4

HOW TO SAY THINGS WITH FORMALISMS

David Auerbach

SUMMARY

Recent attention to "self-consistent" (Rosser-style) systems raises anew the question of the proper interpretation of the Gödel Second Incompleteness Theorem and its effect on Hilbert's Program. The traditional rendering and consequence is defended with new arguments justifying the intensional correctness of the derivability conditions.

I

The conception of formalism as uninterpreted, but interpretable, systems that was wavering into focus during the 1920s achieved clarity in Godel's 1931 paper. The historical irony here is that while this clarity fulfilled one of Hilbert's demands for firm foundations, the Godel results themselves have been seen as frustrating Hilbert's central epistemological desire.

Due in part to Hilbert's vagueness in formulating his demands, and in part to the subtlety of the issues involved, there remains an ongoing debate as to whether the First Incompleteness Theorem scuttled Hilbert's Program, whether only the Second did, or whether neither did. Leaving detailed exegesis of Hilbert aside, I propose to investigate some very general issues that arise in the course of such debate.

Hilbert wanted to save mathematics from Kronecker's constructivist critique, a critique made pointed by the antinomies of set theory. Hilbert, however, shared some of Kronecker's basic epistemological scruples and thus proposed, on one interpretation of his project, to salvage the nonconstructivist part of mathematics by making it meaningless but helpful.

This instrumentalist reading of Hilbert is arguable. Or rather, "meaningless" might be taken only as Fregean meaningless. It is

not that the nonfinitary mathematics is true because, *inter alia*, its terms refer, but that its terms refer because it is true. The Hilbert Program then aims at making sense of this notion of truth prior to reference. This would make Hilbert what we would now call a meaning-holist. For present purposes I need not choose between these Hilberts; the common core is the claim that the theorems of mathematics get all of their acceptability from system-wide considerations and not from the one-by-one truth of the theorems or axioms. His enduring insight was that the representation of theories about infinitistic matters as finitistic objects (formalisms) gave one a finitistic handle on the nonfinitistic. Here is the standard story in brief.

Some parts of mathematics are about finite objects and finitistically establishable properties of them. The "upper bound" on this conception of finitary mathematics is the notion of the "potentially" infinite, as exemplified by the sequence of natural numbers. (It is, of course, a matter of some dispute as to how to mark precisely, in extension, this distinction.) This part of mathematics the Hilbertian treats contentually: questions of truth and belief arise and, roughly speaking, it has a Fregean semantics continuous with that of nonmathematical concrete language.

The rest of mathematics consists of ideal pseudo-statements whose only role is the efficient and secure calculation of contentual statements. (The noninstrumentalist Hilbert would put this: the rest of mathematics consists of ideal statements whose truth is explicated by their role in the efficient and secure calculation of contentual statements.) Hilbert wanted to assuage Kronecker's epistemological doubts that rendered ideal sentences problematic. But Hilbert thought that they possessed a significance insofar as they were useful in deriving real statements. This usefulness arises because a finitistic proof of a real sentence might be unfeasibly long or difficult, whereas a proof that took a shortcut through the infinite might be shorter or at least easier to find.[1] Along with this explanation of the usefulness of the ideal, Hilbert needed a certification of its safety: that is, that there is no ideal proof of a false real proposition. This is real-soundness.

Hilbert saw the then nascent conception of formalism as vital at this point. If ideal proofs could be finitistically represented as manipulations of concrete symbols and strings of symbols, then a real proof about ideal proofs becomes a feasible notion. Hilbert also thought that the combinatorial properties of concrete symbols

were the real content of arithmetic; in this way the entire actual content of mathematics becomes metamathematical. All real mathematics is about symbols (and finitary systems of symbols). This answers Kronecker's complaint that the content of mathematics should be exhausted by the elementary theory of numbers.

The Hilbertian would like a finitistic proof of real-soundness; however, real-soundness of T implies consistency of T and, if T includes finitistic reasoning, T cannot prove T's consistency. So finitistic proofs of real-soundness are a forlorn hope and the epistemological goal of Hilbert's Program is unattainable.[2]

There are two related matters in this brief story that I wish to disinter. The first is the cavalier manner in which talk of mathematics is replaced by talk of formal theories – in which theorems about formalisms are used to motivate conclusions about mathematics proper. The second is the very much related issue of the sense in which consistency is expressed in a formalism. These are not only hard to get straight, but positions on these matters are often difficult to extricate from the viewpoint being argued. I will be arguing that once these buried matters are exhumed and examined, we will be able to safely rebury Hilbert's Program.

II

In an earlier paper,[3] I argued that there is a considerable distance between the technical result about formalisms that Gödel sketches in Theorem XI and since known (in various generalized forms) as the Gödel Second Incompleteness Theorem and its standard gloss. The standard gloss states that no sufficiently strong formal system can prove its own consistency or (as above) T cannot prove T's consistency, for Ts that extend PA. Note that the *first* theorem makes no claim about the content of its underivable sentence. What would smooth the path from the mathematical theorem to the philosophical gloss are the Positive Expressibility Thesis (PET) and the Negative Expressibility Thesis (NET):[4]

PET The underivable (in **T**) sentence of the Gödel Second Incompleteness Theorem does express consistency.

NET No derivable (in **T**) sentence expresses consistency.

Detlefsen (1986), in an intriguing resurrection of Hilbert's Program, argues that NET has not been established. He exploits the fact that special problems arise in looking for a satisfying

account that would establish PET and NET. These problems, and technical machinery for ameliorating them, have recently become more widely known. I sketch them here, focusing on Peano arithmetic and its formalization PA.

In proving the First Incompleteness Theorem a formal predicate is constructed that numeralwise expresses *is a derivation of*, and from this predicate is constructed the famous Gödel sentence. Letting **Prf(x, y)** be an arbitrary predicate that numeralwise expresses *is a derivation of*, the Gödel sentence is $\neg\exists$**xPrf(x, k)**, where k is the Gödel number of a formula provably (in PA) equivalent to $\neg\exists$**xPrf(x, k)**. Many formulas numeralwise express *is a derivation of*, and any of them will suffice for the First Theorem. If we now construct $\neg\exists$**xPrf(x,** $\ulcorner\bot\urcorner$**)** as a plausible candidate for a consistency sentence, we get the problem that some such sentences are trivial theorems. The classic example uses Rosser's derivability predicate[5] but other constructions have been studied.

So either NET is false or we can find some reason to exclude Rosser-type predicates. Now we can find some *ways* to exclude them; but in the insightful and important Chapter IV of Detlefsen (1986) it is argued that these ways are not reasons.

The way that Detlefsen considers descends from the Hilbert–Bernays derivability conditions. In particular, the formalization of what Detlefsen calls local provability completeness, LPC, plays the crucial role. LPC is simply

LPC If $\vdash_T$**A**, then $\vdash_T$**Bew**$_T$**([A])**,

and its formalization is[6]

F-LPC $\vdash_T$**Bew**$_T$**([A])** $\rightarrow$ **Bew**$_T$**([Bew**$_T$**([A])])**.

F-LPC is, of course, familiar in modal guise, as it appears in modal treatments of provability: $\vdash \Box A \rightarrow \Box\Box A$. Detlefsen argues that no argument for the necessity of F-LPC, as a condition on correct derivability predicates, works.

It is instructive to look at an untenable attempt at establishing NET, due to Mostowski, to see what can go wrong. Mostowski's goal is to justify F-LPC by seeing it as part of the project of representing as many truths about provability as possible.

Bew(x) is naturally viewed as the existential generalization of **Prf(x, y)** where **Prf** represents the derivability relation. Here "represents" has the purely technical meaning of "strongly represents" or "numeralwise expresses" – whenever **a** is a derivation

of **b**, $\mathbf{Prf}(\ulcorner \mathbf{a} \urcorner, \ulcorner \mathbf{b} \urcorner)$ is a theorem and whenever **a** is not a proof of **b**, $\neg\mathbf{Prf}(\ulcorner \mathbf{a} \urcorner, \ulcorner \mathbf{b} \urcorner)$ is a theorem. This enforces the extensional adequacy of the representation of *is a derivation of*, but does not yield F-LPC.

But not all truths are particular, that is, numeralwise, truths. So perhaps we should demand that all truths concerning provability be represented. The first stumbling block is that Gödel theorems show the impossibility of this for precisely the case in point. So we weaken the demand to one that as many truths as possible be represented.

But for the latter proposal to make sense we have to construe it as demanding that certain truths be represented. But which? An answer that lists F-LPC is circular with respect to the project at hand, which was justifying F-LPC. (Any variation on a purely quantitative approach that subsumes F-LPC will run into the following problem: we will not be able to codify any truth about *unprovability*. Any assertion of unprovability is tantamount to consistency.)

Detlefsen is certainly right about *this* strategy (as well as those of Prawitz and Kreisel–Takeuti) of justifying the derivability conditions. Those strategies, as Detlefsen demonstrates,[7] attempt to prove too much. We can do better by attempting less and relocating disputes about F-LPC to their proper philosophical arena. I will return to this claim in Section V.

III

Under what conditions do we legitimately ascribe meaning to the formulas of a formalization; and what do these conditions permit by way of inference concerning the subject matter of the ascribed meaning? To make this both concrete and simple: what justifies the (correct) claim that in formalized Peano arithmetic we can prove that 2 is less than 3 or that addition is commutative? In much of what follows there is a tedious dwelling on standard technical moves that in technical contexts are treated off-handedly or invisibly; here, for philosophical purposes, we linger over them.

What makes the formalism we call Peano arithmetic arithmetic is more than the formalism; we do not need something as sophisticated as the Skolem–Löwenheim theorem to tell us that uninterpreted formalisms are uninterpreted.

Formal languages permit many interpretations, and except for

perhaps in some contingent historical sense, no interpretation is privileged other than by stipulation. Furthermore, for purposes that exceed mere consideration of truth conditions in an extensional language, the nonidentity of *ways of giving* the interpretation is relevant. Note that this extends to the interpretation of quantification, via the specification of the domain.[8]

We will want an account of "ways of giving" to extend naturally to allow the usual overview of the First Theorem as proceeding by constructing a formula that says that it is not derivable; and of the Second Theorem as demonstrating the nonderivability of consistency. (The First Theorem actually has very little to do with interpretation. It is best seen as a remark about formalisms, independent of their possible interpretations, and the machinery necessary to prove it is indifferent to any rich sense of interpretation. Thus *Rosser's* proof of the First Theorem does not proceed "by constructing a formula that says it is not derivable." But this is precisely the desirable expected result yielded by the intensional account being developed.) To avoid using the awkward "ways of giving" and the already otherwise employed "interpretation," I will use "reading" to talk about the richer sense of interpreting formulas sketched here.

The recasting of the language of formalized arithmetic so that we can plausibly say that a certain sentence of arithmetic(!) says that $\exists$ is a quantifier, that **G** says that **G** is not provable, that $\mathbf{Con}_{\mathbf{T}}$ says that **T** is consistent, is not accomplished directly, by paralleling the arithmetic case.

As with arithmetic we start with the informal language, a piece of technical but natural language, with which we theorize about the syntax of formalisms. This language has terms for such entities as $\forall$, $\exists$, and predicates like *is a formula, is a derivation from*, etc. The objects that this language treats of include the symbols of **PA** and pairs and sequences of them. Following Boolos, we call the language the language of Syntax and the informal theory we couch in it is Syntax.

We do not directly specify a reading of the language of **PA** in terms of the language of Syntax. The syntactic reading of **PA** is, unimportantly, parasitic on the arithmetic interpretation.[9] We first set up a correspondence between the *objects* of **PA** and the *objects* of Syntax. This is Gödel numbering. This is used to induce a correspondence between the terms and formulas of **PA** and the names and predicates of Syntax. The initial correspondence, at the

level of the primitives of Syntax, is extension respecting – a term of **PA** corresponds to a name in Syntax by denoting the Gödel number of the object denoted by that name. Definitional compounds in Syntax are then merely mimicked in **PA**. The initial coextensiveness guarantees the various useful meta-theorems about this re-reading of **PA**: sentences of **PA** are derivable in **PA** *only if* their counterparts are demonstrable in Syntax.

This correspondence makes no mention of the axiomatization of **PA**; the same old extensional interpretation of **PA** now induces a way of reading the formulas of **PA** as syntactic remarks. It is, of course, a different collection of predicates of (conservative extensions of) **PA** that will be of reading interest; typically, those whose extensions are the usual syntactic categories and which have "natural" readings into Syntax.

We abet and exploit the closeness of the correspondence by using, for certain terms and formulas of **PA**, orthographic cousins of their syntactic counterparts.[10] A rather large amount of Syntax can be mirrored in **PA**; crudely speaking, it is the Gödel Second Incompleteness Theorem that puts an upper bound on what can be mirrored.

It may strike the reader that there is a freedom here that was not present in the arithmetic interpretation of **PA**. Certainly, the strict copying into the language of **PA** of definitions and proofs in Syntax constrains the correspondence between **PA** and Syntax more than mere co-extensiveness would. But no analogous strictures were stated concerning the primitive correspondences from which the whole is built up. But there are such strictures. The term that corresponds to a name in Syntax is a *numeral* of **PA**. Without such a stricture the Gödel theorems would not be forthcoming and certain quantificational facts about provability would not even be statable.

Let **Bew(x)** be a formula of **PA** that results from a long series of definitions in **PA**, as just outlined, such that we can read **Bew(x)** as "x is derivable." We can read it like this because (i) in the standard interpretation it is true of just the (Gödel numbers of) theorems; and (ii) it was built up in **PA** by mimicking the standard definitions of formula, axiom, variable, free for, etc. in Syntax. It is notable that condition (i) can hold without condition (ii) and that some formulas for which condition (i) but not condition (ii) hold are sufficient for a proof of G1 but not of G2.[11]

IV

I have sketched an arithmetic reading of the language of **PA** and a syntactic reading of the same language. Thus the very same formula can be read in either way. While these readings concern, so far, the *language* of **PA**, the justification of their strictures develops from the use to which we put *formalisms* (i.e. language as well as axioms and a notion of derivation).

A use to which we put formalisms is the "capturing" of facts about the intended reading. One aim is to characterize a syntactic mechanism that isolates formulas whose readings are true. (This is the successful component of Hilbert's Program.) We do this by specifying the notion of a derivation. Furthermore, we often have in mind the additional requirement that being a derivation reflect the notion of proof in a straightforward way: readings of derivations should be proofs. This is the epistemological applicability requirement:

EAR If **A** is informally provable from the principles that **T** formalizes, then **A** is derivable in **T**; and derivations of **A** in **T** formalize proofs of **A**.

Our account of "boldfacing," via structure-sensitive readings, yields EAR, establishing PET and NET. The real work, in particular cases, is giving enough sense to the boldface mapping to support EAR. Since *is provable* is an intensional predicate the need for a structure-sensitive notion, sketched above as a "reading," should be no surprise.[12]

Built into our characterization of the reading of formulas of **PA** as arithmetic and as syntactic was a respect for logical form. Thus, if **Fa** is read as "14 is even," then $\exists$**xFx** is read as "something is even." When we add a notion of derivation to that language of **PA** to get **PA** proper we also obtain a useful result: there will be a derivation of $\exists$**xFx** from **Fa**.

We need beware a potential confusion. In giving the syntactic reading of the language of **PA** we did not use a notion of derivation; derivation merely happened to be one of the notions of the theory, Syntax, into which we were reading the formulas of **PA**. Derivation makes a second appearance, which we are now noting, as a part of **PA**, the neutral formalism. More precisely, we now consider the full-fledged notion of *is a derivation* (of **PA**), complete with specification of the usual axioms.

The usefulness of our structure-sensitive readings exceeds the modest consequence that theorems of **PA** tell us truths of Syntax; the surplus value is that the derivations of **PA** correspond to proofs in Syntax. So the appropriate start to a justification for the usual gloss of the Gödel Second Incompleteness Theorem amounts to the observation that any purported proof of $\mathbf{CON}_{\mathbf{PA}}$, using the machinery formalized in **PA**, would give us a derivation of $\mathbf{CON}_{\mathbf{PA}}$, which is what the purely technical result tells us cannot happen.

Here is the situation so far. Extension-respecting readings of **PA** are inadequate even for explicating the representation of arithmetic statements in **PA**. Structure-respecting readings reflect our actual practice in reading formal formulas, and the somewhat devious case of the syntactic reading of the formulas of **PA** was partially detailed. When we add the usual axiomatization in **PA** we get useful meta-theorems linking the derivability of certain formulas in **PA** with the establishment of theorems of (informal) Syntax. (Similarly, but less to the present point, we get useful meta-theorems linking the derivability of certain formulas in **PA** with the establishment of (informal) arithmetic.)

I now need to link the observations of Section II with those of Section III. One fact Syntax can prove about **Bew** is that if $\vdash_{\mathbf{PA}}\mathbf{A}$ then $\vdash_{\mathbf{PA}}\mathbf{A}\ \mathbf{Bew}([\mathbf{A}])$. So **PA**, insofar as it is adequate to Syntax, ought to be able to derive **if** $\vdash_{\mathbf{PA}}\mathbf{A}$ **then** $\vdash_{\mathbf{PA}}\mathbf{Bew}([\mathbf{A}])$. In other words,

F-LPC $\quad \vdash_{\mathbf{PA}}\mathbf{Bew}([\mathbf{A}]) \rightarrow \mathbf{Bew}([\mathbf{Bew}([\mathbf{A}])])$

where the notation used reflects the stipulations of note 11. And, of course, F-LPC holds because the way we constructed **Bew**[13] excludes Rosser-style predicates. Thus, this justification of F-LPC does not depend on the Mostowski, Kreisel–Takeuti or Prawitz arguments that Detlefsen rightly criticizes. However, in considering arguments for F-LPC, and hence for NET, Detlefsen presents positive arguments against F-LPC; it is to those that I turn in the next section.

V

The notion of reading partially[14] developed above is the only viable candidate I know that explains the uses to which we put formalisms. As I read Detlefsen, his real quarrel is not with such a

notion, nor even with constructions of **Bew** that guarantee F-LPC. I think he thinks that even if EAR is true it has no deleterious effect on the instrumentalist Hilbertian. For, quite simply, the Hilbertian is not interested in the same notion of proof that we are.

Hilbertians should grant that F-LPC is an unarguable fact about any **Bew** that really represents derivability – and further grant that this notion of derivability nicely captures the classical notion of proof. The Hilbertian's radical proposal is that our classical notion of proof needs amendment and it is that amended notion that belongs in the construction of a consistency sentence (cf. Detlefsen 1986: 121ff.). But this is prior to any formalization. The derivability conditions appear because Detlefsen finds inspiration for an amended concept of proof in the Rosser predicates.

Detlefsen is careful to say that he is not, for various reasons, proposing any particular Rosser-style concept of proof as a candidate for Hilbertian proof. He does, however, regard them as good models for the right *sort* of approach to proof on an instrumentalist basis; and he defends Rosser proofs against certain objections (see Detlefsen 1986: 122ff.).

Nonetheless, I think the peculiar nature of such a notion of proof is worth dwelling on, particularly since our rich notion of a reading suggests a useful heuristic for thinking about derivability predicates.

Confronted with a derivability predicate that does not satisfy F-LPC we have reason to be wary about statements made that utilize it. Consider Edna, whose set of beliefs concerning formalisms is precisely characterized by our syntactic reading of **PA**. Amongst her beliefs are beliefs about **PA**. When Edna says that if **A** is derivable in **PA** then it is derivable in **PA** that that **A** is derivable in **PA**, we have every reason to believe her. For, by our stipulation about the nature of Edna's beliefs, this amounts to F-LPC. If challenged, Edna can demonstrate F-LPC, having at her command the power of induction and the inductive definitions of Syntax.

Edna's situation differs from Ralph's. Ralph's set of beliefs is also partly characterized by **PA** together with a deviant syntactic reading: Ralph's beliefs about formalisms are characterized in terms of some Rosser-style derivability predicate. That is, when Ralph uses *derivable* in a belief we read it as a Rosser-style predicate. Unlike Edna, Ralph believes that $\mathrm{CON}_{\mathrm{PA}}$, although our report of this would be less misleading if we said: Ralph believes

that $\mathbf{CON}^*_{PA}$. Note that $\mathbf{CON}_{PA}$ and $\mathbf{CON}^*_{PA}$ are distinct, non-equivalent sentences of **PA**. In fact, an even less misleading report of Ralph's belief is:

Ralph believes that the largest consistent subsystem[15] of **PA** is consistent.

That **PA** is consistent and Ralph's belief that the largest consistent subsystem of **PA** is consistent are radically different beliefs, although **PA** *is* the largest consistent subsystem of **PA**. Leopold, a skeptic about PA's consistency, would hardly be reassured by Ralph's assertion of his belief, once he understood the content of Ralph's belief. Nor, of course, can Edna reassure him since she does not believe that **PA** is consistent. Of course, properly understood, neither does Ralph. Neither Edna nor Ralph have the right *de re* belief about **PA** necessary to assuage Leopold's worries.

We know that

1 Ralph can use his notion of derivability to prove anything Edna can, and
2 Ralph knows that his notion of derivability will never produce a proof of $\perp$.

But Ralph does not know 1.

The recommendation that we reform our mathematical practice and replace the canonical notion of derivability with a Rosser-style one will indeed assure us, quite easily, of consistency. But that epistemic gain is offset by the epistemic loss occasioned by not knowing what it is that is consistent.

The epistemic trade-off is most easily seen by looking at Feferman's version of a Rosser system. Taking some enumeration of the (infinitely many) axioms of **PA**, we can define the notion of an *initial* set of axioms. Then let R = {x|x is a finite, initial, consistent set of axioms} be Ralph's version of the axioms of PA. R is the set of axioms of **PA**. Ralph's notion of derivation now has the following character.[16]

In reasoning about the formalism derived from R, Ralph generates ordinary derivations, $d_1, d_2, d_3, \ldots$. If d_i's last line is $\perp$, he goes back and tosses out all derivations using axioms no smaller than the largest one in d_i. He proceeds, always tossing out such derivations, and every time a derivation of $\perp$ is encountered the toss-out procedure is repeated. An R-derivation is a derivation that is never tossed out.

Ralph can decide whether a given derivation is an R-derivation, although neither he nor Edna can know that all derivations are R-derivations. Leopold cannot rationally worry that R-derivation is insecure. But he can, and will, be skeptical about the extent of Ralph's knowledge.

The instrumentalist can reply that this is just fine; a secure system of unknown extent is better than an insecure system of known extent. Without direct quibbling about the insecurity of **PA**, I would point out that our (nontechnical) interest in the system produced by R-derivations (or similar Rosser systems) is proportional to the strength of our belief that it is co-extensive with **PA**. But the stronger this belief, tantamount to consistency of **PA**, the less reason to bother with R-derivations.

The instrumentalist's perfectly coherent option here is to point out a divergence of interests – that "our" interest is not hers. Detlefsen's Hilbertian can live happily with a possibly pared down ideal mathematics, even one of unmappable extent. The instrumentalist is under no obligation to lay claim to our "secret" knowledge of the co-extensiveness of **PA** with the Rosser system. The merits (and demerits) of this choice, however, lie beyond the scope of this paper.[17]

A final point about the double role of **PA**. Unless one takes seriously the way in which the syntactic reading of **PA** yields a notion of a formalism talking about formalisms that include it, needless puzzles arise about the need for F-LPC. A useful way of looking at this involves the related "puzzle" about proofs of the Gödel Second Incompleteness Theorem for systems weaker than **PA**.

Typically, the first theorem is shown to hold for extensions of **Q** (a finitely axiomatizable system) or perhaps for some other weak system like **PRA**. The conditions on a system, needed for a proof of the Gödel First Incompleteness Theorem, are well known and simple to state. The Gödel Second Incompleteness Theorem, however, is stated in full generality for extensions of a much stronger system than **Q**, namely **PA**. Of course the *technical* reason for this is that F-LPC will not be available in weaker systems; but it is worth seeing why, conceptually, this is not a mere *ad hoc* contrivance.

What underlies the presence of F-LPC is an ability to deal with formalisms by being able to comprehend the essentially inductive definition of a formalism. Before Ralph learned as much as he did,

his syntactic beliefs were characterized, not by **PA**, but by **Q**. Similarly with Edna. We maintain, for the moment, the same readings of their beliefs.

What should Leopold make of the beliefs about **PA** or **Q** entertained by the younger Ralph or Edna? They do not have any. What should we make of Edna's possible belief represented by $\mathbf{CON}_{\mathbf{Q}}$? First of all she will not believe it.[18] The delicate question, however, is how we should read it or Ralph's deviant version (which he can derive). In this case neither Edna nor Ralph know what they are talking about; they do not know the first thing about formalisms. Neither of them understand (have beliefs about the defining characteristics of) formalisms in general or **PA** or **Q** in particular. They cannot even entertain the propositions about **PA**'s or **Q**'s consistency. To read these weak systems as making remarks about formalisms would be to misconstrue what they are capable of telling us.

NOTES

I would like to thank Harold Levin, Louise Antony, Joe Levine, and members of the Triangle Language and Mind Group for helpful discussions and Michael Detlefsen for useful comments on an earlier version.

1 Detlefsen (1986) elaborates this in terms of human capabilities and the possible divergence between humanly natural modes of reasoning and correct modes of reasoning. This should be contrasted with the treatment of these same matters by Hallett (1989).

2 Let T be a system of ideal mathematics and let S be a finitistically acceptable theory of real mathematics. There is some question as to whether real-soundness should be taken as: $[\text{Real}(A) \,\&\, \vdash_T A] \rightarrow \vdash_S A]$ or as the weaker $[\text{Real}(A) \,\&\, \vdash_T A] \rightarrow \neg \vdash_S \neg A$. On the second formulation there can be real sentences not decided by S, and if T proves them this does not obligate S to prove them. This is one way for a Hilbertian to argue the irrelevance of the First Incompleteness Theorem – the Gödel sentence, though real, is not finitistically established. Detlefsen (1990) points out that the weaker version of soundness is the plausible one.

3 Auerbach (1985).

4 From this point forward I will use the convention of boldfacing to indicate that a formal object is being referred to as well as, in the appropriate contexts, to indicate that the formal object is the formal representation of the nonbold informal term. Consider the simplest case: **2** names the numeral for 2. We shall be concerned with constraints on boldface mappings, particularly those that yield the

useful **2 ≠ the only even prime**, although **2 = the only even prime**. Ultimately we want:

CON(PA) is not derivable; hence CON(PA) is not provable by methods formalized by PA,

as a consequence of an adequate notion of boldfacing.

The word "derivation" will apply to certain formal objects, while "proof" refers to those unformalized items discovered in the daily practice of working mathematicians.

5 Let **Prf′** be

$$\mathbf{Prf}(x, y) \;\&\; \neg\exists x\, x < y \;\&\; \mathbf{Prf}(x, \mathbf{neg}(y)),$$

which reads "x is a derivation of y and there is no smaller derivation of the negation of y." For consistent formalisms like PA, **Prf′** is co-extensive with **Prf** and numeralwise expresses what **Prf** does. A more stripped down Rosser-style predicate is **Prf″**:

$$\mathbf{Prf}(x, y) \;\&\; \neg\mathbf{Prf}(x, [\mathbf{0} = \mathbf{1}]).$$

The result of replacing **Prf** with either **Prf′** or **Prf″** in the "consistency" formula is a trivial theorem. This dooms numeralwise expressibility as a sufficient condition for capturing *dicta.* See Auerbach (1985) for more details.

6 See note 11.

7 See Chapter IV of Detlefsen (1986).

8 Logic texts and the technical literature are often careless about the intensional aspect of interpretation. Mates contains a brief discussion of this. Boolos, in the chapter on Peano arithmetic in the forthcoming second edition of Boolos (1978), is explicit: "S expresses the commutativity of addition because it is, as we suppose, interpreted in accordance with the usual interpretation *N* of PA, *as we standardly give that* interpretation."

9 It need not be this way. One could directly formalize Syntax in its own suitable language and prove the Gödel theorems directly for (and in) it. As far as I know it is never done quite this way. However, some approaches are certainly in this spirit; Smullyan's various abstract versions of the Gödel theorems are based on stripped down formalizations of Syntax and the detailed framework for dealing with the Gödel results in a purely syntactical manner is supplied by his *Theory of Formal Systems. Computability Theory, Semantics, and Logic Programming* by Melvin Fitting is a recent modern treatment that avoids the arithmetic route.

10 Examples help:

$$\mathbf{AtForm}(x) \text{ is } \exists t < x\, \exists t' < x\, ((\mathbf{Term}(t) \wedge \mathbf{Term}(t') \wedge x = (\ulcorner = \urcorner, t, t')) \vee x = \ulcorner \bot \urcorner)$$

where $\ulcorner = \urcorner$ and $\ulcorner \bot \urcorner$ are the numerals for the Gödel numbers of those symbols.

11 If **F** is a sentence of PA, how do you write in PA that **F** is a theorem?

Well, **F** has a Gödel number; call it f. Furthermore, there is a term in PA that names f, in fact many. We define [**F**] as the *numeral* for f. The standard way to write in PA that **F** is a theorem is **Bew**([**F**]). It is not in general true, if **t** is a term that denotes f, that **Bew**([**F**]) ↔ **Bew**(**t**) is a theorem. If we restrict ourselves to terms for provably recursive functions then the biconditional is a theorem.

Now suppose that **F** is an open sentence of PA. As it stands, both **Bew**(**x** < **y**) and **Bew**([**x** < **y**]) are syntactic nonsense; we would like to give sense to such a formula so that we could say **Bew**(**x** < **y**) is true of ⟨2, 4⟩. Well, what do we want *this* to mean? Presumably that a certain sentence is a theorem. Not just any sentence with terms denoting 2, 4 and a formula whose extension is *is less than*, but the sentence **2** < **4**. So we make **Bew**([**F**]) be true of some sequence of numbers just in case the substitution of the standard numerals for those numbers into **F** results in a theorem. Note that machinery like this is necessary even to make sense of F-LPC and to define all the appropriate varieties of term substitution.

12 Note the following feature of the boldface mapping. Let **A** be a sentence of English in the vocabulary of the arithmetic interpretation of **T**; and let the arithmetic formula it interprets be **A**. Suppose **A** is a theorem of **T**. Now let S be a rewriting of **A** into the vocabulary of the syntactic theory of **T**, constrained (only) by co-extensiveness, with the presumed Gödel numbering as the basis. Facts: Any such S is true. S need not be **A**. Indeed, S need not be a theorem. Moreover, if **A** were a non-theorem, S might be a theorem.

13 My construction of **Bew**, based on faithfulness to the notion of reading the formula syntactically, descends from the rigorous account of Feferman (1960). This assures us, modulo a concern about the representation of the axioms, of F-LPC.

In Feferman's generalization of the Gödel Second Incompleteness Theorem, the boldface mapping (in Feferman dotting an expression corresponds to our boldface) of complex syntactic notions is achieved by straightforward transcription of their (often inductive) definitions. In particular, the derivability predicate is a complex formula that encodes a usual textbook definition of *is a derivation of*. The basis of such a definition is the set of axioms. This definition of *is a derivation of* is the same across all formalisms, save for reference to the axioms. (One assumes a fixed logical apparatus.)

How is reference to the axioms handled? Since there are, in the case of **PA**, infinitely many axioms, they are formalized via an open sentence. Many distinct open sentences will numeralwise express the same set of axioms; this creates the same state of affairs sketched above with respect to variant (and deviant) proof predicates. Only certain of the open sentences that numeralwise express the axioms of **T** *really* express the axioms of **T**. Feferman is able to characterize a property, being an "RE-formula," that guarantees correctness. The "RE" terminology comes from "recursively enumerable." In this case, it roughly means that the formula has the form of an RE definition; it *does not* mean (just) that the set picked out is recursively enumerable,

but that it is picked out in a way that guarantees that the extension is recursively enumerable. Feferman (1960) and Monk (1976) give the details.

Those who have slogged through Gödel's original paper, or some other *detailed* proof, will remember that a great deal of trouble is taken, not merely to define, in arithmetic, the right syntactic categories but to define them in certain ways: in particular, with bounds on the quantifiers. The purpose of this is to insure, not just that the sets defined are numeralwise expressible, but that the very form of the definition guarantees it. So a prover of the First Theorem shows that the definitions pick out numeralwise expressible sets by adverting to the form of the definitions. When the prover in question is **PA** itself, as in the context of the Second Theorem, we need a formalization of appropriate form. This, in effect, is what Feferman gives us with RE-formulas. An RE-formula is one that canonically, as a matter of form, picks out a recursively enumerable set.

This approach individuates formalisms by their "presentation" – and co-extensive presentations are not intersubstitutable in the context of the Second Theorem.

More precisely: if $\alpha(\mathbf{x})$ is a formula that numeralwise expresses the axioms of T, a proof predicate can be constructed in a standard way from α. Since many αs numerically define the same set of axioms, different formal proof predicates will be defined for the same axioms, one for each α. Deviant αs are bizarre ways of presenting the axioms – bizarre enough to carry a trivial assurance of consistency.

14 "Partially," because I do not think enough has been said about the initial steps in the assignment; in particular the role of numerals as proper names has been left unexamined here. I leave that for another paper, where I will take up a somewhat different defense of the derivability conditions. There I will argue that Mostowski should not have aimed at all truths and settled for some, but rather should have aimed at analytic truths and gotten them all.

15 Many different notions of subsystem will do here. Ralph need understand very little about the notion of subsystem; no more, in fact, than the bare terminology suggests. For concreteness the following will do: the nth subsystem is the formalism characterized by the axioms $< n$. **PA** is not finitely axiomatizable, and so will have infinitely many subsystems in this sense. Some more details are supplied below.

16 Cf. Visser [1989].

17 Detlefsen, in a private communication, has pointed out to me that Detlefsen (1990) contains a discussion of this point. Detlefsen emphasizes the conceptual separability of "locative" concerns (what are the theorems) and "quality-control" concerns (soundness). Given this separability it is open to the Hilbertian to demand different sorts of evidence in the two cases. Detlefsen makes the case that the Hilbertian need not give a finitary answer to *both* concerns.

18 See Bezboruah and Shepherdson (1976).

REFERENCES

Auerbach, D. (1985) "Intensionality and the Gödel Theorems," *Philosophical Studies* 48: 337–51.

Bezboruah, A. and Shepherdson, J.C. (1976) "Gödel's Second Incompleteness Theorem for **Q**," *JSL* 41: 503–12.

Boolos, G. (1978) *The Unprovability of Consistency: An Essay in Modal Logic*, Cambridge: Cambridge University Press.

Detlefsen, M. (1986) *Hilbert's Program: An Essay in Mathematical Instrumentalism*, Dordrecht: Reidel.

—— (1990) "On an alleged refutation of Hilbert's Program using Gödel's First Incompleteness Theorem," *Journal of Philosophical Logic* 28, in press. Reprinted in this volume as Chapter 8.

Feferman, S. (1960) "Arithmetization of metamathematics in a general setting," *Fundamenta Mathematicae* 49.

Hallett, M. (1989) "Physicalism, reductionism and Hilbert," in A.D. Irvine (ed.) *Physicalism in Mathematics*, Kluwer Academic, 182–256.

Monk, J. (1976) *Mathematical Logic*, Berlin: Springer-Verlag.

Visser, A. (1989) "Peano's smart children: a provability logical study of systems with built-in consistency," *Notre Dame Journal of Formal Logic* 30: 161–96.