PHILOSOPHIA MATHEMATICA

http://www.UManitoba.CA/pm/

Editor: Robert S. D. Thomas, Department of Mathematics,
University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2
Robert_Thomas@UManitoba.CA or thomas@cc.UManitoba.CA

PHILOSOPHIA MATHEMATICA

PHILOSOPHY OF MATHEMATICS,
ITS LEARNING, AND ITS APPLICATION

PHILOSOPHIE DES MATHÉMATIQUES,
LEURS ÉTUDE ET APPLICATIONS

SERIES III

Founder
J. FANG

Editor
R. S. D. THOMAS

VOLUME NINE

2001

WINNIPEG
THE CANADIAN SOCIETY FOR HISTORY
AND PHILOSOPHY OF MATHEMATICS

LA SOCIÉTÉ CANADIENNE D'HISTOIRE ET
DE PHILOSOPHIE DES MATHÉMATIQUES

DUMMETT, M. [1978]: *Truth and Other Enigmas*. Cambridge, Mass.: Harvard University Press.
FEFERMAN, S. [1997]: 'Does mathematics need new axioms?', *American Mathematical Monthly* **106**, 99–111.
FREGE, G. [1893]: *Grundgesetze der Arithmetik: Begriffsschriftlich abgeleitet* Band I. Jena: H. Pohle. Reprinted in 1962 with Frege: [1903]. Hildesheim: Georg Olms.
———— [1903]: *Grundgesetze der Arithmetik: Begriffsschriftlich abgeleitet*. Band II. Jena: H. Pohle.
GÖDEL, K. [*1951]: 'Some basic theorems on the foundations of mathematics and their implications', in Gödel [1995], pp. 304–323.
———— [1995]: *Collected Works*. Vol. III. Oxford: Oxford University Press.
HERSH, R. [1997]: *What is Mathematics, Really?* Oxford: Oxford University Press.
TAIT, W. [1983]: 'Against intuitionism: Constructive mathematics is part of classical mathematics', *Journal of Philosophical Logic* **12**, 175–195.
———— [1986a]: 'Plato's second best method', *Review of Metaphysics* **39**, 455–482.
———— [1986b]: 'Truth and proof: The Platonism of mathematics', *Synthese* **69**, 341–370.
———— [1992]: 'Reflections on the concept of a *priori* truth and its corruption by Kant', in Detlefsen [1992], pp. 33–64.
———— [1997a]: 'Frege versus Cantor and Dedekind: On the concept of number', in Tait [1997b], pp. 213–248.
———— , ed. [1997b]: *Early Analytic Philosophy: Frege, Russell, Wittgenstein. Essays in honor of Leonard Linsky*. Chicago: Open Court.
———— [n.d.]: 'Noēsis: Plato on exact science', forthcoming in D. Malament, ed., A festschrift for Howard Stein. Chicago: Open Court.

ABSTRACT. This paper contains a defense against anti-realism in mathematics in the light both of incompleteness and of the fact that mathematics is a 'cultural artifact'. Anti-realism (here) is the view that theorems, say, of arithmetic cannot be taken at face value to express true propositions about the system of numbers but must be reconstrued to be about something else or about nothing at all. A 'bite-the-bullet' aspect of the defense is that, adopting new axioms, hitherto independent, is not a matter of recognizing truths which had previously been unrecognized, but of extending the domain of what is true.

What Does Gödel's Second Theorem Say?†

MICHAEL DETLEFSEN*

1. Introduction

The aim of this paper is to improve our understanding of the philosophical application of Gödel's Second Theorem (hereinafter, G2). Most specifically, it is to call attention to certain problems, heretofore largely unnoticed, facing the application of generalized versions of G2. As Bernays noted (cf. Hilbert-Bernays [1939], §§5.1c–5.2e), Gödel's original proof of G2 wanted generalization of two types. One of these is 'system generalization', the purpose of which is to secure the application of G2 to a broader class of theories than is provided for by Gödel's original proof. It consists in the specification of a general set of expressive, logical and arithmetical properties the possession of which by a theory would guarantee the applicability of G2 to it, at least with respect to consistency formulae constructed after the manner of the consistency formula of Gödel's original proof.

The other type of generalization—call it 'expression generalization'—is intended to extend the G2 phenomenon from the particular type of consistency formulae that Gödel used to *all* formulae capable of expressing consistency. The goal is to obtain a result that can be taken to show, of any theory to which it applies, that '*the* formalized expression of its [the theory's] consistency can not be derived within it so long as it is consistent' (cf. Hilbert and Bernays [1939], p. 324, brackets, emphasis and translation mine).

Bernays pursued expression generalization by identifying a set of general conditions (commonly referred to as the 'Hilbert-Bernays Derivability Conditions') on 'genuine' consistency formulae for any of the theories to

† I would like to thank audiences at the Notre Dame Logic Seminar, the Logic Seminar of Indiana University-Bloomington, the George Boolos Memorial Symposium and the Logic Colloquium of the University of Illinois-Urbana/Champaign for useful discussions of (shortened oral versions of) this paper. Among individuals, I am grateful to Andrew Arana, George Boolos, Peter Cholak, Anthony Everett, Matthew Frank, Jacob Heidenreich, Julia Knight, Tim McCarthy and Michael Stob for useful comments and discussion. They are not, of course, responsible for any deficiencies that may remain.

* Department of Philosophy, University of Notre Dame, Notre Dame, Indiana 46556 U. S. A. Detlefsen.1@nd.edu

PHILOSOPHIA MATHEMATICA (3) Vol. 9 (2001), pp. 37–71.

which G2 was to be applied.[1] The idea was that (i) any formula capable of expressing the consistency of the theory would have to satisfy these conditions, and that (ii) any formula satisfying them would be guaranteed to be unprovable in the theory so long as it was consistent.

Satisfaction of (ii) is a matter that can be (and has been) settled by proof. Satisfaction of (i), on the other hand, is not. It requires what we will refer to as *justification* of the Derivability Conditions (DCs). By justification of the DCs, we mean an acceptable argument showing that any formula Con_T capable of expressing the consistency of the system T must satisfy the DCs. To justify the DCs is therefore to establish them as necessary conditions on the ability of a formula to serve as an expression of consistency. Clearly, this is something different from proving the DCs.

Our chief concern in this paper is the justification of the DCs. More particularly, it is the justification of the so-called 'Third Derivability Condition' (the 'Third Condition', for short). More particularly still, it is one particular justification of the Third Condition—a justification we will refer to as the *Reflexivity Defense*.

We will argue that adopting the Reflexivity Defense has serious consequences. More accurately, we will argue that, for suitable system generalizations of G2, use of the Reflexivity Defense induces a Fourth Derivability Condition whose justification is fraught with difficulties. Our conclusion is therefore that the Reflexivity Defense does not provide a satisfactory justification of the Third Condition.

This naturally raises the question of whether there are other plausible justifications of the Third Condition that avoid introduction of the problematic 'Fourth' condition just mentioned. We lack the space to argue this matter properly here. We believe, however, that the answer is 'no', that the Reflexivity Defense is the most satisfying justification of the Third Condition that there is and that there is a high price to be paid for relinquishing it.

2. Clarifications and Qualifications

In order to develop our argument, certain preliminary clarifications need to be made and certain basic distinctions drawn. It is to these that we now turn, beginning with what we will refer to as the 'proto-philosophical' content of G2, or what G2 can be taken to 'say' for general purposes of philosophical application.

This notion is one-half of a distinction between literal and interpreted versions of G2. By literal versions we mean those which proceed by way of a mathematically precise description of various components of a given arithmetization. These components include (i) the theory (referred to here as the *represented* theory and designated throughout this paper by the letter 'T') whose syntax or metamathematics is to be represented by the arithmetization, (ii) the theory (referred to here as the *representing* theory, and typically designated by the letter 'S') in which the representation is to take place, and (iii) various formulae used by the representing theory to represent metamathematical notions or concepts (*e.g.*, the notion of consistency) pertaining to the represented theory. The descriptions of and conditions on these items play a central role in the proof of G2 and they form the main components of what we are referring to as *literal* versions of G2.

The following illustrates what we mean by a literal version of G2.

Literal G2: Let $Prov_T(x)$ be a formula of $\mathcal{L}_T$ (= the language of T) that satisfies the Derivability Conditions (*i.e.*, DC1–DC3 below) and the Diagonalization Lemma (*i.e.*, DL below). And let Con_T be the formula $\forall x(Prov_T(x) \rightarrow \neg Prov_T(neg(x)))$ of $\mathcal{L}_T$ (where '$neg(x)$' is a term of $\mathcal{L}_T$ that represents the negation function). Then, if T is consistent and its logic supports certain inferences and theorems (to be identified later), $\nvdash_T Con_T$.

DL: There is a sentence $\mathcal{G}$ of $\mathcal{L}_T$ such that $\vdash_T \mathcal{G} \leftrightarrow \neg Prov_T(\ulcorner \mathcal{G} \urcorner)$.

DC1: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$, if $\vdash_T \mathcal{A}$, then $\vdash_T Prov_T(\ulcorner \mathcal{A} \urcorner)$.

DC2: For all sentences $\mathcal{A}$, $\mathcal{B}$ of $\mathcal{L}_T$,
$\vdash_T Prov_T(\ulcorner \mathcal{A} \rightarrow \mathcal{B} \urcorner) \rightarrow (Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner \mathcal{B} \urcorner))$.

DC3: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$,
$\vdash_T Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner Prov_T(\ulcorner \mathcal{A} \urcorner) \urcorner)$.

Though illustrative of what we mean by a literal version of G2, Literal G2 is nonetheless not entirely typical of what we have in mind. The reason is that it fails adequately to mark a distinction that is vital to our argument: namely the distinction between the *represented* and *representing* theories of a given arithmetization.

Rather than allow the representing and represented theories to be different theories, which is the general case, Literal G2 assumes that a single theory (denoted by 'T' in the above statement) should play the role of both. This does not threaten the accuracy or demonstrability of Literal G2 since there clearly are *monotheoretic* versions of G2 (*i.e.*, versions in which the representing and represented theories are the same theory).

For certain purposes, however, it is necessary to allow the representing and represented theories to be distinct. Indeed, we consider this to be necessary for what is perhaps the most important philosophical application of G2—namely, that to Hilbert's Program. We are therefore interested in *bitheoretic* versions of G2; versions in which it is allowed that the representing and represented theories be different. We'll give a more careful statement of the precise type of bitheoretic version of Literal G2 we're in-

[1] Somewhat more accurately, they are conditions placed on the formulae expressing the notion of provability—or derivability-in-T—from which the consistency formulae are to be defined.

terested in a little later. For the moment we wish only to note that the literal version of G2 that we will ultimately be interested in is different from Literal G2.

This said, let us now say what we mean by 'interpreted' or 'proto-philosophical' versions of G2. These differ from literal versions in that they replace the formal conditions on the key representing formulae of the representing theory (*e.g.*, the conditions DC1–DC3 placed on the formulae $Prov_T(x)$ and Con_T) by conditions that 'interpret' these in terms of their importance to the proper representation of the metamathematical notions that the formulae involved are to represent (here, the notions of provability-in-T and consistency of T). Thus, a proto-philosophical or interpreted version of G2—and the one we're particularly interested in—is obtained by (i) taking DC1–DC3 as necessary conditions on the adequacy of $Prov_T(x)$ as an expression of the notion of provability-in-T and by (ii) taking the constraint that Con_T be defined as the formula $\forall x(Prov_T(x) \rightarrow \neg Prov_T(neg(x)))$ (or some similar formula) as a necessary condition on the adequacy of Con_T as a representation of T's consistency, *given that* $Prov_T(x)$ adequately expresses the notion of provability-in-T.

Combining these 'interpretations' of the conditions appealed to in Literal G2, we arrive at the following proto-philosophical version of G2.

Phil G2: Let $Prov_T(x)$ be a formula of $\mathcal{L}_T$ that expresses the notion of provability-in-T, and let Con_T be a formula of $\mathcal{L}_T$ that is constructed from $Prov_T(x)$ in such a way that if $Prov_T(x)$ expresses the notion of provability-in-T, then Con_T expresses the notion of T's consistency. Then, if T is consistent and its logic supports certain inferences and theorems, $\nvdash_T Con_T$.

The reader will recognize Phil G2 as a somewhat more careful version of the usual type of informal statement of G2 found in the logical and philosophical literature, including the statement by Bernays quoted above. Again, we call it a 'proto-philosophical' statement of G2 because, though it does not itself constitute a philosophical application of G2, it is the type of statement upon which such an application must be based. So, to illustrate, while Phil G2 does not itself state that G2 refutes Hilbert's Program, it is nonetheless the type of statement to which such an evaluation of Hilbert's Program must needs appeal. It is, in a word, what for philosophical purposes we might regard G2 as 'saying'.

Described thus, the notion of G2's proto-philosophical content is clearly related to the notion of a 'justification' for the Derivability Conditions mentioned earlier. The Derivability Conditions are, most directly, conditions on the choice of formulae to represent the notion of provability-in-T. A justification of a Derivability Condition will therefore consist in an argument establishing a necessary link between the given condition and the proper representation of the notion of provability for the represented theory. Taken

together, such arguments will constitute a larger argument establishing the following.

Pivotal Implication (PI): If $Prov_T(x)$ expresses the notion of provability-in-T in T, then $Prov_T(x)$ satisfies DC1–DC3.

This larger argument will involve analyzing the notions of provability-in-T and/or proper representation of provability-in-T in T to the point of revealing that they require satisfaction of DC1–DC3.

As mentioned above, our concern in this essay is with the justification of the Third Condition. It follows that we are concerned with the proper representation of the key metamathematical notions of proof and/or provability and consistency. But what is the conception of representation that figures here?

We can begin by noting that it is not merely a *semantical* conception. That is, it is not merely a conception according to which a formula achieves its representational end when its semantical interpretation produces an extension, and perhaps also an intension, that 'matches' (*modulo* the relevant encoding/decoding) those of the notion it is supposed to represent. It is rather what I would call an 'epistemic' conception of representation—that is, a conception according to which proper representation requires the representing entity to 'know' or 'grasp' or 'register' various facts concerning the notions it represents. On such a conception, and assuming that the representing device is a theory (hence a device which 'knows' or 'grasps' or 'registers' a given fact by *proving* it), the adequacy of a formula $\mathcal{F}$ as a representation of a set or a notion Φ is determined by what theorems involving $\mathcal{F}$ the representing theory proves. The success or failure of $\mathcal{F}$ as a representation of Φ thus consists in something other than a purely semantical relationship between $\mathcal{F}$ and Φ. It consists as well in a relationship between the 'facts' concerning Φ and the theorems involving $\mathcal{F}$ that the representing theory can prove.

The notion of representation with which we are concerned is therefore not one that is focused exclusively on the semantical interpretation (under an assumed interpretation of the language of the representing theory) of representing formulae. It is also one which distinguishes between the representation, of *sets*, on the one hand, and the representation of *concepts* or *notions*, on the other.

According to this distinction, a *set* Φ is represented by a formula $\mathcal{F}(x)$ in a theory T only if, for each c that qualifies as a possible candidate for membership in Φ, T proves $\mathcal{F}(c)$ (where 'c' is a recognized name for c in the language of T)[2] just in case c is an element of Φ.

[2] To give an exact (and compelling) account of what it should mean to say that a given term is a 'recognized' name for something in a given language is no easy matter. Since, however, the difficulties involved in doing so do not affect the project of this paper in any special way, we will not give such an account here.

Adequate 'intensional' (or 'intensionally adequate') representation of a *property*, *concept*, or *notion* Φ by a formula $\mathcal{F}(x)$ in a theory T seems to require something more than mere adequate representation of the set that is Φ's extension. It might, for example, require as well that T prove certain features of the 'logic' of Φ with respect to $\mathcal{F}(x)$. Or it might make use of a type of meta-condition requiring that certain facts concerning Φ's representation by $\mathcal{F}(x)$ *themselves* be registered as theorems of T.[3] This latter type of constraint figures centrally in the so-called Reflexivity Defense of the Third Condition.

Generally speaking, intensional representation conceives of a theory not merely as a set of theorems but as a set of theorems *given by* a certain concept or property. It therefore requires fidelity not only to the extension of the theory but also to the concept of provability by which it is given. For the most part, it is this intensional conception of theory with which we are concerned in this paper.[4]

The above remarks also suggest another distinction that is important for our discussion. This is the distinction between what we will refer to as the *representing* theory and the *represented* theory of a representational scheme. For a given metamathematical property or set Φ, the theory *to whose* metamathematics Φ pertains (the represented theory) need not be the same as the theory *in which* its representation is given (the representing theory). We want to consider what happens to the DCs and their justification when one makes systematic allowance for such a distinction between representing and represented theories.

There are two reasons why this is important to our discussion. The first is that it points up an element of unclarity in the usual 'monotheoretic' formulations of G2 (*e.g.*, that referred to above as 'Literal G2').[5] In such formulations, some of the references to T are references to it in its capacity *as representing theory* while others are references to it in its capacity *as represented theory*. The *justification* of the Derivability Conditions requires a clear demarcation of these roles. A justifiable constraint on the representing theory of a representational scheme can not generally be expected to be a justifiable constraint on the represented theory of that scheme, and *vice versa*. The justification of representational constraints therefore generally requires a distinction between the representing and represented theories of a representational scheme. In addition, we will argue, observance of the

[3] This type of constraint would hold if the notion of representation were 'internalist' in a certain sense—that is, if, in order to represent a given notion N, a representing formula $\mathcal{R}$ would not only have to register correctly the extension and certain features of the internal logic of N, but also have to 'see' itself as doing so.

[4] The intensional *vs.* extensional terminology was introduced in Feferman [1960]. See Feferman [1982] and [1989] for developments of the analysis begun there.

[5] By a 'monotheoretic' statement of G2, we mean a statement of it in which the representing and represented theories are the same theory.

representing/represented distinction can itself lead to the introduction of substantive additions to the usual conditions on intensional representation that figure in the proof of G2.

The second reason the representing *vs.* represented theory distinction is important for our purposes is that, as indicated above, certain applications of G2 require that we allow the two to be different. The particular application we have in mind is the application of G2 to the evaluation of Hilbert's Program. It requires that we allow the representing theory to become as weak as (some codification of) finitary reasoning while, at the same time, allowing the represented theory to be as strong as the strongest classical theory that possesses the type of instrumental virtues for which Hilbert generally prized classical mathematics (*e.g.*, various systems of set theory). If the G2 phenomenon were to hold only for some environments containing finitary reasoning, and not for all of them, it would not be legitimate to take it as refuting Hilbert's Program because it would not then be an invariant feature of all (proper) representational environments. Justifications of the DCs must therefore be valid not only in the monotheoretic setting but also in the appropriate bitheoretic settings.

We close this section with a final clarification. It concerns a certain relativization that seems to be built into the notion of representation, and which we must therefore expect to be reflected in any 'justification' of the DCs as representational constraints. This relativization consists in the fact that what can and should count towards accuracy of representation will generally be determined by the purpose or set of purposes for which a given representation is wanted. Talk of the justification of the DCs therefore presupposes a (set of) representational purpose(s) that is (are) to be achieved through their institution. It is not to be expected that all such purposes will call for the same constraints or even that they will include some decisive common core of them. To give definition to our investigation of the justification of the DCs, therefore, we must identify a set of purposes with respect to which representational adequacy is to be judged. For the sake of concreteness, we will take this purpose to be that of evaluating Hilbert's Program. At the same time, however, it should be noted that the argument given here is adaptable to a variety of other purposes as well.

3. The Reflexivity Defense of the Third Condition.

P.-G. Odifreddi states the idea behind the Reflexivity Defense of the Third Condition as follows.

> The first condition [DC1] was external to T, saying that any single provable formula can be recognized to be provable by T. This ... condition [*i.e.*, DC3] is internal to T, and says that T is aware of the first condition: *inside* T we know that if a formula is provable then we can prove this fact. (Odifreddi

[1989], p. 169, brackets and emphasis mine)[6]

In order to obtain a suitably clear and general statement of this defense, we must determine which references to T are essentially references to it in its role as *represented* theory (of the given arithmetization or representational scheme) and which are references to it in its role as *representing* theory. To this end, we offer the following restatement of Odifreddi's claim.

(RD-I): The first condition [DC1] is external to T, saying that any single formula provable in T can be recognized by T to be provable in T. This ... condition [*i.e.*, DC3] is internal to T, and says that T is aware of the first condition: inside T we know that if a formula is provable in T then we can prove this fact in T.[7]

Of the nine references to T in the above, the first, third, fifth, sixth, seventh, and ninth seem clearly to be references to T in its capacity as representing theory. Equally clearly, the second, fourth, and eighth are references to T in its capacity as represented theory.

Designating the representing theory of an arithmetization by 'S' and the corresponding represented theory by 'T', we can thus rewrite the above statement as follows.

(RD-II): The first condition is external to S, saying that any single formula provable in T can be recognized by S to be provable in T. The third condition is internal to S, and says that S is aware of the first condition: inside S we know that if a formula is provable in T then we can prove this fact in S.

A (the only?) plausible reading of Odifreddi yields this as the generalized (*i.e.*, bitheoretic) thesis of the Reflexivity Defense. We are left to our own devices to discover the deeper reasoning that is supposed to support such a view.

There seem to be two different directions in which to seek such support. On one of these—what we will refer to as the 'logical' variant—the Third Condition expresses a feature of the internal *logic* of the notion of representation. It maintains, that is, that in order for a formula $\mathcal{F}(x)$ of the

[6] Others also suggest this defense. See, for example, Smorynski [1977], p. 829, and Prawitz [1981], p. 261. It is not clear from the statements in Smornyski and Prawitz, however, whether they think of DC3 as a precept of the 'logic' of the concept of representation or as stemming from some other source. Odifreddi's statement is, of course, somewhat inaccurate. What T is 'aware of' when DC3 is satisfied is not DC1, but each of the several instances of DC1.

[7] This (and later reformulations) preserves the possible 'overstatement' of Odifreddi's original monotheoretic formulation. DC3 does not say that T knows *that* every formula provable in T is such that its provability-in-T is provable in T. Rather, it makes only the more 'local' claim that T knows *of* any formula of T that if it is provable in T, then T can prove this fact.

language of a theory S to *represent* a set or notion Φ of the metamathematics of T in S, it must not only be the case that, for all e, $e \in \Phi$ only if $\vdash_S \mathcal{F}(e)$ but also that S 'sees' this as holding. In other words, S's ability to represent Φ by $\mathcal{F}(x)$ requires not only a coordination of facts concerning Φ with 'beliefs' (*i.e.*, theorems) of S concerning $\mathcal{F}(x)$; it requires as well an 'awareness' or 'grasp' or 'registration' *of* this coordination *by* S. Without this latter, the thinking continues, S could not rightly be thought of as having the capacity to use what it 'believes' (*i.e.*, can prove) about $\mathcal{F}(x)$ to serve as a guide to facts concerning Φ. And without this type of self-reflective capacity on S's part, the defense concludes, S could not rightly be said to *represent* Φ by $\mathcal{F}(x)$.

On this account, then, representation is an inherently reflexive affair: in order for agent α to represent Φ by $\mathcal{F}(x)$, it is required not only that the instances of $\mathcal{F}(x)$ that α believes be 'true' of Φ. It is required as well that α herself grasp or believe in (the instances of) this correlation, by means of some concept she has of herself as a believing agent.

The other variant of the Reflexivity Defense—what we will call the 'evidentiary' variant—is based on the very different idea that in order for the First Condition to play its proper role in S's representation of provability-in-T by $Prov_T(x)$, the several instances of that condition must be verifiable by a certain type of evidence—evidence which, as it happens, is codifiable in S.[8] On this variant of the Reflexivity Defense, the Third Condition is not a consequence of the fact that S is the *representing* theory of the given representational scheme. Rather, it is a consequence of choosing the representing theory (S) to be a formalization of the type of evidence regarded as the proper standard for verification of the instances of the First Condition. This is a very different thing.

Of these two variants of the Reflexivity Defense, the logical variant seems the more basic. We do not insist upon this, however, since our purposes do not require it. We distinguish these two strains of the Reflexivity Defense only to give the reader an idea of its overall breadth and versatility, to inform her of the general fact that it signifies not a single justificatory idea but a (small?) family of such and, finally, to distinguish the Reflexivity Defense, as one broad strategic alternative regarding the justification of the Third Condition, from a very different type of defense that we will now briefly describe.

This alternative is what we call the *Strength Defense*. It sees the Third Condition as a special case of a deeper constraint—namely, the S-provable

[8] Here we are generally thinking of a bitheoretic formulation of the First Condition and not the monotheoretic formulation given in DC1. That is, we are generally thinking of the following condition:

Bi-DC1: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$, if $\vdash_T \mathcal{A}$, then $\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner)$.

Σ_1-completeness of T.[9] Moreover, it sees the justificatory idea behind this condition as entirely different from that underlying (either variant of) the Reflexivity Defense. Roughly, it is that S can not adequately represent T unless it 'sees' certain crucial facts concerning what T proves. Specifically, it can not hope to represent T adequately unless it not only sees of each true Σ_1 formula of the language of S that T proves it, but also sees of each Σ_1 formula σ of the language of S that if σ, then T proves σ.

To be interestingly different from the Reflexivity Defense, the Strength Defense has to be taken as based upon a different conception of the relationship between the First and Third Conditions than that which is assumed by the Reflexivity Defense. In other words, it can't be taken as the mere requirement that S 'see' the several instances of a reconceived version of the First Condition which states that S proves (a formula expressing) every true Σ_1 statement of informal arithmetic. The thought must rather be that certain elements of T's interior are so important to its identity that any good representation of it must see those elements as belonging to it. The challenge in defending such an idea, of course, is to say what it could be about *some* elements of T's interior that makes knowledge of their belonging to it more important to the representation of T than is the same knowledge with respect to other elements of its interior (or, for that matter, the parallel knowledge with respect to elements of T's exterior).

We do not believe that this challenge can be adequately met. Consequently, we do not believe that the Strength Defense provides a successful justification for the Third Condition. To argue for this belief, however, is not our concern here. We mention the Strength Defense for two reasons only. The first is to make clear to the reader that we are aware that there are alternatives to the Reflexivity Defense and that the present essay can not, by itself, be regarded as a conclusive general treatment of the justification of the Third Condition. The second is to point out the existence of bitheoretic treatments of G2 that do not require introduction of a formula $Prov_S(x)$ representing the notion of provability for the *representing* theory of a representational scheme.[10] This constitutes a major difference between the Reflexivity and Strength Defenses and to put the present discussion in proper perspective, it is important to bear this difference in mind.

[9] That is, the condition:

Bi-DC3#: For every Σ_1 sentence σ of $\mathcal{L}_S$, $\vdash_S \sigma \rightarrow Prov_T(\ulcorner \sigma \urcorner)$.

This is accurate, of course, only for theories of the type we take S to be—namely, theories containing an existential quantifier and capable of expressing a genuine notion of provability. For theories with no existential quantifier (*e.g.*, the usual formulations of PRA), hence no genuine notion of provability, a different condition is needed.

[10] On the Strength Defense, a sufficient set of conditions for (a bitheoretic form of) G2 would be Bi-DC1, Bi-DC2 and Bi-DC3#. Notice that, in contrast to the conditions coming from the Reflexivity Defense, no formula $Prov_S(x)$ representing the notion of provability-in-S appears anywhere in these conditions.

The difference becomes apparent when we consider the second clause of the Reflexivity Defense; the clause which says (as in RD-II) '[t]he third condition is internal to S, and says that S is aware of the first condition: inside S we know that if a formula is provable in T then we can prove this fact in S'. This is the clause in the Reflexivity Defense that requires introduction of a formula '$Prov_S(x)$' expressing (S's conception of) the notion of provability-in-S. It mandates a reflection by S on its own ability to detect and register the fundamental facts concerning the extension of the notion of provability-in-T.

The Reflexivity Defense thus suggests the following bitheoretic generalizations of DC1 and DC3.

Bi-DC1: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$, if $\vdash_T \mathcal{A}$, then $\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner)$.

Bi-DC3: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$,
$\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_S(\ulcorner Prov_T(\ulcorner \mathcal{A} \urcorner) \urcorner)$.

The antecedent of Bi-DC3 (*i.e.*, the formula '$Prov_T(\ulcorner \mathcal{A} \urcorner)$') should express the antecedent of Bi-DC1 (*i.e.*, the metatheoretic statement '$\vdash_T \mathcal{A}$') and its consequent should express the consequent of Bi-DC1. Since this latter is the metatheoretic statement '$\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner)$', it follows that the consequent of the Third Condition should be a formula '$Prov_S(\ulcorner Prov_T(\ulcorner \mathcal{A} \urcorner) \urcorner)$', where '$Prov_S(x)$' expresses '$\vdash_S$'.

Generally speaking, we will use the letters 'S' and 'T' to stand for sets of sentences—sets of sentences which form the extensions of theories. For convenience sake, however, we will generally speak of S and T as 'theories' rather than 'extensions of theories'. Also, in order to capture the type of relationship between S and T that we are generally interested in, we will assume that the language of S is a sublanguage of the language of T and that S is a subtheory of T ('$S \subseteq T$', in our notation).[11] We will argue that the conditions required for production of the G2 phenomenon in this type of bitheoretic setting differ significantly from those required for its production in the monotheoretic setting.

To this end, we will now consider what happens to G2 and its proof when, as per the dictates of the Reflexivity Defense, the forms of the First and Third Conditions used are those given in Bi-DC1 and Bi-DC3.

4. Effects of the Reflexivity Defense on the Proof of G2

To secure a proof of a bitheoretic version of G2 (Bi-G2) under the Reflexivity Defense of the Third Condition, we need not only Bi-DC1 and Bi-DC3, but also the following bitheoretic modifications of the DL and DC2.

Bi-DL: There is a sentence $\mathcal{G}$ of $\mathcal{L}_S$ such that $\vdash_S \mathcal{G} \leftrightarrow \neg Prov_T(\ulcorner \mathcal{G} \urcorner)$.

Bi-DC2: For all sentences $\mathcal{A}$, $\mathcal{B}$ of $\mathcal{L}_T$,
$\vdash_S Prov_T(\ulcorner \mathcal{A} \rightarrow \mathcal{B} \urcorner) \rightarrow (Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner \mathcal{B} \urcorner))$.

[11] We do not, however, assume that all proofs of S are proofs of T.

In addition, we need a Fourth Condition that assures S 'access' to the relationship (*viz.*, $S \subseteq T$) which, in the hypothesis of our bitheoretic version of G2, we assume to exist between S and T. Specifically, we need

Bi-DC4: For all sentences $\mathcal{A}$ of $\mathcal{L}_S$, $\vdash_S Prov_S(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner \mathcal{A} \urcorner)$.

With Bi-DL and Bi-DC1–Bi-DC4 at our disposal, we can prove the core lemma needed for the proof of (a bitheoretic version of) G2—namely

Bi-G2 Lemma: Let $S \subseteq T$ and let $Prov_T(x)$ and $Prov_S(x)$ be formulae of $\mathcal{L}_S$ that satisfy Bi-DC1–Bi-DC4 and Bi-DL. In addition, let Con_T be the formula $\forall x(Prov_T(x) \rightarrow \neg Prov_T(neg(x)))$ of $\mathcal{L}_S$. Then, if the logic of S supports certain inferences and theorems (made clear in the proof below), $\vdash_S Con_T \rightarrow \mathcal{G}$.

Proof (description of how to build a proof in S of '$Con_T \rightarrow \mathcal{G}$'):

(1) $\vdash_S \neg\mathcal{G} \rightarrow \neg\mathcal{G}$	Logic of S
(2) $\vdash_S \neg\mathcal{G} \rightarrow Prov_T(\ulcorner \mathcal{G} \urcorner)$	(1), Bi-DL, logic of S
(3) $\vdash_S Prov_T(\ulcorner \mathcal{G} \urcorner) \rightarrow Prov_S(\ulcorner Prov_T(\ulcorner \mathcal{G} \urcorner) \urcorner)$	Bi-DC3
(4) $\vdash_S \neg\mathcal{G} \rightarrow Prov_S(\ulcorner Prov_T(\ulcorner \mathcal{G} \urcorner) \urcorner)$	(2), (3), logic of S
(5) $\vdash_S \neg\mathcal{G} \rightarrow Prov_T(\ulcorner Prov_T(\ulcorner \mathcal{G} \urcorner) \urcorner)$	(4), Bi-DC4
(6) $\vdash_S \mathcal{G} \leftrightarrow \neg Prov_T(\ulcorner \mathcal{G} \urcorner)$	Bi-DL
(7) $\vdash_S Prov_T(\ulcorner \mathcal{G} \leftrightarrow \neg Prov_T(\ulcorner \mathcal{G} \urcorner) \urcorner)$	(6), $S \subseteq T$, Bi-DC1
(8) $\vdash_S Prov_T(\ulcorner \neg\neg Prov_T(\ulcorner \mathcal{G} \urcorner) \leftrightarrow \neg\mathcal{G} \urcorner)$	(7), $Prov_T$-$\leftrightarrow$ contraposition
(9) $\vdash_S Prov_T(\ulcorner \neg\neg Prov_T(\ulcorner \mathcal{G} \urcorner) \rightarrow \neg\mathcal{G} \urcorner)$	(8), $Prov_T$-$\leftrightarrow$ simplification
(10) $\vdash_S Prov_T(\ulcorner \neg\neg Prov_T(\ulcorner \mathcal{G} \urcorner) \urcorner) \rightarrow Prov_T(\ulcorner \neg\mathcal{G} \urcorner)$	(9), Bi-DC2
(11) $\vdash_S Prov_T(\ulcorner Prov_T(\ulcorner \mathcal{G} \urcorner) \urcorner) \rightarrow Prov_T(\ulcorner \neg\mathcal{G} \urcorner)$	(10), $Prov_T$-$\neg\neg$-intro.
(12) $\vdash_S \neg\mathcal{G} \rightarrow Prov_T(\ulcorner \neg\mathcal{G} \urcorner)$	(5), (11), logic of S
(13) $\vdash_S \neg\mathcal{G} \rightarrow (Prov_T(\ulcorner \mathcal{G} \urcorner) \,\&\, Prov_T(\ulcorner \neg\mathcal{G} \urcorner))$	(6), (12), logic of S
(14) $\vdash_S (Prov_T(\ulcorner \mathcal{G} \urcorner) \,\&\, Prov_T(\ulcorner \neg\mathcal{G} \urcorner)) \rightarrow \neg Con_T$	Def. of Con_T, logic of S
(15) $\vdash_S \neg\mathcal{G} \rightarrow \neg Con_T$	(13), (14), logic of S
(16) $\vdash_S Con_T \rightarrow \mathcal{G}$	(15), logic of S

□

Bi-G2 Lemma together with

Bi-G1.1: Let S be a consistent theory such that $S \subseteq T$, and let $Prov_T(x)$ and $\mathcal{G}$ be formulae of $\mathcal{L}_S$ which satisfy Bi-DL and Bi-DC1. Then $\nvdash_S \mathcal{G}$.

then yields the following generalized 'literal' version of G2

Bi-G2: Let $S \subseteq T$ and let $Prov_T(x)$ and $Prov_S(x)$ be formulae of $\mathcal{L}_S$ that together satisfy Bi-DC1–Bi-DC4 and the Bi-DL. In addition, let Con_T be the formula $\forall x(Prov_T(x) \rightarrow \neg Prov_T(neg(x)))$ of $\mathcal{L}_S$. Then, if S is consistent and the logic of S supports certain inferences and theorems, $\nvdash_S Con_T$.

These are the basic results with which we will be concerned. Before passing to our main argument, however, we'd like to call attention to a certain feature of the proof of Bi-G2 Lemma—specifically, the segment composed of lines (3)–(5). What we would like to note is that it is in lines (3)–(5) that Bi-DC3 and Bi-DC4 do their work. In effect, they take us from an application of Bi-DC3 to what is essentially an S-theoretic version of DC3, namely

Bi-DC3Δ: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$,

$$\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner Prov_T(\ulcorner \mathcal{A} \urcorner) \urcorner).$$

It is Bi-DC3Δ that is critical to the proof of Bi-G2 Lemma. Bi-DC3 (*i.e.*, the version of the Third Condition yielded by the Reflexivity Defense) thus gains its effect in the proof of Bi-G2 Lemma by being supplemented by a condition (*viz.*, Bi-DC4) that allows it to be extended to Bi-DC3Δ.

We point this out to guard against a possible misunderstanding of our employment of Bi-DC3. We use Bi-DC3 rather than Bi-DC3Δ as our variant of the Third Condition because our interest is in the Reflexivity Defense. Given Bi-DC1 as the form in which we have the First Condition, it is Bi-DC3 rather than Bi-DC3Δ that is *justified* by the Reflexivity Defense. Hence, if our proof of Bi-G2 is to fit the Reflexivity Defense, it must derive Bi-DC3Δ from Bi-DC3; that is, it must see Bi-DC3 rather than Bi-DC3Δ as the basic form of the Third Derivability Condition. This despite the fact that it is Bi-DC3Δ rather than Bi-DC3 that is crucial to the *proof* of Bi-G2. The Reflexivity Defense is therefore what necessitates the detour through Bi-DC3 and the extra condition Bi-DC4 in order to obtain Bi-DC3Δ. As we will see, the use of this extra condition is not without consequence.

5. The Reflexivity Defense and the Proto-philosophical Content of G2.

Bi-G2 is thus the bitheoretic form of G2 provided by the Reflexivity Defense. Our question now is what proto-philosophical content Bi-G2 should be seen as sustaining? In particular, we would like to know whether it can plausibly be regarded as sustaining the following:

Bi-Phil G2: Let $S \subseteq T$ and let $Prov_T(x)$ be a formula of $\mathcal{L}_S$ that expresses the notion of provability-in-T in S and Con_T a formula of $\mathcal{L}_S$ that is constructed from $Prov_T(x)$ in such a way that if $Prov_T(x)$ expresses the notion of provability-in-T in S, then Con_T expresses the notion of T's consistency in S. Then, if S is consistent, and the logic of S supports certain inferences and theorems, $\nvdash_S Con_T$.

We believe that Bi-Phil G2 is what is commonly regarded as the proto-philosophical content of Bi-G2. We will argue that it can not plausibly be so regarded—at least not if the justification used for Bi-DC3 is the Reflexivity Defense. The reason is that the Reflexivity Defense does not seem to promote a plausible form of Pivotal Implication to support the inference from Bi-G2 to Bi-Phil G2.

What form of Pivotal Implication (PI) does the Reflexivity Defense provide? It is possible to distinguish at least five different elements involved in such a PI. What is perhaps the core element is:

(Element 1): If there is a formula $Prov_T(x)$ of $\mathcal{L}_S$, that expresses the notion of provability-in-T in S, there is a formula $Prov_S(x)$ of $\mathcal{L}_S$ such that $Prov_S(x)$ expresses the notion of provability-in-S (*i.e.*, the notion of provability-in-the-representing-theory) in S.

Element 1 represents what we think is a common belief concerning the capacities of the representing theories that figure in our discussion: namely, that if they are capable of representing the notion of provability for the represented theory then they are capable of representing their own notion of provability. This is one of two key elements of the Reflexivity Defense of Bi-DC3. The other is:

(Element 2): For any formulae $Prov_T(x)$ and $Prov_S(x)$ of $\mathcal{L}_S$, if $Prov_T(x)$ expresses the notion of provability-in-T in S and $Prov_S(x)$ expresses the notion of provability-in-S in S, then $Prov_S(x)$ and $Prov_T(x)$ satisfy Bi-DC3.

In addition to these elements, we can assume that the Reflexivity Defense of Bi-DC3 is augmented by a defense of Bi-DC1 and Bi-DC2 that implies that:

(Element 3): For every formula $Prov_T(x)$ of $\mathcal{L}_S$, if $Prov_T(x)$ expresses the notion of provability-in-T in S, then $Prov_T(x)$ satisfies Bi-DC1 and Bi-DC2.

A defense of Bi-DC1 is indeed presupposed by the Reflexivity Defense. A defense of Bi-DC2 is not, but we will suppose for the sake of argument that there is such a defense and that it can be added to the Reflexivity Defense to form a defense of the larger set of conditions needed for the proof of Bi-G2 Lemma.

This leaves Bi-DC4 to consider. No justification of it is implied either by the Reflexivity Defense proper or by a defense of Bi-DC1 and Bi-DC2. It requires a justification of its own, one which says that if S is a subtheory of T, then S should be able to 'see' this with respect to its expressions of S and T. In other words, it should be the case that

(Element 4): For S and T such that $S \subseteq T$, if $Prov_S(x)$ and $Prov_T(x)$ are formulae of $\mathcal{L}_S$ that express the notions of provability-in-S and provability-in-T, respectively, then $Prov_S(x)$ and $Prov_T(x)$ together satisfy Bi-DC4.

Taken together, Elements 1–4 are the chief ingredients of our answer to the question regarding the form of Pivotal Implication that is to be provided by the Reflexivity Defense. One less central question remains.

It concerns the Bi-DL and its justification. So long as it is plausible to think that a formula representing the notion of provability-in-T will be a formula of one free variable, and so long as it is assumed that S is a fragment of arithmetic in which the numeralization, substitution and diagonalization functions are representable, it will be plausible to maintain the Bi-DL. These are all constraints that we are willing to grant the advocate of the Reflexivity Defense. Hence, we are willing to grant that

(Element 5): For any formula $Prov_T(x)$ of $\mathcal{L}_S$, if $Prov_T(x)$ expresses the notion of provability-in-T in S, then $Prov_T(x)$ satisfies Bi-DL.

The above five elements thus constitute the (expanded form of) Pivotal Implication sponsored by the Reflexivity Defense. For convenience's sake we now condense these five elements into the following two principles:

(Bi-PIΔ1): For S and T such that $S \subseteq T$ and formulae $Prov_T(x)$ and $Prov_S(x)$ of $\mathcal{L}_S$ that express (in S) the notions of provability-in-T and provability-in-S, respectively, $Prov_T(x)$ and $Prov_S(x)$ together satisfy Bi-DC1–Bi-DC4 and the Bi-DL.

(Bi-PIΔ2): If there is a formula $Prov_T(x)$ of $\mathcal{L}_S$ that expresses the notion of provability-in-T in S, then there is a formula $Prov_S(x)$ of $\mathcal{L}_S$ that expresses the notion of provability-in-S in S.[12]

We effect this condensation to highlight two importantly different types of conditions that figure in Elements 1–5. The one type, embodied in Bi-PIΔ1, identifies Bi-DL and Bi-DC1–Bi-DC4 as necessary conditions on the ability of formulae to express the notions of provability-in-T and provability-in-S. It comes from items 2–5 of the elements enumerated above. The other type, embodied in Bi-PIΔ2 takes S's ability to represent T as sufficient for its ability to represent itself.

We will not mount an independent challenge to either the first or second types of conditions in isolation.[13] Rather, we question their joint plausibility. In addition we would note that both seem to be necessary in order to secure Bi-Phil G2 as an entailment of Bi-G2. Specifically, Bi-PIΔ1, taken by itself, is not enough to secure the connection. Taken together with

[12] We understand Bi-PIΔ2 to assume that the represented theory T is recursively axiomatizable. This is in keeping with the idea in Hilbert's Program that the represented theory be treated as a formal object. This is not to deny, of course, that Hilbert's ideas might be extended to a larger class of represented theories along such lines, say, as those described in Schütte [1960] and [1977].

[13] We believe that both conditions can be challenged, however. In particular, we believe that neither Bi-DC3 nor Bi-DC4 is plausible.

Bi-G2, it entails only the very different proto-philosophical claim[14]

(Bi-Phil G2Δ): Let $S \subseteq T$ and let $Prov_T(x)$ and $Prov_S(x)$ be formulae of $\mathcal{L}_S$ that express in S the notions of provability-in-T and provability-in-S, respectively. Finally, let Con_T be a formula of $\mathcal{L}_S$ constructed from $Prov_T(x)$ in such a way that if $Prov_T(x)$ expresses the notion of provability-in-T in S, then Con_T expresses the notion of T's consistency in S. Then, if S is consistent, and the logic of S supports certain inferences and theorems, $\nvdash_S Con_T$.

If, therefore, Bi-Phil G2 is to be defended as the proto-philosophical reading of Bi-G2 under the Reflexivity Defense, then both Bi-PIΔ1 and Bi-PIΔ2 must be defended.

Bi-PIΔ2, moreover, requires a defense of a different type from that of Bi-PIΔ1. The reason is that it is a different type of condition. Instead of linking the ability of formulae to express the notions of provability-in-T and provability-in-S to their satisfaction of the Derivability Conditions, as Bi-PIΔ1 does, Bi-PIΔ2 links the ability of S to express the notion of provability-in-T to its ability to express the notion of provability-in-S. As we will presently see, such a linkage seems dubious.

Before developing this argument further, however, we want to say a little about our identification of Elements 1 and 2 as the core elements of the Reflexivity Defense. The reader may wonder why, instead of Elements 1 and 2, we did not take the following single claim as the core element of the Reflexivity Defense.

(Element 1Δ): For every formula $Prov_T(x)$ of $\mathcal{L}_S$, if $Prov_T(x)$ expresses the notion of provability-in-T in S, there is a formula $Prov_S(x)$ of $\mathcal{L}_S$ such that $Prov_S(x)$ expresses the notion of provability-in-S (*i.e.*, the notion of provability-in-the-representing theory) in S and $Prov_S(x)$ and $Prov_T(x)$ together satisfy Bi-DC3.

Element 1Δ is implied by Elements 1–2, but it does not in turn imply them. We mention this because certain of our criticisms of Bi-PIΔ1 and Bi-PIΔ2 apply directly only to the conjunction of Elements 1 and 2 and not to Element 1Δ. The question thus arises: Why take Elements 1 and 2 rather than Element 1Δ as constituting the core of the Reflexivity Defense?

The answer derives from our view of the structure of the Reflexivity Defense. We see it as saying that (i) if there is a formula $Prov_T(x)$ that expresses the notion of provability-in-T in S, then there is a formula $Prov_S(x)$ that expresses the notion of provability-in-S in S, and that (ii) *any* pair

[14] There is actually more that's required. One needs a premise to the effect that if Con_T is a formula of $\mathcal{L}_S$ constructed from $Prov_T(x)$ in such a way that if $Prov_T(x)$ expresses the notion of provability-in-T, then Con_T expresses the notion of T's consistency, where Con_T is the formula $\forall x(Prov_T(x) \rightarrow \neg Prov_T(neg(x)))$, or some formula that is S-equivalent to it. Having noted this, however, we will, for simplicity's sake, suppress mention of this condition in the proto-philosophical statements of G2.

of such formulae—that is, any pair of formulae expressing provability-in-T and provability-in-S respectively—will satisfy Bi-DC3. In other words, we see the Reflexivity Defense as saying that Bi-DC3 expresses a condition on the proper representation of provability-in-T and provability-in-S by *any* formulae of $\mathcal{L}_S$. We therefore take the following claim to follow from the Reflexivity Defense.

(Element 1†): For every formula $Prov_T(x)$ of $\mathcal{L}_S$, if $Prov_T(x)$ expresses the notion of provability-in-T in S, there is a formula $Prov_S(x)$ of $\mathcal{L}_S$ such that $Prov_S(x)$ and $Prov_T(x)$ together satisfy Bi-DC3.

In particular, and unlike the defender of Element 1Δ, we see this as following from Elements 1 and 2 and, so, from the ability of $Prov_S(x)$ and $Prov_T(x)$ to serve as genuine expressions of provability-in-S and provability-in-T.

One who takes Element 1Δ to be the core element of the Reflexivity Defense would have to see its structure in a different way. She would have to reason thusly: for every $Prov_T(x)$ that expresses the notion of provability-in-T in S, there is a $Prov_S(x)$ which both expresses the notion of provability-in-S in S *and* which has the unrelated auxiliary property that, when taken together with $Prov_T(x)$, it satisfies Bi-DC3. Seen this way, Bi-DC3 would be a purely contingent product of certain formulae that express the notion of provability-in-S. It would not be taken as a necessary condition on the ability of a pair of formulae to express provability-in-T and provability-in-S, respectively.

We don't find this a plausible interpretation of the Reflexivity Defense—or any other defense of Bi-DC3, for that matter. Accordingly, we will not consider it further. We mention it only to explain to the reader why we take Elements 1 and 2 rather than Element 1Δ as the core of the Reflexivity Defense. This done, we now turn our attention to Bi-PIΔ1 and Bi-PIΔ2, where we will argue that they are jointly implausible.

Our arguments are of three types. All are intended to call Bi-PIΔ2 into question *given* the conditions on proper representation laid down in Bi-PIΔ1. In the first argument the condition featured is Bi-DC1. In the second and third arguments, the focal condition is Bi-DC4—the 'extra' condition made necessary by the generalization of the Reflexivity Defense to bitheoretic settings. The arguments are strategically related.

The first accepts the general idea—suggested by the use of Bi-DC1 in Bi-PIΔ1—that proper representation of a set σ in S requires the enumeration in S of σ.[15] It then observes the lack of any verified, general connection

[15] For the reader who may not be familiar with the terminology, a set of n-tuples of numbers θ is said to be *weakly represented* in S by the formula $\tau(x_1, \ldots, x_n)$ just in case for every n-tuple of numbers $\langle \mathsf{k}_1, \ldots, \mathsf{k}_n \rangle$, $\langle \mathsf{k}_1, \ldots, \mathsf{k}_n \rangle \in \theta$ iff $\vdash_S \tau(k_1, \ldots, k_n)$, where k_i is a canonical term in S for k_i. θ is weakly representable in S just in case there is some formula that weakly represents it in S. Enumeration is just weak representation

between S's ability to enumerate a recursively axiomatizable theory T and its ability to enumerate itself if its arithmetic type (*i.e.*, its place in the arithmetic hierarchy) is different from T's. Given that there are legitimate questions concerning the formalizability of finitary reasoning (*i.e.*, its exact codification into a recursively axiomatizable theory), this serves at least to raise questions concerning the joint general plausibility of Bi-PIΔ1 and Bi-PIΔ2 for cases of S assumed to contain finitary reasoning.

It only raises a question, though, and doesn't settle anything. Our second argument therefore seeks to go beyond this by locating a weakness in Bi-PIΔ1 and Bi-PIΔ2 that rests upon something more than the mere uncertainty of the formalizability of finitary reasoning. It therefore grants the formalizability of finitary reasoning as a strategic concession and goes on to show that even if this is assumed there is ample room to doubt the plausibility of Bi-PIΔ2 given the use (in Bi-PIΔ1) of Bi-DC4 as a necessary condition on the proper representation of the notions of provability-in-S and provability-in-T. The argument is that Bi-DC4 does not generally hold for pairs of formulae expressing the *notions* or *concepts* of provability-in-S and provability-in-T for S and T such that $S \subseteq T$. The conclusion is that Bi-PIΔ1 and Bi-PIΔ2 can not both be generally maintained as conditions governing the representation of concepts.

The third argument seeks to take the claim of the second argument beyond the level of concept representation to the (more basic?) level of set representation. It argues that Bi-DC4 does not generally hold for pairs of recursively enumerable *sets* S and T such that $S \subseteq T$. If correct, it shows that Bi-PIΔ1 and Bi-PIΔ2 can not be maintained as conditions governing the representation of sets—even recursively enumerable sets—generally.

This, in outline, is our argument. We now proceed with the details.

First Argument

We begin by considering the case where T is a recursively axiomatizable theory and S a subtheory of T, though not a recursively axiomatizable one. The chief fact to be borne in mind here is that S's having the ability to enumerate Σ_1 or recursively enumerable sets (*i.e.*, sets of the arithmetic type that T is assumed to be) does not guarantee it the ability to enumerate sets of other arithmetic types (*e.g.*, sets of type Σ_n, $n > 1$). If it is granted that S's overall ability to represent or express itself requires that it enumerate itself,[16] it then follows that S's ability to represent T does not

in the left-to-right direction. Hence, $\tau(x_1, \ldots, x_n)$ *enumerates* θ in S just in case, for every n-tuple of numbers $\langle k_1, \ldots, k_n \rangle$, $\langle k_1, \ldots, k_n \rangle \in \theta$ only if $\vdash_S \tau(k_1, \ldots, k_n)$. θ is *enumerable* in S just in case there is a formula that enumerates it in S.

[16] The advocate of the inference from Bi-G2 to Bi-Phil G2 has to grant this; otherwise, she loses her *justification* for Bi-DC1. The justification for Bi-DC1 is general. That is, it assumes that for *any* set or property Φ and *any* formula ϕ of the language of S, ϕ represents Φ in S only if it enumerates Φ in S.

imply a similar ability on its part to represent itself. That being the case, Bi-PIΔ2 would appear to be groundless ... at least insofar as there are serious reasons for believing that S might be something other than a Σ_1 or recursively enumerable theory. It is to a consideration of such reasons that we now turn.

Our reasons stem from a closer consideration of finitary reasoning and of what can be regarded as its 'internal' structure—a structure which, it seems, is suggested by a suitably refined understanding of the nature of finitary reasoning, albeit one the likes of which one seldom finds mentioned in the literature.[17]

The structure arises from the fact that not all finitary evidence is on the same epistemological footing. In particular, there is a type of finitary evidence that functions as the 'data' of finitary thought and another type that can be thought of as less basic. The former is expressed by those statements that are decidable solely by means of finitary judgments concerning particular finitary objects (or finite assemblages thereof). In the context of a first-order arithmetic language, these are the statements expressed by variable-free sentences—the statements making up the decidable fragment of first-order arithmetic.

The latter is expressed by sentences containing variables; sentences which serve not as judgments proper but as 'judgment-schemata' or 'judgment-forms'. These are devices which become genuine judgments or assertions when, and, in Hilbert's view, only when, numerals or other closed terms are substituted for the variables that occur in them. Despite their schematic status, these judgment-schemata were treated by Hilbert as capable of some type of 'acceptance' (resp. 'rejection') by finitary reasoners; namely, that which is constituted by a disposition to affirm each of their instances.

Just what might serve as the rational basis for such a disposition, how a *disposition* to assert all instances of a schema might differ from a simple brute capacity to assert them all, and how we might come to be in possession of such a disposition are points on which Hilbert said nothing of significance.

This notwithstanding, the acceptability of a finitary schema should be seen as depending upon its compatibility with the more basic propositions of finitary thought—the so-called singular judgments or propositions. This induces an internal epistemic structure among finitary judgments. In addition, finitary schemata might admit of an internal structuring that represents an ordering of relative acceptability among them. Such an ordering could in part arise from strict differences in relative logical strength (*i.e.*, from the fact that one schema is strictly stronger than another). It could also arise from and reflect non-logically based differences—epistemic differences between distinguishable types of evidence all of which count as

[17] A partial exception is Smorynski ([1988], pp. 38–40).

finitary. There is nothing in the traditional formulations of finitism to prohibit such differences. Indeed, they could result naturally from taking a realistic approach to differences in the relative complexity of different finitary objects. None of the usual formulations of finitism (nor any plausible formulation of it) institutes homogeneity constraints so powerful as to prohibit such differences.

To the extent that such a view of finitism is correct, finitary reasoning could easily possess an internal epistemic structure that is not itself that of a Σ_1 theory. Indeed, such an internal structure is suggested by the following general condition on finitary provability: $\mathcal{A}$ is finitarily provable just in case (i) $\mathcal{A}$ is an acceptable (*i.e.*, finitarily knowable) singular proposition, (ii) $\mathcal{A}$ is a schema and it can be known finitarily that no instance of $\mathcal{A}$ is either the denial of or is denied by an acceptable singular proposition, and (iii) $\mathcal{A}$ is a schema and it can be known finitarily that $\mathcal{A}$ neither denies nor is the denial of a finitary schema that occupies a place of greater elementariness in the ordering of finitary schemata according to their relative plausibility. Clauses (ii) and (iii) would appear to place a type of consistency constraint on even the choice of finitary *axioms*. Assuming that this could be part of an intensionally correct description of the internal structure of finitary reasoning, and assuming that the logic used to execute the consistency check is subject to Church's theorem, it would seem to follow that neither the notion of axiom nor the notion of proof for finitary reasoning is necessarily a recursive one.

The above notwithstanding, let us hasten to add that we do not know of a way to demonstrate that the full set of finitarily acceptable statements and schemata can not be formalized by *some* recursive set of formulae. The fact that, modulo its epistemically most natural or basic description, the structure of the finitarily acceptable propositions and proposition-schemata is not evidently Σ_1 does not imply that there is no recursive axiomatization of it. Our claim is therefore only that there is room to doubt that finitary reasoning is adequately formalizable as a Σ_1 theory. To the extent that this is correct, there is room to doubt that all cases of S in which we are interested can safely be assumed to be Σ_1 theories. Furthermore, to the extent that this is so and S's ability to represent itself requires its ability to enumerate itself, it is not clear that S should have the ability to represent itself just because it has the ability to represent a Σ_1-theory T. In other words, it is not clear that Bi-PIΔ2 holds.

We realize that the above argument is somewhat vague and inconclusive and that it is essentially an argument from what we don't know. It would be nice to have something more precise and conclusive to offer in its place. In particular, it would be nice to have a way of judging the plausibility of Bi-PIΔ2 regardless of what the exact arithmetic type of S might turn out to be. To provide that, however, we would have to have either

a better characterization of the ways in which a Σ_n theory might achieve Σ_n-completeness or a better description of the ways (if any) in which a Σ_n theory might achieve self-enumeration without being Σ_n-complete.

Such characterizations would, in turn, require a fuller specification of the set of the conditions necessary for S's representation of T (modulo some specified foundational purpose) as well as an analysis linking these conditions to S's ability to enumerate itself. At present, however, neither of these seems to be available. What we know is that Σ_1 extensions of Q enumerate all Σ_1 sets and therefore enumerate themselves. In addition, we know that all Σ_n-complete Σ_n extensions of Q enumerate themselves. What we do not know (or at least what I do not know)—either for n in general or for specific cases of n—is a useful characterization (*i.e.*, a characterization related to a specification of conditions known to be necessary for the representation of T by S) of which, if any, Σ_n extensions of Q *do not* contain all Σ_n truths but *do* enumerate themselves. A significant problem is therefore to ...

(Problem 1): Characterize those Σ_n extensions of Q (if any) which enumerate themselves but which are not themselves Σ_n-complete.

Second Argument

We will now set the preceding argument aside and suppose that it is proper to restrict both S and T to Σ_1 or recursively axiomatizable theories. We thus consider the following restriction of Bi-PIΔ2.

(Bi-PIΔ2*): For recursively axiomatizable S and T, if there is a formula $Prov_T(x)$ of $\mathcal{L}_S$ that expresses the notion of provability-in-T in S, then there is a formula $Prov_S(x)$ of $\mathcal{L}_S$ that expresses the notion of provability-in-S in S.

Perhaps the first thing to note in this connection is that Bi-G2 itself requires no such restriction. Hence, strictly speaking, the restricted form of Bi-Phil G2 obtainable from Bi-G2 via Bi-PIΔ2* can not be viewed as giving the proto-philosophical content of the full Bi-G2 but only the proto-philosophical content of a narrower theorem which results from restricting the choice of S in Bi-G2 to recursively axiomatizable theories. This notwithstanding, we will consider the plausibility of Bi-PIΔ2* and grant, for the sake of argument, that it is the narrower theorem just mentioned that ought to be our chief concern.

We will therefore consider Bi-PIΔ2* and its justification. Throughout our discussion, however, it will be important to keep in mind a distinction that was noted earlier. This is the distinction between the representation of *concepts* or *notions*, on the one hand, and the representation of *sets*, on the other. As formulated, both Bi-PIΔ2* and Bi-PIΔ2 concern, at least on their surfaces, the representation of *concepts* or *notions* and not (simply) the representation of *sets*. We will argue that Bi-PIΔ2* is unjustified if, as is asserted by Bi-PIΔ1, Bi-DC4 is taken to be a necessary condition on the

adequate representation of the *concepts* of provability-in-S and provability-in-T. In other words, we will argue that Bi-DC4 does not appear to be a legitimate constraint to place on the proper representation of the *concepts* of provability-in-S and provability-in-T.

Various examples can be used to illustrate this point. The following simple one suits our purposes well enough.

Let S and M be recursively axiomatizable theories in $\mathcal{L}_S$ such that $S \subseteq M$. *Define* T to be $\{\tau : \tau \in M \& \tau \neq \nu\}$, where ν is a sentence of $\mathcal{L}_S$ that is not a theorem of M. In addition, let $Prov_S(x)$ be a formula of $\mathcal{L}_S$ that expresses the notion of provability-in-S and $Prov_M(x)$ be a formula of $\mathcal{L}_S$ that expresses the notion of provability-in-M. Finally, let $Prov_T(x)$ be the formula $Prov_M(x) \& x \neq n$, where 'n' is the numeral in $\mathcal{L}_S$ for the gödel number n of the sentence ν.

We claim that $Prov_T(x)$ expresses the *concept* of provability-in-T adequately; that is, it follows and reflects the *definition* of T in the way that a formula expressing the *concept* of provability-in-T ought to. Furthermore, since $S \subseteq M$, the definition of T, which guarantees that $M \subseteq T$, also guarantees that $S \subseteq T$.

On the other hand, however, Bi-DC4 does not seem to hold for $Prov_S(x)$ and $Prov_T(x)$. In particular, the instance '$Prov_S(n) \rightarrow Prov_T(n)$' does not seem to hold. The reasoning is as follows. If we assume, as above, that $Prov_S(x)$ is a formula of $\mathcal{L}_S$ that expresses the notion of provability-in-S, and we assume that the monotheoretic version of G2 holds for S (for $Cons_S$ defined in the usual way from $Prov_S(x)$), and we assume that the logic of S preserves certain principles and inferences that we may assume it to preserve, then we may conclude both that

$$(1) \qquad \nvdash_S \neg Prov_S(n).$$

and that

$$(2) \qquad \vdash_S n = n.$$

From (2) and the definition of $Prov_T(x)$, it then follows that

$$(3) \qquad \vdash_S \neg Prov_T(n).$$

However, from (3) and the hypothesis that

$$(4) \qquad \vdash_S Prov_S(n) \rightarrow Prov_T(n),$$

it follows that

$$(5) \qquad \vdash_S \neg Prov_S(n).$$

Assuming the consistency of S, then, the hypothesis in (4)—which is an instance of Bi-DC4—can not be accepted.

It follows that if the standard for the representation of *concepts* articulated in Bi-PIΔ1 is accepted—specifically, that Bi-DC4 is accepted as a necessary condition on the ability of $Prov_S(x)$ and $Prov_T(x)$ to express the *concepts* of provability-in-S and provability-in-T—then Bi-PIΔ2* can not generally be accepted as a condition on the adequate representation of provability concepts.[18] From this it follows in turn that the inference from Bi-G2 to Bi-Phil G2 can not be accepted, and this means that Bi-G2 can not rightly be regarded as 'saying' Bi-Phil G2.

It would appear, then, that there is a problem concerning the justification of Bi-DC4. Specifically, if $\mathfrak{Prov}_S$ and $\mathfrak{Prov}_T$ are pre-arithmetic, informal provability concepts that present the theorem-sets S and T respectively,[19] then the following principle can not be used to justify it.

(Proto-Bi-DC4): If S and T are theories such that $S \subseteq T$, and $\mathfrak{Prov}_S$ and $\mathfrak{Prov}_T$ are the *concepts* by which S and T (respectively) are intensionally given, then if $Prov_S(x)$ and $Prov_T(x)$ are to represent, respectively, $\mathfrak{Prov}_S$ and $\mathfrak{Prov}_T$ in S, Bi-DC4 must hold of $Prov_S(x)$ and $Prov_T(x)$.

Nor can the more general standard of concept representation—a standard that applies to concepts generally, and not just to provability concepts—that stands behind Proto-Bi-DC4 be accepted. According to this more general standard, if the formulae $\mathcal{F}$ and $\mathcal{G}$ of a given representing theory S are to represent the informal concepts $\mathcal{C}_\mathrm{F}$ and $\mathcal{C}_\mathrm{G}$, whose extensions are F and G, respectively, S must capture or express the subset relations that obtain between F and G as theorems involving $\mathcal{F}$ and $\mathcal{G}$. Somewhat more precisely:

(Gen-Proto-Bi-DC4): If S and T are sets such that $S \subseteq T$, and $\mathcal{C}_S$ and $\mathcal{C}_T$ are the *concepts* by which S and T are intensionally given, then, if the formulae $C_S(x)$ and $C_T(x)$ are to adequately represent or express $\mathcal{C}_S$ and $\mathcal{C}_T$ in a theory θ, it must be the case that for every n, $\vdash_\theta C_S(n) \rightarrow C_T(n)$ (where 'n' is a term in the language of θ that is acknowledged to be a designator of n).

As the case constructed earlier shows, Gen-Proto-Bi-DC4 can not be

[18] I have benefited from discussions of this and related matters (*i.e.*, conditions properly regarded as governing the representation of *concepts*) with George Boolos, Julia Knight, Mike Stob and Peter Cholak.

[19] In calling $\mathfrak{Prov}_S$ and $\mathfrak{Prov}_T$ 'pre-arithmetic' provability concepts presenting S and T, I mean only that they are the provability concepts of the theories S and T as they (the theories and their provability concepts) are given prior to arithmetization. They are therefore the concepts that the formulae $Prov_S$ and $Prov_T$ are supposed to represent in S.

accepted as a generally valid constraint on the representation of concepts. The formulae $Prov_S(x)$ and $Prov_T(x)$ of that case are acceptable expressions of the pre-arithmetic concepts of provability-in-S and provability-in-T defined there, but they do not satisfy the appropriate instance of Gen-Proto-Bi-DC4.

We conclude that it is not generally necessary that formulae expressing or representing concepts in a given theory S should instance-wise capture (as theorems of the representing theory) the subset relations that obtain between the extensions of those concepts. There are, I believe, two basic facts which account for this. The first is that for a formula $\mathcal{F}$ to serve as a proper representation of a concept $\mathcal{C}$ in a theory θ, what is necessary is that the *defining* characteristics of $\mathcal{C}$ be registered in θ as theorems concerning $\mathcal{F}$. The second is that not everything that may be known or proved about given concepts (including facts regarding their extensions) by those ordinary users of the concepts who grasp them is rightly regarded as a defining feature of them.

Evidently, these are points of some subtlety since the following reasoning seems to be widely used without comment: (i) for many pairs of theories S, T, it is possible, using the pre-arithmetic informal provability concepts by which they are presented, to prove (in the informal metamathematics of S and T) that every theorem of the one is a theorem of the other; therefore, (ii) for such pairs of theories, it is proper to take Bi-DC4 as a necessary condition on the ability of the formulae $Prov_S(x)$ and $Prov_T(x)$ properly to represent those pre-arithmetic provability concepts in S. If the argument given above is correct, however, such an attitude is wrong; the inference from (i) to (ii) can not be accepted without further justification.

There are, I believe, two principal strategies that one might pursue in attempting to provide such justification. Each, however, appears to require solutions to some fairly difficult problems.

The first strategy concedes that Gen-Proto-Bi-DC4 does not hold for concepts generally and even that it does not hold for provability concepts generally (*i.e.*, it concedes that Proto-Bi-DC4 does not generally hold). It seeks, nonetheless, to make a case for Proto-Bi-DC4 as a restricted principle—one that holds for certain pairs of provability concepts. The motivating idea behind this strategy is that there seem to be cases where the subsumption of the extension of one concept by that of another is immediate from their definitions. In such cases, it may be reasonable to regard knowledge of such subsumption as partially constitutive of what is involved in grasp of the concepts.

Perhaps the clearest case of this type is one where the axioms and rules of inference in terms of which one of the provability concepts is defined are themselves explicitly included among the axioms and rules of inference in terms of which the other provability concept is defined.[20] In such cases, it may well be that sheer grasp (hence, sheer representation) of the provability concepts involved necessitates knowledge (or instance-wise knowledge) of the subsumption of the one concept's extension by that of the other. The problem, however, is that the statement, proof and applications of Bi-G2 are generally not restricted to such cases. This means that one advocating the usual interpreted version of Gödel's Second Theorem (*i.e.*, the Bi-Phil G2 reading of Bi-G2) must provide an argument that does not look to be easy to provide; specifically, she must

(Problem 2): Explain, for the full range of theories S and T such that $S \subseteq T$ and the full range of provability concepts $\mathbf{Prov}_S$ and $\mathbf{Prov}_T$ for which Bi-G2 can be proved, why a proper grasp of $\mathbf{Prov}_S$ and $\mathbf{Prov}_T$ requires knowledge (or instance-wise knowledge) that $S \subseteq T$, even though this is not *generally* required for grasp of concepts one of whose extensions subsumes that of the other.

In other words, the advocate of the traditional interpretation of Bi-G2 must say what it is about the provability concepts in Bi-G2 that makes knowledge (or instance-wise knowledge) of the subsumption relation that exists between their extensions a necessary condition of their proper grasp. I do not believe that a satisfactory solution to this problem can be given. Hence, I do not believe that there is anything about Bi-G2's holding of a pair of provability concepts $\mathbf{Prov}_S$ and $\mathbf{Prov}_T$ such that $S \subseteq T$ that justifies placing a condition like Bi-DC4 on formulae purporting to represent or express $\mathbf{Prov}_S$ and $\mathbf{Prov}_T$. What can be, and often is, true is that Bi-DC4 is *provable* for formulae representing such pairs of provability concepts. That Bi-DC4 may be provable, however, does not imply that it is justifiable. To prove Bi-DC4 does not show that it is a necessary condition on the ability of formulae of the language of S to properly *represent* or *express* the concepts $\mathbf{Prov}_S$ and $\mathbf{Prov}_T$.

Third Argument

On the basis of the above argument, then, we claim that Bi-PIΔ2*—and, hence, the usual interpretation of Bi-G2—can not be maintained when the criteria for representing the *concepts* of provability-in-T and provability-in-S are taken to include satisfaction of Bi-DC4. This notwithstanding, some might still argue that Bi-DC4 holds when $Prov_S(x)$ and $Prov_T(x)$ are formulae that express the provability concepts which 'naturally' (or perhaps 'canonically') present the theorem-sets T and S. It is this possibility that we will now consider.

The basic ideas behind this attempt to justify Bi-DC4 are that (i) *con-*

[20] As specific examples, consider the usual formal axiomatic definitions of the well known systems of first-order arithmetic Q and PA.

cepts are ways of presenting *sets* and that (ii) for any given *set*, though there may be a variety of intensionally non-equivalent concepts by which it can (in some sense) be presented, there is a single concept or a restricted class of concepts that present it *best*. We will refer to this supposed optimal presentation of a set as its 'preferred' or 'canonical' concept.

Given this idea of *canonically* presenting concepts for sets, the following might be offered as a general principle justifying Bi-DC4.

(Gen-Proto-Bi-DC4Δ): Where S and T are sets, and $\mathcal{C}(S)$ and $\mathcal{C}(T)$ are concepts that canonically present S and T, and $S \subseteq T$, it should be the case that for any representing theory θ and any formulae $C_S(x)$ and $C_T(x)$ of the language of θ, if $C_S(x)$ and $C_T(x)$ adequately represent $\mathcal{C}(S)$ and $\mathcal{C}(T)$ (respectively) in θ, then, for every n, $\vdash_\theta C_S(n) \rightarrow C_T(n)$ (where 'n' is a term in the language of θ acknowledged to be a designator of n).

Advocacy of Gen-Proto-Bi-DC4Δ, however, exacts a tribute; namely: (1) to identify a set of general characteristics of concepts that are those characteristics that are responsible for their being the canonical presentations of the sets they present; (2) to explain what it is about the general characteristics identified that should make concepts possessing them the 'preferred' or 'canonical' ways of presenting the sets that they present; and (3) to explain why it is that adequate representations of canonical concepts ought always to capture any subset relations that exist between their extensions.

None of these are easy tasks to manage. Indeed, they may not be manageable at all. Adoption of Gen-Proto-Bi-DC4Δ as the justificatory principle for Bi-DC4 is therefore not to be taken lightly. It brings heavy responsibilities with it. To see the kind of difficulties it encounters, consider the following example. Let S be the set of numbers x such that there is no odd perfect number greater than x. And let T be the empty set of numbers presented by whatever concept might be seen as the canonical concept presenting it (e.g., the concept of being non-self-identical). For all we know, no reasonable choice of representing theory θ will decide the question of whether there is an odd perfect number. At the same time, however, it can safely be assumed that a reasonable choice of θ will, for each n, prove '$n = n$', which we may assume to be (provably equivalent in θ to) '$\neg C_T(n)$'. If S is presented by the concept indicated above, then we may take $C_S(n)$ to be the formula '$\neg\exists x(x > n$ & Odd(x) & Perfect $(x))$'.[21] Hence, assuming, as above, that θ does not decide the question of whether there is an odd perfect number, it follows that θ can not prove '$C_S(n) \rightarrow C_T(n)$' for any n.

[21] Here, of course, 'Odd(x)' and 'Perfect(x)' are short for longer, more complicated formulae of first-order arithmetic.

If, therefore, we were to accept Gen-Proto-Bi-DC4Δ as the justification of Bi-DC4, we would be obliged also to accept either that (a) the concept of being non-self-identical does not canonically present the empty set of numbers, or that (b) the formula '$x \neq x$' does not express this concept in θ, or that (c) the concept of n's being a number such that there is no odd perfect number greater than n does not canonically present the set which is, as a matter of fact, its extension, or that (d) the formula '$\neg\exists x(x > n$ & Odd(x) & Perfect$(x))$' does not express in θ the concept of n's being a number such that there is no odd perfect number greater than n. None of these alternatives seems more attractive, however, than rejection of Gen-Proto-Bi-DC4Δ.

The prospects do not improve when we pass from the justification problem for Bi-DC4 to the justification problem for the type of variant of Bi-DC4 that is needed if Bi-G2 is to be used to evaluate Hilbert's Program.

The variant we have in mind is one which interprets Bi-DC4 finitarily. Specifically, it is:

(Fin-Bi-DC4): For all m, n, $\vdash_S Prf_S(m, n) \rightarrow Prf_T(f(m), n)$.

Here m, n are supposed to be numerals (or other suitably canonical terms) in $\mathcal{L}_S$ for m, n, respectively, and f is taken to be a term of $\mathcal{L}_S$ for which there are means of 'primitive recursive computation' in S (*i.e.*, means which 'express', in the language and proof-apparatus of S, the primitive recursive computation of the function f, which is supposed to be the function expressed by f). This is the type of condition that Hilbert's finitism would require in the place of the otherwise finitarily unintelligible Bi-DC4 with its unbounded existential quantifiers in $Prov_S(x)$ and $Prov_T(x)$.

The justification of Fin-Bi-DC4 requires a principle different from Gen-Proto-Bi-DC4Δ. In particular, it calls for a principle in which reference to the sets S and T is replaced by reference to the fields of relations and the possible containment relationships that may exist between them. What adoption of Fin-Bi-DC4 as a justifiable constraint on our choice of proof expressions for S and T seems to commit us to is the idea that formulae capable of expressing relations must capture the relations of inclusion that exist among their various fields. One idea for a justifying principle for Fin-Bi-DC4 is therefore the following:

(Proto-Fin-Bi-DC4-I): When $\mathcal{C}(R^1(\mathrm{x}_1, \ldots, \mathrm{x}_n))$ and $\mathcal{C}(R^2(\mathrm{x}_1, \ldots, \mathrm{x}_n))$ are concepts that canonically present the relations $R^1(\mathrm{x}_1, \ldots, \mathrm{x}_n)$ and $R^2(\mathrm{x}_1, \ldots, \mathrm{x}_n)$, respectively, and the j^{th} field $(1 \leq \mathrm{j} \leq \mathrm{n})$ of R^1 is a subset of the j^{th} field of R^2, then there is a primitive recursive function f such that for any $\mathrm{m}_1, \ldots, \mathrm{m}_{j-1}, \mathrm{m}_j, \mathrm{m}_{j+1}, \ldots, \mathrm{m}_n$, if $R^1(\mathrm{m}_1, \ldots, \mathrm{m}_{j-1}, \mathrm{m}_j, \mathrm{m}_{j+1}, \ldots, \mathrm{m}_n)$, then $R^2(\mathrm{f}(\mathrm{m}_1), \ldots, \mathrm{f}(\mathrm{m}_{j-1}), \mathrm{m}_j, \mathrm{f}(\mathrm{m}_{j+1}), \ldots, \mathrm{f}(\mathrm{m}_n))$ and, for any formulae $C_{R^1}(x_1, \ldots, x_n)$ and

$C_{R^2}(x_1, \ldots, x_n)$ of the language of θ capable of representing $\mathcal{C}(R^1(x_1, \ldots, x_n))$ and $\mathcal{C}(R^2(x_1, \ldots, x_n))$ in θ, there are terms f and $m_1, \ldots, m_{j-1}, m_j, m_{j+1}, \ldots, m_n$ of the language of θ such that, for any $m_1, \ldots, m_{j-1}, m_j, m_{j+1}, \ldots, m_n$,

$$\vdash_\theta C_{R^1}(m_1, \ldots, m_{j-1}, m_j, m_{j+1}, \ldots, m_n) \rightarrow C_{R^2}(f(m_1), \ldots, f(m_{j-1}), m_j, f(m_{j+1}), \ldots, f(m_n)).$$[22]

Proto-Fin-Bi-DC4-I generalizes Gen-Proto-Bi-DC4Δ by extending it from single-field relations to many-field relations. It requires that formulae taken to express relations instance-wise capture subset relations that exist between parallel fields of them. Specifically, it requires a primitive recursive way of constructing, for R^2, a parallel relational 'environment' for any field of R^1 that is contained in the parallel field of R^2. This amounts to a kind of finitary 'constructibility' requirement—that any containment of a field of R^1 in the parallel field of R^2 should be inducible through a primitive recursive transformation of a relational environment that preserves the contained field of R^1 as the parallel field of R^2.

It is then further required that this 'constructibility' be part of what is registered by any formulae of the representing theory capable of properly representing R^1 and R^2.

Seen this way, the justification of Fin-Bi-DC4 (*i.e.*, Proto-Fin-Bi-DC4-I) is a composite of the following two principles:

(Finitary Constructibility): If $\mathcal{C}(R^1(x_1, \ldots, x_n))$ and $\mathcal{C}(R^2(x_1, \ldots, x_n))$ are concepts that canonically present the relations $R^1(x_1, \ldots, x_n)$ and $R^2(x_1, \ldots, x_n)$, respectively, and the j^{th} field ($1 \leq j \leq n$) of R^1 is a subset of the j^{th} field of R^2, then there should be a primitive recursive function f such that, for any r, if $R^1(x_1, \ldots, x_{j-1}, r, x_{j+1}, \ldots, x_n)$, then $R^2(f(x_1), \ldots, f(x_{j-1}), r, f(x_{j+1}), \ldots, f(x_n))$,

and

(Constructibility Registration): If $\mathcal{C}(R^1(x_1, \ldots, x_n))$ and $\mathcal{C}(R^2(x_1, \ldots, x_n))$ are concepts that canonically present the relations $R^1(x_1, \ldots, x_n)$ and $R^2(x_1, \ldots, x_n)$, respectively, and there is a primitive recursive function f such that for any r, if $R^1(x_1, \ldots, x_{j-1}, r, x_{j+1} \ldots x_n)$, then $R^2(f(x_1), \ldots, f(x_{j-1}), r, f(x_{j+1}), \ldots, f(x_n))$, then for any representing theory θ and any formulae $C_{R^1}(x_1, \ldots, x_n)$ and $C_{R^2}(x_1, \ldots, x_n)$ of the language of θ, if $C_{R^1}(x_1, \ldots, x_n))$ and $C_{R^2}(x_1, \ldots, x_n))$ properly represent $\mathcal{C}(R^1(x_1, \ldots, x_n))$ and $\mathcal{C}(R^2(x_1, \ldots, x_n))$ in θ, there

[22] Actually, Proto-Fin-Bi-DC4-I may not be as general as it ought to be. It only requires capturing of subset relations between what we here call 'parallel' fields—that is, the j^{th} field of R^1 and the j^{th} field of R^2. There is no evident reason, however, why capturing containment relations between parallel fields of two relations ought to be more vital to their proper representation than capturing containment relations between their non-parallel fields.

should be a primitive recursive term f of the language of θ such that for any $m_1, \ldots, m_{j-1}, m_j, m_{j+1}, \ldots, m_n$,

$$\vdash_\theta C_{R^1}(m_1, \ldots, m_{j-1}, m_j, m_{j+1}, \ldots, m_n) \rightarrow C_{R^2}(f(m_1), \ldots, f(m_{j-1}), m_j, f(m_{j+1}), \ldots, f(m_n)).$$

It is possible to show, however, that there are binary primitive recursive relations $\rho(x, y)$ and $\sigma(x, y)$ such that FC does not hold. A simple example (one of many) which illustrates this is the following: let $R^1(x, y)$ be defined as $y=2x$ or $y=2x+1$ and let $R^2(x, y)$ be defined as $x=y$. The y such that $\exists x R^1(x, y)$ form a subset of the y such that $\exists x R^2(x, y)$, since for k in the left field of R^1, (k, 2k) and (k, 2k+1) are both in R^1, and both (2k, 2k) and (2k+1, 2k+1) are in R^2. Since, however, no primitive recursive function (indeed, no function of any kind) maps k to both 2k and 2k+1, it follows that FC is not satisfied for this choice of R^1 and R^2.

It would seem, then, that FC is not generally an acceptable constraint to place on concepts that canonically represent relations. Either that, or there are recursive (even primitive recursive) relationships that are simply not adequately representable in any reasonable choice of representing theory. In either case, we believe, we are left with no adequate justification of Fin-Bi-DC4. In the former case, we are forced to relinquish our current candidate for a justificatory principle (*viz.*, Proto-Fin-Bi-DC4-I); in the latter, we are forced to renounce something perhaps even more fundamental to the justification of Fin-Bi-DC4—namely, the idea that all primitive recursive relations are adequately representable for reasonable choices of S.

It might be replied that the counter-example offered above to FC (hence to Proto-Fin-Bi-DC4-I) overlooks a feature of the particular relations—*proof* relations—that are of direct concern to Fin-Bi-DC4; namely, that for each value of their left fields there is a *unique* value for their right fields. It might therefore be thought that a variant of FC reformulated so as to reflect this fact would be an acceptable justificatory principle for Fin-Bi-DC4. Such a variant would run as follows.

(Proto-Fin-Bi-DC4-II): For primitive recursive relations $R^1(x, y)$ and $R^2(x, y)$ such that (i) for every x there is a unique y such that $R^1(x, y)$ and a unique y such that $R^2(x, y)$, and (ii) the y such that $\exists x R^1(x, y)$ form a subset of the y such that $\exists x R^2(x, y)$, there is a primitive recursive function f such that for all m, n, if $R^1(m, n)$, then $R^2(f(m), n)$.

I have two responses to this. The first is to argue that Proto-Fin-Bi-DC4-II does not hold. The second is to argue that even if it did, the envisioned justification of Fin-Bi-DC4 would still face grave difficulties.

The first argument begins with an appeal to Kleene's Normal Form Theorem to obtain the existence of a non-primitive recursive function with a

primitive recursive graph.[23] We let $R^1(x, y)$ be the graph of such a function $r_1(x)=y$. Since, defined in this way, $R^1(x, y)$ is the graph of a function, the condition that for every x there is a unique y such that $R^1(x, y)$ is guaranteed to be satisfied. We then let $R^2(x, y)$ be the graph of the identity function $r_2(x) = x$. By this definition of $R^2(x, y)$, satisfaction of the condition that for every x there is a unique y such that $R^2(x, y)$ is likewise guaranteed. Given this way of defining $R^1(x, y)$ and $R^2(x, y)$, we are also clearly guaranteed that the y such that $\exists x R^1(x, y)$ form a subset of the y such that $\exists x R^2(x, y)$. There can, however, be no primitive recursive function f such that

(1) for all m, n, if $R^1(m, n)$, then $R^2(f(m), n)$.

For suppose there were. Then, by (1) and the definition of $R^2(x, y)$ as the identity relation, it would follow that

(2) for all m, n, if $R^1(m, n)$, then $f(m)=n$.

But by the logic of identity we would then have it that

(3) for all m, n, if $f(m)=n$, then $R^1(m, n)$ iff $R^1(m, f(m))$.

By (2) and (3) it would thus follow that

(4) for all m, n, if $R^1(m, n)$, then $R^1(m, f(m))$,

and from (4) and the definition of $R^1(x, y)$ (as the graph of $r_1(x)$) it would follow further that

(5) for all m, n, if $r_1(m)=n$, then $r_1(m)=f(m)$.

(5), however, implies that for all m, $r_1(m)=f(m)$, and this contradicts the original assumption that $r_1(x)$ is a recursive but not primitive recursive function. Hence, Proto-Fin-Bi-DC4-II does not generally hold.

The hoped-for justificatory principle for Fin-Bi-DC4—namely, prototypical condition Proto-Fin-Bi-DC4-II—is therefore unacceptable. Hence, the currently envisioned justification of Fin-Bi-DC4 fails. This leaves the defender of Fin-Bi-DC4 with a dilemma: either (a) maintain Fin-Bi-DC4, but relinquish the more general prototypical principle Proto-Bi-DC4-II upon which to base it, or (b) relinquish Fin-Bi-DC4 as an appropriate constraint to place on the representation of the proof relations for S and T.

To grant (b), of course, is to admit our claim: namely, that the finitary correlates of Bi-PIΔ1 and Bi-PIΔ2 can not both be maintained and that a version of Bi-G2 modified so as to apply to the evaluation of Hilbert's Program can not therefore rightly be interpreted as 'saying' that no formula of S that expresses the finitary consistency of T is provable in S (in cases where S is a theory into which finitary reasoning may be embedded and $S \subseteq T$).

[23] I am grateful to Stan Wainer for suggesting this type of application of Kleene's Normal Form Theorem.

If, on the other hand, (a) is adopted, the defender of Fin-Bi-DC4 must either (1) maintain that though Fin-Bi-DC4 is a legitimate condition to place on the representation of the particular primitive recursive relations that are the proof-relations for S and T, the parallel conditions for other primitive recursive relations can not generally be maintained to be valid conditions on their representation, or (2) hold that not all primitive recursive relations are properly representable in S. Both (1) and (2) seem unattractive.

Even without these problems, however, advocacy of Proto-Fin-Bi-DC4-II as a justificatory principle for Fin-Bi-DC4 is problematic. The reason is that the shift from Fin-Bi-DC4-I to Fin-Bi-DC4-II seems arbitrary. Why should it be more important for the proper representation of $R^1(x, y)$ and $R^2(x, y)$ that their representing formulae capture subset relations between their right fields when the values of their right fields are unique (with respect to a given value of their left fields) than when they are not unique? Why should uniqueness have (or be granted) this type of normative force in the representation of relations? For that matter, why should it only be subset relations between *parallel* fields that need be captured in order to adequately represent relations? Why not require that subset relationships between non-parallel fields of relations be captured as well? But would anyone seriously maintain that subset relations between *any* two fields, parallel or not, of *any* two relations (of any arity whatsoever) must be instance-wise registered by any formulae capable of representing those relations?

Such questions seem to challenge our understanding of the appropriate standards for representation. I therefore leave the reader with the following problem(s).

(Problem 3): (i) Specify which relationships of containment among the various fields of a given family of relations must be registered by formulae capable of adequately representing those relations. (ii) Explain why it is that adequate representation of these relations requires that it be exactly these containment relationships and no others that need to be registered.

It seems, then, that even for the representation of sets—as distinguished from the representation of notions or concepts—there is no general reason to accept Bi-PIΔ2* and Bi-PIΔ1 (resp. their finitary counterparts) if one also accepts Bi-DC4 (resp. its finitary counterpart).

I would like to close this section with two disclaimers. The first is that I do not deny that there are many pairs of recursively enumerable sets and/or recursive (or primitive recursive) relations for which Bi-DC4 and/or Fin-Bi-DC4 *hold*. My argument requires only that the *holding* of Bi-DC4 or Fin-Bi-DC4 be distinguished from their being *justified*—that is, from their being properly regarded as necessary conditions on the *adequate representation* of S and T and/or their proof-relations. The difference is important since, as

noted before, formulae can and do satisfy conditions that are not necessary to their ability to represent the sets and/or relations they may represent.

Secondly, I do not deny that Bi-DC4 and/or Fin-Bi-DC4 may be proper conditions to place on the representation of T and S in certain specific cases. As mentioned earlier in connection with Problem 2, there are cases where $S \subseteq T$ holds because of some more thorough-going relationship of similarity between S and T (*e.g.*, when all proofs in S are proofs in T). In certain such cases, a Bi-DC4-like condition may be justifiable. Similarly with certain cases where Bi-G2 obtains not because $S \subseteq T$, but because of a relationship between S and T that is more general than $S \subseteq T$ (*e.g.*, cases where, though S is not a subtheory of T, there is nonetheless a 'translation'—one which preserves such things as proof and/or theoremhood—from proofs and/or theorems of S to proofs and/or theorems of T).[24]

None of this, however, changes the fact that not all relationships between S and T that permit a *proof* of Bi-G2 also permit a *justification of* the conditions (*viz.* the particular Bi-DC4-like condition) that such proofs place on the representation of the proof and/or provability relations of S and T. We have argued this in the particular case of the relationship $S \subseteq T$. But the same type of problem arises in the case of other relationships as well. They all show a need for a Bi-DC4-like condition at the level of 'arithmetized' metamathematics in order to obtain a *proof* of a Bi-G2-like theorem; and they all run into problems similar to the ones mentioned above when it comes to *justifying* this condition.

The following problem therefore arises:

(Problem 4): (i) Characterize the full range of relationships between S and T that both support a *proof* of a bitheoretic version of G2 and for which there is a *justification* of the Bi-DC4-like condition used in that proof. (ii) For each different relationship between S and T that supports a proof of a bitheoretic version of G2, give an exact statement of the Bi-DC4-like condition that is needed and provide a justification for it.

Only when Problem 4 is solved will we be in a position to determine what are the justified interpretations of G2 (including its bitheoretic variants) and to discern what it is that G2 truly 'says'. I have tried to indicate in this section, and throughout this paper generally, how far we are from having a solution to this problem.

6. Discussion

It has been objected that the entire line of argumentation developed in the preceding sections rests on a failure properly to reckon the logical form

[24] Cases of this latter type require certain minor modifications of Bi-DC4 and/or Fin-Bi-DC4.

of Bi-G2.[25] Specifically, it has been claimed that all that proof of Bi-G2 requires is that $Prov_T(x)$ satisfy Bi-DC1 and Bi-DC2 and that *there exist some formula* $\mathcal{F}(x)$ of $\mathcal{L}_S$ which, together with $Prov_T(x)$, satisfies the following conditions.

Bi-DC3*: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$,
$$\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow \mathcal{F}(\ulcorner Prov_T(\ulcorner \mathcal{A} \urcorner) \urcorner);$$

Bi-DC4*: For every sentence $\mathcal{A}$ of $\mathcal{L}_S$, $\vdash_S \mathcal{F}(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner \mathcal{A} \urcorner)$.

$\mathcal{F}(x)$ need *not* express provability-in-S in order to satisfy Bi-DC3* and Bi-DC4*. It might, for example, be replaced by a formula $Prov_T(x)$ that expresses the notion of provability-in-T. There is therefore no need, the reasoning continues, to introduce a formula $Prov_S(x)$ that expresses the notion of provability-in-S. From this it follows that there is also no need to introduce the conditions Bi-DC3 and Bi-DC4 (or Bi-DC3* and Bi-DC4*). One might just as well revert to the more nearly 'monotheoretic'

Bi-DC3†: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$,
$$\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner Prov_T(\ulcorner \mathcal{A} \urcorner) \urcorner)$$

as the 'instantiation' of Bi-DC3* and to the tautologous

Bi-DC4†: For every sentence $\mathcal{A}$ of $\mathcal{L}_T$, $\vdash_S Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner \mathcal{A} \urcorner)$

as the 'instantiation' of Bi-DC4*.[26]

So long as $\mathcal{F}(x)$ satisfies Bi-DC3* and Bi-DC4*, the reasoning goes, it does what it has to do; namely, provide a means of getting from (3) to (5) in the proof of Bi-G2 Lemma. Once one realizes this, the argument concludes, one sees that the entire argument of this paper is based upon a false assumption—namely, that to prove a version of Bi-G2 in which the conditions on the key metamathematical notions are all *justified*, one must introduce a formula $Prov_S(x)$ whose task is to express the notion of provability-in-S.

In reply to this objection, I have two related points to make. Before giving them, however, I want to make it clear that I agree with the *logical* claim that the objection makes—namely, that all that is needed for the *proof* of (a variant of) Bi-G2 is the existence of an $\mathcal{F}(x)$ that satisfies Bi-DC3* and Bi-DC4*. What I do not agree with is that this fact casts doubt on the analysis and argument of this paper.

The ground of my disagreement is simple and, in the end, it boils down to a point I have already made repeatedly: namely, that to obtain an interpreted version of G2 from a literal version of G2, it is not enough

[25] Warren Goldfarb and Matthew Frank raised this objection during presentation of a condensed oral version of this paper at the Boolos Symposium in April of 1998.

[26] Bi-DC1, Bi-DC2 and Bi-DC3† suffice for the proof of Bi-G2 Lemma. Bi-DC3† is not an entirely monotheoretic version of DC3 since it requires that '$Prov_T(\ulcorner \mathcal{A} \urcorner) \rightarrow Prov_T(\ulcorner Prov_T(\ulcorner \mathcal{A} \urcorner) \urcorner)$' be provable in S rather than T. Bi-DC3† is the same condition we referred to earlier as Bi-DC3Δ.

simply to *prove* G2; one must also *justify* the conditions on the formulae representing the key metamathematical notions that are used in the proof. In this paper, I have concerned myself with one particular justification of one particular such condition—namely, the Reflexivity Defense of the Third Condition. My claim has been that this justification requires introduction of a formula that expresses the notion of provability-in-S and that it does so because of the relationship that it asserts to exist between the First and Third Conditions.

This relationship does not obtain between Bi-DC3* and Bi-DC1. It is therefore a characteristic intrinsic to the Reflexivity Defense that demands replacement of Bi-DC3* and Bi-DC4* by Bi-DC3 and Bi-DC4.

It may be, however, that a broader, more radical critique of our argument is intended in the suggestion that Bi-DC3* and Bi-DC4* be substituted for Bi-DC3 and Bi-DC4. This is a critique which holds that the entire line of thinking behind the Reflexivity Defense, and all other defenses of versions of the Third and Fourth Conditions that are less general than Bi-DC3* and Bi-DC4*, are fundamentally misguided and that they all misrepresent the logical form of G2. Only the version framed in terms of Bi-DC3* and Bi-DC4*, the critique continues, attributes to G2 a properly general form.

In response to this, I would ask a question: 'What *is* the *justification* of the general condition whose two clauses are Bi-DC3* and Bi-DC4*?'. It is, after all, interpreted or proto-philosophical versions of G2, and not merely literal versions, that we are interested in here. And while it may be agreed that one can obtain (*i.e.*, *prove*) a literal version of G2 using only Bi-DC3* and Bi-DC4*, it does not thereby follow that one can obtain an interpreted version of G2 from it. To put it plainly, I do not believe that there is a justification of Bi-DC3* and Bi-DC4*. The replacement of $Prov_S(x)$ by $\mathcal{F}(x)$ leaves one with a condition that is simply too 'abstract'—or, better, too indefinite in content—to admit of justification. Or perhaps I ought rather to say that it leaves one with conditions that are too indefinite to admit of basic or non-derivative justification. A justification of Bi-DC3*–Bi-DC4* might be derived from some more specific set of conditions such as Bi-DC3 and Bi-DC4 or Bi-DC3† and Bi-DC4†. In such cases, however, it is the justification of these more definite conditions and not the justification of Bi-DC3* and Bi-DC4* that is ultimately at issue, and the analysis and argument presented above then apply.[27]

[27] Justification of Bi-DC3* and Bi-DC4* by appeal to Bi-DC3 and Bi-DC4 would, of course, require introduction of a formula expressing provability-in-S, and the analysis and argument given above would then apply. Justification of Bi-DC3* and Bi-DC4* by appeal to Bi-DC3† and Bi-DC4†, on the other hand, would not seem to fit with the thinking behind the Reflexivity Defense since Bi-DC3† lacks the requisite relationship to Bi-DC1. The Reflexivity Defense requires that the formula making up the consequent of the Third Condition express the property or concept in terms of which the consequent of Bi-DC1 is stated. Justification of Bi-DC3* and Bi-DC4* via justification of Bi-DC3† and Bi-DC4† is therefore justification outside the bounds of the Reflexivity Defense, and so justification outside the scope of interest of this paper.

References

DETLEFSEN, MICHAEL [1990]: 'On an alleged refutation of Hilbert's Program using Gödel's first incompleteness theorem', *Journal of Philosophical Logic* **19**, 343–377.

FEFERMAN, SOLOMON [1960]: 'Arithmetization of metamathematics in a general setting', *Fundamenta Mathematicae* **49**, 35–92.

_____ [1982]: 'Inductively presented systems and the formalization of meta-mathematics', in D. van Dalen *et al.*, eds. *Logic Colloquium '80*. Amsterdam: North-Holland, pp. 95–128.

_____ [1989]: 'Finitary inductively presented logics', in R. Ferro *et al.*, eds. *Logic Colloquium '88*. Amsterdam: North-Holland, pp. 191–220.

HILBERT, DAVID, and PAUL BERNAYS [1939]: *Grundlagen der Mathematik*, II. Berlin: Verlag Julius Springer.

ODIFREDDI, P.-G. [1989]: *Classical Recursion Theory*. Amsterdam: North-Holland.

PRAWITZ, DAG [1981]: 'Philosophical aspects of proof theory', in G. Fløistad, ed., *Contemporary Philosophy: A New Survey*. Vol. 1. The Hague: Martinus Nijhof.

SCHÜTTE, K. [1960]: *Beweistheorie*. Berlin-Heidelberg-New York: Springer-Verlag.

_____ [1977]: *Proof Theory*. Berlin-Heidelberg-New York: Springer-Verlag.

SMORYNSKI, C. [1977]: 'The incompleteness theorems', in J. Barwise, ed. *Handbook of Mathematical Logic*. Amsterdam: North-Holland,.

_____ [1988]: 'Hilbert's programme', *CWI Quarterly* **1**, 3–59.

ABSTRACT. We consider a seemingly popular justification (we call it the Reflexivity Defense) for the third derivability condition of the Hilbert-Bernays-Löb generalization of Gödel's Second Incompleteness Theorem (G2). We argue that (i) in certain settings (roughly, those where the representing theory of an arithmetization is allowed to be a proper subtheory of the represented theory), use of the Reflexivity Defense to justify the third condition induces a fourth condition, and that (ii) the justification of this fourth condition faces serious obstacles. We conclude that, in the types of settings mentioned, the Reflexivity Defense does not justify the usual 'reading' of G2—namely, that the consistency of the represented theory is not provable in the representing theory.