The Gödelian Inferences

CURTIS FRANKS
University of Notre Dame

I attribute an 'intensional reading' of the second incompleteness theorem to its author, Kurt Gödel. My argument builds partially on an analysis of intensional and extensional conceptions of meta-mathematics and partially on the context in which Gödel drew two familiar inferences from his theorem. Those inferences, and in particular the way that they appear in Gödel's writing, are so dubious on the extensional conception that one must doubt that Gödel could have understood his theorem extensionally. However, on the intensional conception the inferences are straightforward. For that reason I conclude that Gödel had an intensional understanding of his theorem. Since this conclusion is in tension with the generally accepted view of Gödel's understanding of mathematical truth, I explain how to reconcile that view with the intensional reading of the theorem that I attribute to Gödel. The result is a more detailed account of Gödel's conception of meta-mathematics than is currently available.

1. Introduction

Gödel's second incompleteness theorem (G2) produces for any ω-consistent, reasonably strong formal system S a specific formula that is neither provable nor refutable in S. Constructive versions of the first incompleteness theorem (G1) do the same. What makes G2 noteworthy, rather than redundant, is the fact that the G2 formula is a formalization of the statement that S is consistent. By contrast, the G1 formula is not the formalization of any inherently interesting meta-mathematical statement.

Because of the correspondence between the G2 formula and the consistency statement, writers routinely infer from G2 that the systems S do not prove their own consistency[1] and that no mathematical proof of the consistency of S can be formalized in S. It is not apparent whether these inferences are justified or, if they are, what features of the formalization of meta-mathematics warrant them. It is not even apparent whether they are the same inference. Some philosophical attention has been directed at these questions, but by and large those discussions have entangled the question about the validity

[1]It suffices that S be consistent (and not ω-consistent) for the G2 formula to be unprovable in S. Thus the inferred claim is that consistent mathematical systems cannot prove their own consistency.

of these inferences with another one about the relationship between G2 and the search for consistency proofs adequate to answer the foundational concerns of the Hilbert school. Since the interpretation of Hilbert's foundational views is itself a contentious matter, this entanglement is unhelpful.

In this paper I present the inferences as a philosophical topic independent of considerations about foundational epistemology and sketch the most promising ways of thinking about them. Since the first writer to draw these inferences was Gödel, I call them the Gödelian inferences. I begin in the next section by presenting the inferences as they appear in Gödel's original *1931* paper. By making their role in that paper's dialectic clear I hope to disentangle them from considerations of the viability of Hilbert's program. In §3 and §4 I proceed by describing two approaches to understanding meta-mathematics, surveying the proposals for how to justify the Gödelian inferences native to each approach, and discussing the assorted obstacles lying in the way of each proposal. My hope there is to present a philosophical problem, central to understanding fundamental techniques of modern logic and distinct from the impact of Gödel's theorems on the foundational program of the Hilbert school, in a way that is accessible to a broad philosophical audience. Finally, I consider what can be said about Gödel's own reasons for drawing the inferences. In §5 I formulate a conjecture about Gödel's reasoning that I hope sheds some light on his philosophical perspective. I consider Gödel's philosophical views worthy of careful attention (independent of their philosophical merits or faults) because of their role in his logical discoveries.

2. The inferences

Consider Gödel's statement of G2 in *1931*:

> Theorem XI. Let κ be any recursive consistent class of FORMULAS; then the SENTENTIAL FORMULA stating that κ is consistent is not κ-PROVABLE; in particular, the consistency of P is not provable in P. (p. 193)

After sketching a proof of this theorem Gödel wrote:

> The entire proof of Theorem XI carries over word for word to the axiom system of set theory, M, and to that of classical mathematics, A, and here, too, it yields this result: There is no consistency proof for M, or for A, that could be formalized in M, or A, respectively, provided M, or A, is consistent. (p. 195)

These passages from the last two pages of Gödel's landmark paper invite many questions. In the statement of Theorem XI, Gödel refers to 'the SENTENTIAL FORMULA

stating that κ is consistent'. Already in this first sentence Gödel appears to hold two commitments in no way justified by the proof that follows. First, in what sense does the formula under consideration 'state' that the formal system of which it is a formula is consistent? Second, in whatever sense it does, how do we know that there aren't other formulas that also do, so that it makes sense to speak of '*the* SENTENTIAL FORMULA' with this feature? The next clause invites further thought, still. Why, from the unprovability of a (or 'the') formula that states (in some way yet to be specified) that P is consistent, did Gödel infer that the consistency of P is not provable in P? Is there no way for a system like P to prove its own consistency other than by having among its theorems the appropriate SENTENTIAL FORMULA?

Finally, in the second passage Gödel drew a related but not obviously identical inference when he wrote that, if κ (standing again for P, M, or A) is consistent, then '[t]here is no consistency proof for [κ] that could be formalized in [κ]'. According to what Gödel proved, a formula Con_κ associated with the consistency of κ is unprovable if κ is consistent. In order for this result to rule out the possibility of formalizing a proof of κ's consistency in κ it must further be argued that any such formalization would involve a κ-proof of Con_κ. What was Gödel's operative notion of formalization of proofs such that this inference is so natural as to need no explanation?

Immediately following this passage, Gödel added the following remark:

> I wish to note expressly that Theorem XI (and the corresponding results for M and A) do not contradict Hilbert's formalistic viewpoint. For this viewpoint presupposes only the existence of a consistency proof in which nothing but finitary means of proof is used, and it is conceivable that there exist finitary proofs that *cannot* be expressed in the formalism of P (or M or A). (p. 195)

Because Gödel issued this disclaimer about the Hilbert program just after his remarks about the impossibility of certain formalizations of consistency proofs, it is now customary to associate the questions that arise naturally from the first two passages with the epistemological concerns of foundational programs. This has further led to two unfortunate sociological phenomena. First, most writers who study Gödel's work have come to identify the philosophical significance of Gödel's theorems with whatever bearing they have on the foundational goals of the Hilbert school. Second, writers who attend to the questions that arise naturally out of the first two passages above tend to approach them against the backdrop of whatever understanding of Hilbert's foundational concerns they have adopted.[2] Though the questions might well be crucial to foundational issues, they

[2] *Detlefsen 1986* is the most extended study of these questions. But his evaluation of the Gödelian inferences is explicitly in terms of the impact of G2 on Hilbert's program, given an 'instrumentalist' interpretation of the latter (see especially pp. 117, 123, and 129). Similarly, Gödel's own brief reflection

need not be posed with an eye to such concerns. The questions are about how properly to understand the techniques of representing a system's meta-theory within the system itself. Since such techniques are at the heart of meta-mathematics, understanding them is a central problem for the philosophy of modern logic.

Although Gödel's own statement of the theorem is considerably richer than this,[3] I shall refer to the claim that $S \nvdash Con_S$ when S is consistent and sufficiently strong as G2, and distinguish from this theorem the two inferences that Gödel drew from it: first, that S does not prove its own consistency, and second, that no proof of the consistency of S can be formalized in S. These are the Gödelian inferences. I now turn to the problem of their justification.

3. Extensionality

On an *extensional* understanding of meta-mathematics, statements about formal systems are theory-independent, mathematical facts. Since they are second-order statements, their truth and falsity typically depend on what mathematical systems prove. For example, the statement that the formula ϕ is a theorem of the formal system S, written '$S \vdash \phi$', is true precisely when there is a proof in S of ϕ. However, the statements themselves need not be theorems of any mathematical system in order to be true. Thus '$S \vdash \phi$' is true so long as S proves ϕ, regardless of whether there is a proof in S or in any other formal system of any formalization of the statement '$S \vdash \phi$'.

Suppose one wants to know whether the system S proves its own consistency. For the present illustration, let the consistency statement of S be the meta-mathematical statement that $\perp$ is not derivable in S. Then the question of S's consistency is of the truth or falsity of the statement '$S \nvdash \perp$'. But the question about S proving its own consistency is about the truth or falsity of the statement '$S \vdash \overline{S \nvdash \perp}$', where '$\overline{S \nvdash \perp}$' is a proper formalization of the statement '$S \nvdash \perp$'. If this statement is true, i.e. if $S \vdash \overline{S \nvdash \perp}$ for some proper formalization of '$S \nvdash \perp$', then S proves its own consistency. And if $S \nvdash \overline{S \nvdash \perp}$ for all proper formalizations of '$S \nvdash \perp$', then S doesn't prove its own consistency, at least not in the form of the underivability of $\perp$. Thus to determine whether or not S can prove its own consistency, one needs a standard of propriety on formalizations of meta-mathematical claims.

---

on these questions in *Gödel 1972* is in terms of the sufficiency of the underivability of a formalization of 'outer consistency' to refute the Hilbert program. Whether or not an instrumentalist understanding of Hilbert's foundational aims wards off the threat that G2 poses to those aims, and whether or not Gödel's notion of 'outer consistency' really is what the Hilbert school needed a finitary proof of are interesting questions. But answering them would not illuminate the more basic questions about Gödel's remarks in *1931*, which are independent of foundational concerns.

[3] Gödel draws the first of the two inferences distinguished here in his official statement of Theorem XI.

Viewing meta-mathematics extensionally is natural, but hardly methodologically neutral. Suppose that S doesn't prove $\bot$ and is sufficiently strong so that G2 applies to S. May one draw the first Gödelian inference and say that S does not prove its own consistency? This depends on whether the unprovability in S of the formula Con_S suffices to show that all proper formalizations of '$S \vdash \bot$' are unprovable in S and also that all proper formalizations of other consistency statements are unprovable in S. One way to justify the Gödelian inference is to show that Con_S itself is a proper formalization of '$S \nvdash \bot$', that all other proper formalizations of '$S \nvdash \bot$' are not only equivalent to Con_S but S-provably so, and finally that all proper formalizations of all other statements of S's consistency are not only equivalent to Con_S but S-provably so. Then one will have shown that the statement of the consistency of S is not properly formalized by any formula provable in S. However, this is a non-trivial inference for the extensionalist, according to whom the various ways of stating the consistency of S are equivalent regardless of whether S proves such equivalences. Are the conditions sufficient for G2 also sufficient for S to be able to prove all such equivalences? If not, then it is consistent with G2 to assume that S proves its own consistency in the form of a proper formalization of a meta-mathematical statement of the consistency of S, and so the Gödelian inference is erroneous.

There is reason to doubt that on the extensional view one can answer this question affirmatively. It is possible to construct formulas analogous to Con_S that are actually provable in S. Their analogy to Con_S is that they are 'numeralwise correct' in the same sense that Con_S is. There are essentially four ways of doing this. The first is to change the way that the notion of provability is formulated within S: in *1936* Rosser showed that S proves a 'consistency formula' built up from a predicate extensionally equivalent (assuming that S is consistent) to the provability predicate Gödel used. *Mostowski 1966* constructed an example (p. 24) similar to Rosser's. Mostowski's example is easy to understand and for that reason is an illustrative example of the general phenomenon: Let $Proof_S(x, y)$ be the usual proof relation so that $Con_S \equiv \forall x \neg Proof_S(x, \overline{\bot})$. Then define $MPrf_S(x, y) \equiv Proof_S(x, y) \wedge \neg Proof_S(x, \overline{\bot})$ and $MCon_S \equiv \forall x \neg MPrf_S(x, \overline{\bot})$. For any consistent theory S, Con_S and $MCon_S$ are extensionally equivalent. But while Con_S is unprovable in S, $MCon_S$ is easily seen to be a theorem of S. Similarly *Takeuti 1955* showed that changing S's derivation rules can result in an equivalent theory S', but that the formula $Con_{S'}$ is provable in S'. *Feferman 1960* (and *Kreisel 1965*) showed that the same phenomenon occurs if one defines the axioms (respectively proofs) of S with a non-standard predicate. These results show that the first Gödelian inference is invalid if one takes 'numeralwise correctness' as one's standard of propriety.

An extensionalist response to this dilemma is to turn to generalized versions of G2 that exhibit the underivability of entire classes of formulas.[4] According to such results,

[4]This approach was pioneered by Bernays in *Hilbert and Bernays 1939* (p. 286) and has been built

under the usual assumption that S is in fact consistent, all formulas that meet a set of 'derivability conditions' are unprovable in S (since one such formula is Con_S, these results are strengthenings of Gödel's original version of G2). The Gödelian inference would be supported if some such conditions turned out not only to be sufficient for unprovability in S but also necessary conditions on a proper formalization of consistency. The extensionalist thus seeks a standard of propriety for formalizations captured by such a set of conditions.

Thus the challenge is to find a generalized version of G2 and show that its derivability conditions are met by all proper formalizations of consistency. In *1986* Detlefsen calls this challenge the 'stability problem'.[5] Solving it would secure the first Gödelian inference. It is not clear what a solution would look like, though, for there is no list of properties generally agreed upon as constitutive of S-consistency, in the sense that any formal statement of that notion must exhibit these properties. Indeed, an extensionalist might have supposed that numeralwise correctness was such a property on its own had it not been for the counter-evidence mentioned above. But it is perverse to back away from a standard of expressive adequacy simply because it proves to be insufficient to secure the Gödelian inference—perverse because doing so is to assume the Gödelian inference rather than to try to establish it. Something like a theory of meaning is needed in order to establish some standard or another up front before the stability problem can be approached scientifically.

On the extensional reading, the second Gödelian inference is problematic for similar reasons. The idea is to infer from G2 or from a generalized version of it that no proof of the consistency of S can be formalized in S. Clearly, if the first Gödelian inference is not justified, then neither is this second one. This is because the first inference fails just in case there is a proper formulation of the statement '$S \nvdash \bot$' or some equivalent consistency statement that is provable in S. In this case the extensionalist would agree that a proof of S's consistency is formalizable in S. It is less evident, though, that the failure of the first inference would follow from the failure of the second one. Proving some proper formulation of a statement of S's consistency in S is presumably one way of formalizing in S a proof of the consistency of S. But it is not obvious that it is the only way, given the breadth of the extensionalist's vision. One could, for example, translate a proof into S by reinterpreting the language of S. Then those familiar with the

upon in *Löb 1955* and *Jeroslow 1973*.

[5]Detlefsen defines the problem as the need 'to show that every set of properties sufficient to make a formula of T a fit expression of T's consistency is also sufficient to make that formula unprovable in T (if T is consistent)' (p. 81). An early, explicit recognition of the problem can be found in *Mostowski 1966* where Mostowski wrote, 'There are many formulae ... strongly representing [T's proof relation] in T. [G2] is valid only for some such formulae. It is not immediately obvious why the theorem proved for just this formula should have a philosophical importance while a similar theorem obtained by a different choice of a formula strongly representing the same set ... is simply false' (p. 23).

translation could view a proof in S of the formula ϕ as serving double duty by also being a proof of the formula that translates ϕ (and of whatever this object is a formalization of). Another reasonable sense in which a proof of ϕ could be formalized in S would be simply to define an interpretation in S of a stronger theory T that proves ϕ. Therefore, it is not clear initially whether the two Gödelian inferences are even saying the same thing, as Gödel seems to think they are. To show that they are, the extensionalist must argue for another standard of propriety and explain why the only proper way (according to this new standard) to formalize a proof of a meta-mathematical fact within S is to show that S proves a proper (according to the old standard) formalization of that fact.

Such is the way the Gödelian inferences look if one views meta-mathematics extensionally. On this view Gödel's mention of '*the* SENTENTIAL FORMULA stating that κ is consistent' appears unwarranted. While Gödel in fact constructed a single formula that allegedly expresses the consistency of P, the extensionalist sees no reason why there cannot be others, whose unprovability is not guaranteed by G2 as Gödel proves it. Meanwhile, the sense in which *any* formula is supposed to state that a formal system is consistent depends on a extensionalist theory of meaning or analysis of the notion of formal provability into a set of necessary conditions C_1. The first Gödelian inference then depends on the stability of a set of derivability conditions C_2 sufficient for a generalized version of G2, where their stability amounts to the demonstration that C_2 imply C_1. The second Gödelian inference depends on all this and more: either an explanation for why all proper formalizations in S of a consistency proof for S come in the form of S proving a proper formalization of its own consistency statement, or an explanation for why a sufficiently general version of G2 also rules out the possibility of formalizing a proof in some other way.

4. Intensionality

On the *intensional* conception of meta-mathematics, statements about mathematics are always part of a mathematical theory.[6] According to this conception, whether or not two formulas are equivalent is not a proper meta-mathematical question unless one is speaking about provable equivalence in a specific theory. Similarly, whether or not a formula defines a function is always to be determined by demonstrating within a theory the relevant totality and uniqueness conditions of functions for the formula. Reviving the example from the previous section, even the statement that the formula ϕ is a theorem of the formal system S is theory dependent: It is true for a theory T just in case T proves a proper formalization of the statement '$S \vdash \phi$'. Only now the question of propriety takes on a new form.

[6]The word 'intensionality' has been used in various ways. I don't intend to provide an analysis of any of these (e.g. Frege's) so much as to discuss a conception of meta-mathematics that is natural but different from the extensional conception canvassed above.

To see this, consider again the question whether the system S proves its own consistency. The question, in other words, is whether S proves that S doesn't prove '$\perp$'. As before, the immediate task one faces when approaching this question is to find a proper formalization of the statement '$S \nvdash \perp$'. But in the intensional context it is no longer enough merely to produce a formula and demonstrate, as the extensionalist must, that it adequately captures the intended notion. Because of the inherent theory-dependence of meta-mathematical questions, there is no general sense of adequacy for the intensionalist to invoke. There are only the various notions associated with individual theories. Thus one must show *in some system* T that a formula is a proper expression of the underivability of contradiction in S. Whether it is will depend on the theory T within which one poses the question. For example, if one wants to know whether PA proves the consistency of S, then one will want a formula that according to PA expresses the consistency of S. If one wants to know whether S proves its own consistency, then the natural choice for T is S.[7]

It is here that the notion of intensionality, as I have been using it, touches on the term's familiar use. The intensionalist thinks the verb 'prove', like the verbs 'know' and 'believe', creates an oblique linguistic context: A theory has not proved its own consistency unless it recognizes the proved formula as stating its consistency. Thus one must tailor the notion of adequate capture to S, and one also must show that the demonstration that a formula adequately captures the notion of consistency can be carried out in S. When these conditions are met, the formula is said to be *intensionally correct* for S—a label more suggestive than the extensional notion of 'adequacy'.

The intensional view just described is rather austere, for it seems to deny there being objective meta-mathematical facts. I shall take up this issue in the next section as part of an analysis of Gödel's views about meta-mathematics. For now I want to point out that one need not adopt a full-blooded intensionalist view in order to insist on formalizing meta-mathematical claims in the context of individual theories. There have been many motivations behind intensional views—from skepticism about the soundness of higher-level techniques to principled arguments against Platonism. But one need neither to be a skeptic nor to have a philosophical agenda in order to adopt an intensional reading of the question about a system proving its own consistency. For one could believe that there are theory-independent meta-mathematical questions and for that reason believe that there is a general sense of 'capture' to invoke at any time, and simply not invoke it when one understands questions in an inherently theory-dependent way. Thus even an extensionalist might *adopt* an intensional attitude when moving from the question 'is S

[7]This raises a question about what it would even mean for a mathematical system to recognize a formula as a consistency statement. I believe Kreisel was the first person to raise this question. Below I sketch two proposals for how to address it and their shortcomings. The one found in *Feferman 1960* was the first explicit attempt to address the question. The second proposal is based on ideas found in Kreisel's earlier work.

consistent?' to the question 'does S prove that it is?'.

On an intensional reading of this question, there is no question of 'stability'. If S fails to prove one formula that it recognizes as a statement of its own consistency, then necessarily S fails to prove any such formula. This is because the only intensional notion of equivalence is provable equivalence in a theory. Thus consider the usual consistency formula 'Con_S' and another '$CFCon_S$' that encodes the statement that there is no cut-free sequent calculus proof in S of '$\perp$'. If we assume that Gentzen's theorem about the admissibility of the cut-rule applies to S, then these formulas are extensionally equivalent. However, S still may be too weak to prove a formalized version of Gentzen's theorem and for that reason will not prove their equivalence.[8] Therefore they are in fact not equivalent from the point of view of that theory and, intensionally speaking, cannot express the same thing. At most one of these formulas can be intensionally correct for S. Thus S could not reasonably be said to recognize two formulas each as statements of its consistency and yet fail to prove their equivalence. So the first Gödelian inference is immediate as soon as one establishes an intensional version of G2. That is, once it is shown that one formula recognized by S as a statement of S's consistency is unprovable in S, one can justifiably infer that S does not prove its own consistency. If any putative consistency formula is provable by S, then the intensional version of G2 for S would be grounds for dismissing it as intensionally incorrect.

Of course the ordinary way to represent meta-theoretical notions with arithmetical formulas is by way of mapping signs onto numbers. This mapping necessarily takes place outside of the theory one investigates, as Georg Kreisel pointed out in *1958*: 'Gödel's work on formulae expressing the consistency of classical arithmetic goes beyond arithmetic concepts because it uses metamathematical interpretation' (p. 177). But intensionality requires that arithmetical systems themselves decode formulas. How is this to be done?

In *1960* Feferman suggested that the generalized versions of G2 mentioned in the last section are a way of getting at the notion of intensional correctness. Underlying his suggestion is the fact that the conditions sufficient for a formula's unprovability are relative to the system under consideration. For example, one common derivability condition is that the consistency formula Con_S be built out of a provability predicate Thm_S for which $S \vdash Thm_S(\phi) \wedge Thm_S(\phi \rightarrow \psi) \rightarrow Thm_S(\psi)$. That is, S itself should be able to verify that *modus ponens* holds for theorems under the Thm_S definition of provability. If closure under *modus ponens* is one of the conditions constitutive of theorem-hood, then this derivability condition shows not only that Thm_S captures that aspect of S-theoremhood, but also that S recognizes that it does.

However, if one pursues intensional correctness in this way, then one encounters anew

[8] Arithmetical systems that do not prove the totality of functions with super-exponential growth rate will be in this category.

one of the problems faced by the extensionalist. The problem is the need to justify a set of conditions as properly constitutive of theorem-hood so that their verification by S amounts to a demonstration of a formula's intensional correctness. To address this problem, one again must appeal to some theory of meaning, widely accepted examples of which are no more available to the intensionalist than to the extensionalist.

There is another problem the intensionalist faces in this pursuit which strikes me as more worrisome than this, though. This is the need to explain why the 'conditions constitutive of S-theoremhood' are identical regardless of the particulars of the system S. Intuitively one expects weaker systems to think of provability and consistency in quite more elementary ways than stronger theories do. An obvious example is that theories unable to prove Gentzen's result on the admissibility of the cut-rule might not think of provability in much the same way as do theories which do prove Gentzen's theorem. In particular they might not think that indirect 'proofs' involving cut are actually proofs. Since such theories are unable to prove the equivalence of the (extensionally equivalent) formulas '$Thm_S(x)$' and '$CFThm_S(x)$', at most one of them can be intensionally correct. In this case, the latter formula is more likely the correct one, since it encodes only facts about theorem-hood recognizable by S. But for the same reason that S cannot prove Gentzen's theorem, S will be unable to prove that '$CFThm_S(x)$' is closed under *modus ponens*. According to the analysis of *Feferman 1960* this speaks against this formula's correctness. However, on an intensional reading it seems more reasonable to conclude that closure under *modus ponens* is not one of the conditions constitutive of S-theoremhood after all. (Whether S proves its own consistency would thus be better viewed in terms of the S-provability of '$CFCon_S$', the unprovability of which in S does not follow from the unprovability in S of the intensionally stronger formula 'Con_S'.)

An alternative way to pursue intensionality can be found in *Kreisel 1951, 1952*. Kreisel's approach stems from work of *Herbrand 1930* and has been revitalized in *Franks 2009*. Herbrand proved that meta-mathematical questions of the form 'is ϕ a theorem of S?' are equivalent to certain Diophantine problems. Kreisel proposes that this gives one a way to view the 'constructive content' of meta-mathematical theorems: If the Diophantine problem is solvable by an effective substitution of numerical terms, then the meta-mathematical question can be answered within any theory whose provably-total functions make up such terms. According to this approach, a Diophantine equation produces for every theory T an intensionally correct formulation of the consistency of S: One asks whether the equation has a solution in the terms of T. If one formalizes this statement in the manner of *Gödel 1931*, the result is a formula of pure number theory that corresponds to the statement of S's consistency 'as', one might say, 'T thinks about the question'. Thus S could be shown not to prove its own consistency by showing that the formula constructed with S in place of T is not provable in S.

This second approach sidesteps the need for a theory of meaning, but at some apparent cost. In place of an analysis of what it means to be consistent or to be a proof,

the intensionalist builds formal consistency statements directly out of the resources of S. Whether S proves that the formulas constructed in this way have any of the expected properties of a consistency formula is then a secondary question. When S is strong, they do, but as the choice of S gets weaker and the consistency formula changes, more and more of the expected properties go unrecognized by S. This can be read as a vindication of the intuition that weak theories should be expected to think about what it means to be consistent in more elementary ways than strong theories do. On the other hand, one worries about pushing this too far, to theories so weak that they prove essentially nothing about their 'intensional' provability predicates and consistency statements. It is unclear at what point to stop thinking that one has recovered an even more basic notion of consistency ('consistency as S understands it') and to accept instead that one has failed to produce a formal statement of consistency at all.

On the intensional view, the second Gödelian inference is the same as the first. The reason is that if one views the formalization of a proof of S's consistency intensionally, then there is no reasonable way to proceed other than by proving in S the proper formalization of the statement '$S \nvdash \bot$'. Moreover the notion of propriety must be the same here as in the analysis of the first Gödelian inference. For if S proves a formula ϕ that it does not recognize as its own consistency statement, and one must verify that it is such a statement in a richer setting than S, then on the intensional view no proof of S's consistency has been formalized in S. Specifically, the interpretation of ϕ as a statement of S's consistency has not been formalized. Since this crucial step must be carried out in some theory stronger than S, it is only in this stronger theory that the entire proof of consistency has been formalized.[9] Thus the second Gödelian inference, like the first, is justified and in fact trivial as soon as the unprovability in S of an intensionally correct

[9] An anonymous referee asked a natural question about the generality of this last assertion. If the stronger theory is interpretable in S, then why wouldn't one say that S 'recognizes' the relevant formula as a consistency statement—just as the stronger theory does?

To see why, consider what happens if one adds to PA an axiom $\neg Con_{PA}$ saying that PA is inconsistent. The resulting theory is interpretable in PA, but one would not want to conclude from this that PA proves its own inconsistency! For a more subtle example, consider the fact that one can build an interpretation of $I\Delta_0 + \Omega_1$ in Q by relativizing all quantifiers to some predicate J defining an 'inductive cut' in Q's numbers. Thus one can relativize all the quantifiers in the usual consistency statement Con_Q to get a formula Con_Q^J that 'says' that there is no proof of contradiction encoded by a number in a model of $I\Delta_0 + \Omega_1$. Now, suppose that one thinks the usual formulation of Q's consistency is intensionally correct for $I\Delta_0 + \Omega_1$ because $I\Delta_0 + \Omega_1$ proves certain intensionality conditions for that formula. Then one might expect that Q will prove the same intensionality conditions for the formula Con_Q^J, so that the unprovability of this formula will attest to Q not proving its own consistency. But in general, a good deal more work needs to be done to verify this expectation (see §5.4 of *Franks 2009* for more of the details of this example). Moreover, even when the expectation is correct, one still faces the question whether the appropriate intensionality conditions for Q might not differ from those for $I\Delta_0 + \Omega$. So the interpretability of one theory in another is a way to recover all of the first theory's theorems, but not necessarily their meanings.

statement of S's consistency has been proved.

The conclusion to draw is that the Gödelian inferences are essentially forced on one if one views the statement of G2 intensionally. The first inference is immediate for the same reason that the stability problem doesn't arise: Since there cannot be multiple formalizations of S's consistency that are intensionally correct in S but not provably equivalent in S, the unprovability of one such formula suffices to secure S's inability to prove its own consistency. For the same reason, it is natural (even if not perfectly accurate) to speak of '*the* SENTENTIAL FORMULA stating that κ is consistent', since whatever κ proves about any one of them it proves about them all. The second inference follows once one realizes that one hasn't fully formalized a proof in S, in the intensional sense, unless the reasoning needed in order to tell what the proof is a proof of has also been formalized in S. Since the intensionalist considers such reasoning part of a consistency proof, the impossibility of formalizing a proof of S's consistency in S is not a new idea over and above the idea of S not proving its own consistency. It is a more precise statement of the old idea.

5. Gödel's view

In this final section, I argue that Gödel viewed his second incompleteness theorem intensionally. My evidence, largely, is the fact that Gödel drew the Gödelian inferences without fanfare or explanation. Since the inferences are not at all trivial on the extensional view, and indeed are fallacious on that view if one draws them, as Gödel did in *1931*, from a version of G2 that deals with a single formula, it would be uncharitable to attribute an extensional conception of the theorem to Gödel. It is also an implausible attribution in light of the famously extreme care that Gödel took in his writing. I do not take this as an argument for the 'correctness' of the intensional conception in any deep sense, but if I am right, then I think that the intensional conception of meta-mathematics has been justified in the pragmatic sense that it apparently was the way of thinking about consistency proofs that led to Gödel's great discovery. Before presenting my evidence in detail, I believe it is worthwhile to fend off an obvious objection and in the process make my conjecture more precise.

It is well known that Gödel defended a version of mathematical Platonism and even argued that his incompleteness theorems supported this view (e.g. in *Gödel 1951*, pp. 311-312). This distinguishes Gödel sharply from other pioneers of meta-mathematics, who variously espoused versions of constructivism, intuitionism, and finitism.[10] It is easy to see how these latter doctrines might have led to advances in meta-mathematics, since

[10] As the passages quoted below indicate, Gödel's own opinion was that his Platonic views were nearly unique among logicians of the era. It is unclear how accurate this self-assessment is, though. An anonymous referee for this journal cites Alfred Tarski as a logician whose practice of meta-mathematics was indicative of something of an objectivist conception, along the lines of Gödel's. On the other hand,

meta-mathematics tends to present formulas and proofs as concrete objects and meta-mathematical technique tends to involve constructive manipulations of these objects. On the other hand, in *1952* Dreben and more recently Buss in *1995* have argued that Herbrand's constructivism prevented him from formulating and proving the completeness of quantificational theory and that Gödel's Platonism is what guided him to the result.

In a letter to Hao Wang (reproduced in *Wang 1974*), Gödel himself claimed that his unique views about the nature of mathematics, and specifically about the nature of *meta-mathematics*, played a role in his discoveries:

> I would like to add that there was another reason which hampered logicians in the application to metamathematics, not only of transfinite reasoning, but of metamathematical reasoning in general. It consists in the fact that, largely, metamathematics was not considered [by them] as a science describing objective mathematical states of affairs, but rather as a theory of the human activity of handling symbols. (pp. 9-10)

And lest one think that Gödel had in mind only non-constructive results like his completeness theorem when he spoke of the value of his 'objective' view of meta-mathematics, he explicitly referred to his incompleteness theorems:

> . . . it should be noted that the heuristic principle of my construction of undecidable number theoretical propositions in the formal systems of mathematics is the highly transfinite concept of 'objective mathematical truth', as *opposed* to that of 'demonstrability.' . . . the use of this transfinite concept eventually leads to finitarily provable results, e.g., the general theorems about the existence of undecidable propositions in consistent formal systems. (p. 9)

Since Gödel viewed meta-mathematics 'as a science describing objective mathematical states of affairs', it would be odd to attribute to him 'the inherent theory-dependence of meta-mathematical questions'. But since this latter view is a commitment of the intensional position as I have described it, Gödel's explicit objectivism seems to rule out intensionality. Thus one might object to my claim that Gödel had an intensional understanding of G2 and especially to my claim that an intensional conception of meta-mathematics might have led Gödel to this theorem. Thus in *Mostowski 1966* one finds the following purely extensional depiction of Gödel's thought:

> The method invented by Gödel was to compare intuitively true properties of mathematical objects with properties expressible in the formal system

in a lecture delivered in 1966, Tarski described Platonism as 'so foreign and strange to [him]' (cited in *Sinaceur 2001*, p. 58). Elsewhere, he described himself as an 'extreme anti-Platonist' (cited in *Feferman 1999*, p. 61).

> under consideration. The sharp division of reasoning into intuitive meta-mathematics and formal mathematics was rejected on principle by the intuitionists; in the hands of Gödel this very division turned out to be an extremely valuable tool for establishing properties of formal systems. (p. 26)

I think that this objection is right in pointing out that Gödel was not a full-blooded intensionalist of the sort described in the last section.[11] It seems to me that Gödel had a largely extensional understanding of meta-mathematics consistent with his conception of meta-mathematics as an objective science. But I think the objection errs in concluding that he could therefore not have viewed G2 intensionally or have been led to G2 by intensional views of meta-mathematics. What I believe the Gödelian inferences show is that Gödel did not think the propriety of the G2 formula consisted simply in that formula *in fact* stating that the system P is consistent. Though his largely extensional conception of meta-mathematics made available to him such a theory-independent notion, the relevant fact for his purposes was that P *itself* recognized the G2 formula as a statement of P's consistency. Thus, although Gödel could have opted for a transparent reading of the provability relation, according to which all that matters is that the formula 'adequately capture' the statement of P's consistency, he chose not to. Gödel thought that the G2 formula was intensionally correct for P and, for that reason, also that G2 showed that P could not prove its own consistency in the intensional sense. This intensional reading of G2 is what secures the Gödelian inferences and therefore what I wish to attribute to Gödel.

In addition to the fact that Gödel drew the Gödelian inferences, there are four features of the broader context of his discussions of G2 that make those inferences particularly revealing of his intensional understanding of the theorem.

First is the fact that in *1931* Gödel suggested that the two inferences are synonymous. When Gödel discussed G2 in the context of P, he never drew the second inference nor summarized the impact of the theorem in terms of the formalizability of consistency proofs. He said only that 'the consistency of P is not provable in P'. But when he proceeded to generalize his theorem to other settings he wrote:

> The entire proof of Theorem XI carries over word for word to the axiom system of set theory, M, and to that of classical mathematics, A, *and here, too*, it yields this result: There is no consistency proof for M, or for A, that could be formalized in M, or A, respectively, provided M, or A, is consistent. (p. 195, my emphases)

This is the first instance of the second Gödelian inference in *1931*. Yet Gödel presents it as a reiteration of his remarks about P in the setting of M and A. If one understands G2

[11] Herbrand is an example of a prominent logician with this austere outlook. See, for example, his *1931*.

intensionally, then this presentation is natural. But on an extensional understanding it is not at all obvious that all formalizations in P of proofs of P's consistency will simply be instances of P proving its own consistency. At the very least one has to say something about the notion of formalizations of proofs. Gödel's silence on the matter speaks for intensionality.

There is also a statistical anomaly. When explaining the impact of G2, Gödel almost always preferred the formulation of the second inference. In fact, the *only* place he drew the first inference other than in *1931*, where it appears as part of his statement of G2, is in a note about Turing's precisification of the notion of formal system that he appended to the texts of *1931* and *1934* in the summers of 1963 and 1964, respectively.[12] By contrast, he used the second inference to explain G2 repeatedly, not only in *1931* but in *1930*, *1931a*, *1931?*, *1934*, and *1958*. In none of these cases did Gödel argue his way to the inference; he always drew it immediately.[13] This again would be odd if he understood his theorem extensionally, since the second inference is the more problematic one from that point of view. But on the view of meta-mathematics I am attributing to Gödel, his preference for the second formulation of the inference is understandable. Since his broadest conception of meta-mathematics *was* extensional, the statement that S doesn't prove its own consistency is ambiguous between S failing to prove any formula that actually expresses a statement of S's consistency and S failing to prove any formula that it recognizes as its own consistency statement. It is likely that Gödel preferred the second inference because he thought it cut through all ambiguity by forcing an intensional reading: One hasn't formalized a proof unless one has formalized every step of the proof including, in the case of a consistency proof, the step where one concludes that what has been proved is the system's consistency.

Not only did Gödel rarely draw the first inference, he even seems to have denied it in a remark he added as a note to the reprinting of his *1932* in *van Heijenoort 1967* and repeated as follows in *Gödel 1972*:

> Under the sole hypothesis that Z (number theory) is recursively one-to-one translatable into S, with demonstrability preserved in this direction, the consistency (in the sense of non-demonstrability of both a proposition and its negation), even of very strong systems S, *may* be provable in S, and even in primitive recursive number theory. (p. 305)

In seeming direct opposition to the claim of *Gödel 1931* that the consistency of P is not provable in P, here Gödel says that the consistency 'even of very strong systems S' may

[12] These notes appear at the ends of the cited versions of these papers from the *Collected Works*.

[13] On page 143 of *Gödel 1930* he presents the second inference with the word 'hence' [*also*], treating it properly as a consequence of, and not simply as the content of, G2. Even here, though, the inference is drawn directly from a statement of the theorem without any explanation.

be provable in S. In *1990* Feferman has, rightly I think, conjectured that Gödel was referring to the various counter-examples of Takeuti, Kreisel, et al. mentioned in §3 that make the stability problem so poignant. Did Gödel take back his earlier claim in light of these examples? I think that he did not. If he was harboring doubts about the stability of the first Gödelian inference in these later writings, then he was conspicuously silent about them. A better explanation of this remark is simply that he was assuming the extensional point of view in order to highlight the value of the 'outer consistency' notion that he went on to formulate.[14] Far from showing that he had an extensional reading of his theorem, though, I believe that this passage shows that both the extensional and intensional points of view were available to Gödel, so that from the former it is quite possible for a system to prove its own consistency while from the latter it is not. This is precisely the ambiguity that, granted that Gödel had an intensional reading of his theorem, would explain his preference for the second inference in almost all correspondence. The quoted passage indicates that Gödel was vividly aware of that ambiguity, which makes his preference for the second inference even more understandable.

Finally, Gödel occasionally shied away from the claim that the G2 formula is a statement of consistency. An example is his restatement of G2 in *1951*:

> ... for any well-defined system of axioms and rules, the proposition stating their consistency (or rather the equivalent number-theoretical proposition) is undemonstrable from these axioms and rules, provided these axioms and rules are consistent and suffice to derive a certain portion of the finitistic arithmetic of integers. (p. 308)

The same parenthetical qualification appears on page 327 of *Gödel 1932*. These passages indicate Gödel's uneasiness with the idea that the formal expressions of number theory are literally statements of consistency—an uneasiness that I think arises from the need of a theory of meaning to explain the sense in which formal expressions are supposed to 'state' anything about informal meta-mathematics at all. Gödel did point out that the formal expressions are equivalent to the appropriate sentences of informal meta-mathematics, but as discussed above, so too are many provable expressions. Gödel was able to issue these caveats with impunity, I think, because only on an extensional understanding of G2 does one need to explain why some of these formal expressions are consistency statements and some of them are not. It would indeed be awkward for anyone with an extensional reading of G2 both to draw the Gödelian inferences and to

[14] Gödel's discussion of 'outer consistency' is not helpful in sorting out his thinking on this point, since his target is again the viability of the Hilbert program and not the Gödelian inferences themselves (see note 1 above). His 'inner/outer' distinction is not directly related to the intensional/extensional distinction. Gödel's point was that even if one is interested merely in extensional results, because one thinks that this is all that is needed in order to realize Hilbert's aims, the unprovability in S of a Π_1-reflection principle for S of the form $Thm_S(\overline{\phi}) \rightarrow \phi$ suffices to refute these aims.

back away from the view that the G2 formula is literally a statement of consistency. But if one understands G2 intensionally, then one doesn't need to divide the expressions in this way. On the assumption that the G2 formula is intensionally adequate, Gödel could safely remain uncommitted to any theory about what arithmetical formulas 'mean' and replace this vague notion with a direct, simple claim about what the unprovability in S of that formula implies: that consistency proofs of S cannot be formalized in S.

Such is the evidence for the unique view of meta-mathematics that I attribute to Gödel. According to this view, G2 is properly understood intensionally although meta-mathematical statements in general describe 'objective mathematical states of affairs' and accordingly should be understood extensionally. The Gödelian inferences—both the simple fact that Gödel drew them and the specific contexts in which he did—make this view evident. In closing, I reiterate my hope that appreciating the subtlety in this view may draw us closer to understanding the line of thought that led Gödel to the discovery of his second incompleteness theorem. This is not to say that the intensional conception of meta-mathematics has been vindicated. Just as there are problems in the way of an extensional solution to the stability problem, there remain problems in the way of articulating a defensible notion of intensional correctness. Gödel's writings give no indication for why he thought the G2 formula was intensionally correct, only compelling evidence that he thought it was. Still, it is significant that Gödel's own reading of G2 can be reconstructed from textual clues in his various discussions of the theorem.

Acknowledgments

This paper grew out of my presentation to a seminar that Michael Detlefsen and I led at Notre Dame in the Fall of 2007. Michael Detlefsen, Graham Leach-Krouse, Charles Pence, Tony Strimple, and Vítězslav Švejdar had many helpful comments and suggestions. Patricia Blanchette, Timothy Bays, and three referees for this journal read earlier drafts of this essay and offered helpful advice.

References

Buss, S. 1995. 'On Herbrand's theorem', in *Logic and Computational Complexity* Lecture Notes in Computer Science, **960**, 195-209.

Detlefsen, M. 1986. *Hilbert's Program: An Essay on Mathematical Instrumentalism*, Dordrecht: D. Reidel Publishing Company.

Dreben, B. 1952. 'On the completeness of quantificational theory', in *Proceedings of the National Academy of Sciences of the United States of America*, 1047-1052.

Feferman, S. 1960. 'The arithmetization of metamathematics in a general setting', in *Fundamenta Mathematica*, **19**, 35-92.

Feferman, S. 1990. 'Remark 1' from the 'Introductory note to *Gödel 1972*', in *Gödel 1990*, 282-287.

Feferman, S. 1999. 'Tarksi and Gödel: between the lines', in *Wolenski and Köhler 1999*, 53-63.

Franks, C. 2009. *The Autonomy of Mathematical Knowledge: Hilbert's Program Revisited*. Cambridge: Cambridge University Press.

Gödel, K. 1986. *Collected Works, Vol I: Publications 1929-1936*. Edited by S. Feferman, et al. New York: Oxford University Press.

Gödel, K. 1990. *Collected Works, Vol II: Publications 1938-1974*. Edited by S. Feferman, et al. New York: Oxford University Press.

Gödel, K. 1995. *Collected Works, Vol III: Unpublished essays and lectures*. Edited by S. Feferman, et al. New York: Oxford University Press.

Gödel, K. 1930. 'Einige metamathematische Resultate über Entscheidungsdefinitheit und Widerspruchsfreiheit', translation by S. Bauer-Mengleberg as 'Some metamathematical results on completeness and consistency' reprinted in *Gödel 1986*, 140-143.

Gödel, K. 1931. 'Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I', translation by J. van Heijenoort as 'On formally undecidable propositions of *Principia Mathematica* and related systems I' reprinted in *Gödel 1986*, 144-195.

Gödel, K. 1931a. 'Diskussion zur Grundlegung der Mathematik', translation by J. Dawson as 'Discussion on providing a foundation for mathematics' reprinted in *Gödel 1986*, 200-205.

Gödel, K. 1931?. 'Über unentscheidbare Sätze', translation by S. Kleene as 'On undecidable sentences' in *Gödel 1995*, 30-35.

Gödel, K. 1932. 'Über Vollständigkeit und Widerspruchsfreiheit', translation by J. van Heijenoort as 'On completeness and consistency' reprinted in *Gödel 1986*, 234-237.

Gödel, K. 1934. 'On undecidable propositions of formal mathematical systems', reprinted in *Gödel 1986*, 346-369.

Gödel, K. 1951. 'Some basic theorems on the foundations of mathematics and their implications', in *Gödel 1995*, 304-323.

Gödel, K. 1958. 'Über eine bisher noch nicht benütte Erweiterung der finiten Standpunktes', translation by S. Bauer-Mengleberg and J. van Heijenoort as 'On an hitherto unutilized extension of the finitary standpoint' reprinted in *Gödel 1990*, 241-251.

Gödel, K. 1972. 'Some remarks on the undecidability results', in *Gödel 1990*, 305-306.

Herbrand, J. 1930. *Recherches sur la théorie de la démonstration.* Herbrand's doctoral thesis at the University of Paris. Translated by W. Goldfarb, except pp. 133-88 translated by B. Dreben and J. van Heijenoort, as 'Investigations in proof theory' in *Jacques Herbrand: Logical Writings*, edited by W. Goldfarb, Cambridge: Harvard Univeristy Press, 44-202.

Herbrand, J. 1931. 'Sur le problème fondamental de la logique mathématique', translation by W. Goldfarb as 'On the fundamental problem of mathematical logic' in *Jacques Herbrand: Logical Writings*, edited by W. Goldfarb, Cambridge: Harvard Univeristy Press, 215-271.

Hilbert, D. and P. Bernays. 1939. *Grundlagen der Mathematik, Vol. 2*, Berlin: Springer.

Jeroslow, R. G. 1973. 'Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem', *Journal of Symbolic Logic*, **38**, no. 3, 359-367.

Kreisel, G. 1951. 'On the interpretation of non-finitist proofs-part I', in *Journal of Symbolic Logic*, **16**, no. 4, 241-267.

Kreisel, G. 1951. 'On the interpretation of non-finitist proofs-part II', in *Journal of Symbolic Logic*, **17**, no. 1, 43-58.

Kreisel, G. 'Mathematical significance of consistency proofs', in *Journal of Symbolic Logic*, **23**, 159-182.

Kreisel, G. 1965. 'Mathematical logic', in *Lectures on Modern Mathematics, Vol. 3.* Edited by T. Saaty, New York: Wiley, 95-195.

Löb, M. H. 1955. 'Solution of a problem of Leon Henkin', in *Journal of Symbolic Logic*, **20**, no. 2, 115-118.

Mostowski, A. 1966. 'The incompleteness of arithmetic', in A. Mostowski, *Thirty Years of Foundational Studies*, New York: Barnes and Noble, 18-26.

Rosser, J. B. 1936. 'Extensions of some theorems of Gödel and Church', in *Journal of Symbolic Logic*, **1**, 87-91.

Sinaceur, H. 2001. 'Alfred Tarksi: semantic shift, heuristic shift in metamathematics', in *Synthese*, **126**, 49-65.

Takeuti, G. 1955. 'On the fundamental conjecture of GLC I', in *Journal of the Mathematical Society of Japan*, **7**, 249-275.

van Heijenoort, J. 1967. *From Frege to Gödel: A Sourcebook in Mathematical Logic*, Cambridge: Harvard University Press.

Wang, H. 1974. *From Mathematics to Philosophy*, London: Routledge & Kegan Paul.

Wolenski, J. and E. Kohler (eds.). 1999. *Alfred Tarski and the Vienna Circle. Austrio-Polish Connections in Logical Empiricism, vol. 6.* Vienna Circle Institute Yearbook. Dordrecht: Kluwer.