools which help us to understand languages and theories, or are languages just combinatorial objects which help us to analyze and understand antecedently given structures? For now, therefore, I'll eschew this defense and simply focus on the role these views play in my disagreement with Floyd and Putnam.

[27] On page 109 of BSWN, Floyd and Putnam suggest that my focus on the model theory of $\mathbb{N}$ involves a failure to recognize distinctions between the notions of *model, interpretation,* and *translation.* This seems quite misleading. At several places in FPWG, I explicitly discuss the fact that there are different ways thinking about the notion of interpretation for arithmetical

From their paper, I would guess that Floyd and Putnam are somewhat skeptical of this kind of naiveté. It's not that they reject it *a priori,* nor that their argument for KC presupposes such a rejection. Rather, I think their argument for KC is supposed to put a kind of *conditional* pressure on views like mine. We come to Gödel's theorem with the naive idea that P can be interpreted on "the natural numbers alone" and that, so interpreted, it can be translated as "P is unprovable." To our surprise, we discover that PM $\vdash \neg$ P, and we then realize that things are "more nuanced and complicated" than we at first supposed. We can no longer assume that our Gödel numbering uses "the natural numbers alone," and we find that the translation of P as "P is unprovable" has to be "given up." In short: *if* PM $\vdash \neg$ P, *then* naive realism is in trouble.[28]

The point of my analysis of KC was simply to resist this final move. If we are antecedently inclined to view $\mathbb{N}$ as the canonical model for the language of arithmetic, then the mere discovery that PM $\vdash \neg$ P (or that PA $\vdash \neg$ P, for that matter) doesn't put any pressure on this view. It is perfectly coherent to respond to the discovery that PM $\vdash \neg$ P by abandoning our commitment to PM while maintaining our commitment to $\mathbb{N}$. Further, and as I argued in FPWG, I think this is exactly what the mathematical community would actually do should they discover that PM $\vdash \neg$ P (and I think they would be *right* to take this approach). Hence, I don't find Floyd and Putnam's own analysis of KC very plausible.

Let me emphasize here that nothing in this argument—or in my original paper, for that matter—is supposed to constitute a positive case for taking a naive attitude towards the natural numbers. Nor does it (even) constitute a substantial objection to the underlying content of any of the less-naive views of arithmetic that we can find in the literature—e.g., formalism, instrumentalism, etc. Instead, its purpose is quite limited. It shows that the situation Floyd and Putnam envisage in their discussion of KC doesn't—or wouldn't—cause any real problems for those of us who are already inclined towards some kind of naive realism. Whatever problems our naivité may face (and there are certainly many!), the various "nuances" and "complications" which Floyd and Putnam find in KC simply aren't among them.[29]

languages (see, for instance, p. 203 n. 15 and p. 207). We can do it using model-theoretic machinery; we can do it using Tarski's original apparatus of language, meta-language, meta-meta-language, etc. (with its associated notion of translation); we can do it using Gödel's own mechanism of arithmetic expressibility. Clearly these approaches are technically different, and, for some purposes, the differences are philosophically significant. But, I don't think they are significant for the purposes of my argument in FPWG. Nor do Floyd and Putnam isolate any *particular* place in the argument where these differences would make a difference. Given this, I don't see any problem with my decision to use model-theoretic machinery for the sake of convenience and standardization. For a bit more on Putnam's own approach to models and interpretations, see Hilary Putnam, "Models and Reality," in *Realism and Reason* (Cambridge: Cambridge UP, 1983): 1–25, Timothy Bays, "On Putnam and his Models," *The Journal of Philosophy* XCVIII (2001): 331–50, and Timothy Bays, "Two Arguments Against Realism," (in preparation).

[28] Let me reiterate that I don't think that Floyd and Putnam's argument simply *presupposes* a rejection of naive realism. There are both textual and philosophical issues here. On the textual side, Floyd and Putnam repeatedly emphasize that they don't start with any such presupposition (see BSWN p. 102–104, 110). On the more philosophical side, such a presupposition would render Floyd and Putnam's larger discussion of KC somewhat pointless: unless we start with the naive idea that P can be interpreted on "the natural numbers alone" and that, so interpreted, it comes out as "P is unprovable," there isn't anything to "give up" when we eventually discover that PM $\vdash \neg$ P . If we simply started with a view that, say, tied truth to derivability or to truth-in-all-models, then we would have no reason accept the translation of P as "P is unprovable" in the first place.

[29] A clarificatory point may be in order here. In section 3 of FPWG, I discussed several methods for rigorously defining the

I'll end with a final comment on these matters. In the last few paragraphs of their latest paper, Floyd and Putnam note that several important logicians have rejected the kind of naive realism that I'm espousing—and, more specifically, the view that the language of arithmetic is fully bivalent. They ask whether I want to read figures like Brouwer, Heyting, and Goodman "out of the camp of philosophy" (p. 110). The answer, of course, is "no." I would no more want to read these figures out of philosophy than I would want to read Floyd and Putnam themselves out of philosophy. That being said, I do disagree with these figures—and with Floyd and Putnam—about some important philosophical issues. As far as I can see, nothing in Floyd and Putnam's discussion of KC puts any real pressure on my side of this controversy (or even lays out a hypothetical situation in which such pressure should be thought to arise). Showing this was the burden of my last paper, and it's where I will end this one.

References

Bays, Timothy. "On Floyd and Putnam of Wittgenstein on Gödel." *The Journal of Philosophy* CI.4 (2004): 197–210.

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

Bays, Timothy. "Two Arguments Against Realism." (In Preparation).

Floyd, Juliet and Hilary Putnam. "Bays, Steiner and Wittgenstein's "Notorious" Paragraph about the Gödel theorem." *The Journal of Philosophy* 103 (2006): 101–110.

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.

Putnam, Hilary. "Models and Reality." In *Realism and Reason.* Cambridge: Cambridge UP, 1983.

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

notion of truth in $\mathbb{N}$. This discussion had two purposes: 1.) to show that the claim "P is true but unprovable" can be given a rigorous mathematical formalization (so, there's nothing intrinsically "metaphysical" about the claim), and 2.) to show that the notion of truth involved in this claim is relevant to thinking about the incompleteness theorem (and, in particular, that it played a role in Gödel's *own* thinking about the theorem).

Of course, none of this really assuages the concerns of someone who's skeptical about our initial grasp on the natural numbers: any concerns we have about the numbers will simply carry over to concerns about the metatheory in which the notion "true in $\mathbb{N}$" gets formalized. But assuaging such concerns was not the purpose of FPWG. As noted above, FPWG simply presupposed that we have a firm grasp on $\mathbb{N}$, and section 3 then discussed some mathematically rigorous ways of adding a truth predicate to $\mathbb{N}$. *That's* very different from the project of addressing skeptical worries about $\mathbb{N}$ itself.