DAVID D. AUERBACH

INTENSIONALITY AND THE GÖDEL THEOREMS

(Received 4 September, 1984)

I

Philosophers of language have drawn on metamathematical results in varied ways. Extensionalist philosophers have been particularly impressed with two, not unrelated, facts: the existence, due to Frege/Tarski, of a certain sort of semantics, and the seeming absence of intensional contexts from mathematical discourse. The philosophical import of these facts is at best murky. Extensionalists will emphasize the success and clarity of the model theoretic semantics; others will emphasize the relative poverty of the mathematical idiom; still others will question the aptness of the standard extensional semantics for mathematics.

In this paper I will investigate some implications of the Gödel Second Incompleteness Theorem for these positions. I shall argue that the realm of mathematics, proof theory in particular, has been a breeding ground for intensionality and that satisfactory intensional semantic theories are implicit in certain rigorous technical accounts. One moral to be drawn is that intensionality does not, as a matter of course, involve incoherence.

The central argument will conclude that an extensional semantics would attribute falsity to the Gödel Second Incompleteness Theorem. Since we have good reason to believe the theorem true, the correct inference is to the insufficiency of extensional semantics. The intensional context lurking in the Second Theorem is that of indirect discourse. This indirect discourse is, *contra* Quine, not "at variance with the characteristic objectivity of science". In fact, rigorous generalizations of the Second Theorem provide the proper non-extensionalist treatment.

Each of the matters just broached involves significant subtleties – even the matter of justifying the truth of the Second Theorem. Moreover, the relevant technical work is not well reported in the philosophical literature[1] and has only recently been assimilated into logic texts.[2] In what follows, therefore, the methodological and metamathematical surveys are necessary preliminaries.

Philosophical Studies **48** (1985) 337–351. 0031–8116/85.10

II

This section illustrates some methodological aspects of the interpretation of metamathematical results, the Gödel Theorems in particular. Among the persistent concerns of philosophers of mathematics are the implications of various technical results. This can be overt, as in 'What are the philosophical implications of A?', where A is some technical result. Such questions raise the further question: what theses are needed to get from A, a purely mathematical proposition, to B, a non-mathematical proposition? This latter question arises covertly when theorems are rephrased or glossed. The following sketch of an answer to one such question is only intended to highlight the special features of my subsequent treatment of the Gödel Second Incompleteness Theorem.

In 1930 Gödel proved that a certain formal system, which he called P, is either incomplete or ω-inconsistent. This is not what is generally referred to by 'Gödel's First Incompleteness Theorem'. The 1930 result gains importance because P is important and because the proof of the result is clearly generalizable. A theorem to the effect that a large and important class of formal systems shares the property of incompleteness with P would seem of greater import. A refinement of Rosser's is needed to yield the familiar Incompleteness Theorem.

(1) There is no consistent complete axiomatizable extension of Q.

(1)[3] expresses an up-to-date generalization of the results of the 1930s and certainly obtains for us the 'large' class of formal systems we asked for.

(1) is a provable mathematical result.

(2) Any sufficiently strong consistent formal system of arithmetic is incomplete.

(2) is often used as an expression of the Gödel result. Since (1) and (2) are not *prima facie* synonymous, nor does (2) look wholly mathematical, what warrants both the assertion of (2) and the claim that it is an expression of the First Incompleteness Theorem?

Getting from (1) to (2) is a special case of the problem I characterized above – getting from mathematical A to nonmathematical B. One thesis needed is: In this context 'x is an extension of Q' implies 'x is sufficiently strong'.[4] In fact the result extends to theories into which Q is interpretable, but this is a refinement which need not detain us.

Additionally, some account of 'of arithmetic' is required. Such an account should yield the thesis: Q is a formal system of arithmetic and all its extensions in the language of arithmetic are formal systems of arithmetic.[5]

Each such thesis, dredged up to create a non-enthymematic valid argument from (1) to (2), needs to be justified. Why is it true that Q and its extensions are 'of arithmetic'? I shall not pause over this interesting, and not so simple, question. The kind of theses I have been talking about (call them 'connective theses') often assert that a particular formalism is an adequate formalization of some notion. Church's Thesis is a good example. Other examples include claims that a formal semantics corresponds in a certain way to an intended interpretation. Semantics of first-order predicate logic/logical validity is such a pair, Kripke semantics/Leibnizian possibility another.

These connective theses are the additional premises needed to produce a valid argument from a mathematical theorem to a philosophical claim. What has often been stressed in regard to Church's Thesis is also true of many important connective theses – they are not mathematical truths and do not partake of mathematics' clarion certainty and precision. Associated with Gödel's Second Incompleteness Theorem are connective theses of a somewhat surprising character. Consideration of the Second Theorem is perhaps best begun by looking at the analogue to (2) – the proposition that is not a mathematical theorem.

(3) If T is a sufficiently strong consistent formal system of arithmetic, then any sentence of T that says that T is consistent is not derivable in T.

(3) is a relatively careful rendering of one of the many things that Gödel is often said to have proved. Sometimes it is put: 'The consistency of (a formal system of) arithmetic is underivable in (that formal system of) arithmetic'. In many texts such words are offered as glosses of some technical result (usually proved only sketchily, if at all) labeled 'Gödel's Second Incompleteness Theorem'. Since what are derivable, or fail to be, are the formulas of a formalism, I take it that (3) is an acceptable rendering of the slightly looser remarks.[6]

These looser remarks, or sometimes (3), are the usual material for philosophical writings concerning the philosophical significance, or consequences, of the Gödel Second Theorem.[7] Since (3) isn't the mathematically proved Second Theorem, such writings would be helped by an argument that (3)

is true – the Gödel Second Theorem being a premise in such an argument – as well as by evidence that the second Theorem is a special sort of premise in such an argument. By this last I simply mean something that would justify taking (3) as a gloss, not a mere consequence, of the Second Theorem. Thus, in discussing (1), the First Theorem, and its relation to (2), one notes that the additional premises are either definitional or trivial. (1) is thereby 'special', being neither definitional nor trivial. We need to take a closer look at the differences between the two Gödel theorems, to begin to uncover what stands to (3) as (1) does to (2).

III

An early reference to the Second Theorem is found in Gödel's 1931 paper, in which a proof of what he calls Theorem XI is sketched. What does Gödel's sketch of a proof of Theorem XI show? Briefly, an underivable formula is exhibited, different from the one (17Gen r) exhibited in the proof of his First Incompleteness Theorem. It should be recalled that in the proof of the First Theorem Gödel constructs a formula that he shows, on hypothesis of consistency of P, to be underivable (in P). That is, he shows that there is a proof that the consistency of P implies that a certain formula is underivable. Corresponding to this, Gödel's *proof*, there is a *derivation* in P of a conditional formula of P corresponding to Gödel's implication: a conditional whose antecedent is a sentence which is the formalization of the assertion that P is consistent, and whose consequent is the formal sentence saying that the Gödel sentence is not derivable. By the construction of the First Theorem this consequent is equivalent in P to the Gödel sentence itself. Hence, the conditional whose antecedent is the consistency sentence and whose consequent is the Gödel sentence is a theorem of P. And since *modus ponens* is a rule of inference of P, the formalized statement of consistency cannot be derivable in P if P is consistent.

This is a rough and deliberately innocent sketch of the idea of the proof of XI. A detailed proof would involve constructing (or, at least, showing how to construct) the crucial formal derivation. This would include making adequate sense of the notion of formalization that infests the above sketch; it is here that the complications of the Second Theorem reside. However, the semantic flavor of this theorem can be preliminary appreciated by noticing that what is underivable is a formula and that this formula is *said to say* that the formal-

ism is consistent (recall (3)). Unlike the situation with some of the informal intuitive descriptions of the First Theorem (and with my description of the 'consequent' above) that are often given, this apparent semantic character is, I shall argue, intrinsic. The constraints on the antecedent are different in kind from those on the consequent.

To draw the contrast between the two Theorems more sharply: The First Theorem predicates a simple syntactic property of members of a large class of formal systems. This property, incompleteness, is simple in at least the following respect – it is definable in terms that do not invoke anything akin to a translation, or formalization, relation; T is *incomplete* just in case there is a sentence of T such that neither it nor its formal denial is derivable in T.

The Second Theorem (as glossed by (3)) predicates a certain more complex property of members of a large class of formal systems. Call this property 'KURT'. T has KURT if and only if any sentence of T that says that T is consistent is not derivable in T. We will be seeing what it takes to render this definition coherent, for 'T' taken as a variable.

What is the underivable formula of the Gödel Second Incompleteness Theorem? Let us call it, restricting our attention for the moment to just one formalism, CON(P)[8]. Call the underivable formula of the First Theorem 'G'. A proof of the Second Theorem establishes that $\vdash \mathrm{CON(P)} \rightarrow \mathrm{G}$ (in fact $\vdash \mathrm{CON(P)} \leftrightarrow \mathrm{G}$) and so, not $\vdash \mathrm{CON(P)}$. In the light of these facts what would make CON(P) importantly different from G?

Let $\mathrm{Pf}(y, x)$ be an open sentence of P that numeralwise expresses[9] the proof relation (assuming some fixed Gödel numbering of syntactic objects, and sequences of them, satisfying the usual constraints) and $\mathrm{Fm}(x)$ an open sentence of P that numeralwise expresses sentencehood. Using PF (and a symbol representing a substitution function), a proof of the First Theorem constructs G, a formula equivalent in P to $\forall y \neg \mathrm{Pf}(y, \overline{g})$, where g is the Gödel number of that very formula, G. $\exists x(\mathrm{Fm}(x) \,\&\, \neg \exists y\, \mathrm{Pf}(y, x))$ mimics, in quantificational structure, a standard definition of consistency. Let such a formula be (temporarily) CON(P). Note that this does not pick out a particular sentence G nor a particular sentence CON(P) since many formulas of P will numeralwise express the proof relation. This produces a fatal equivocation in the preceding paragraph.

This construction does not guarantee that such a CON(P) says that the formalism is consistent. It is not essential to a proof of the First Theorem that the *Gödel sentence* say of itself that it is underivable. (Indeed, it is only

the case the Gödel sentence is equivalent in Q to a sentence H which asserts *something* about G. But even this need not be that G is underivable.)[10] That it may seem to be one is an artifact of certain informal, motivating, semantic accounts of the First Theorem. Certain entailments of the proposition that G says of itself that it is underivable are all that are used and these are sufficiently captured by the relation of numeralwise expressibility. The plausibility of regarding G as saying of itself that it is not derivable arises from considering the standard interpretation, the Gödel numbering, *and* regarding Pf as expressing the proof relation. In the case where Pf only satisfies the entailments mentioned, however, this last is at best a pun on 'numeralwise expressible'. A sentence so constructed is true (in the standard interpretation) if and only if it is not derivable; but nothing stronger than this extensional agreement is forthcoming.[11] This is all the First Theorem requires, even in the general case where P is replaced by an arbitrary sufficiently strong theory T.

Requiring only extensional agreement with respect to the proof predicate will not, however, in the general case, yield extensional agreement with respect to the consistency statement. That is, for some T, and some CON(T) constructed as above, it is false that: CON(T) is true if and only if T is consistent. Moreover, even in the case of P there are two sentences, $CON_1(P)$ and $CON_2(P)$, constructed as above, but such that they are not logically equivalent; and one of them is derivable in P.

IV

I have claimed that the difference between G and CON(P) is that *a* G may be constructed from an open sentence that numeralwise expresses the proof relation and that numeralwise expressibility is the only constraint needed to show such a G to be underivable; *a* CON(P), constructed as above, may be derivable. Concerning *saying that*, I have pointed out that the plausibility of regarding G as saying that G is not derivable arises, *inter alia*, from regarding Pf as expressing the proof relation, but that this plays no role in establishing (1) or (2). But suppose we take this plausibility at face value and benefit from the heuristic value of being able to explain the workings of G1 by adverting to the simple logic of the antinomies. Then a CON(P), constructed as above, should say that P is consistent. But then, I have claimed just above and will show below, either (3) is false or the account of *says that* is defective. Furthermore, abandoning all accounts of *says that* would eviscerate

the Second Theorem. I will now establish and illuminate these claims.

Rosser exploited numeralwise expressibility's inability to distinguish coextensive relations. By constructing a new open sentence that numeralwise expresses *y is a proof of x* but which had special properties as well, Rosser was able to improve the Gödel result. Letting Pf numeralwise express the proof relation, Rosser used

$$\mathrm{Pf}(y, x) \,\&\, \neg \exists y(y < x \,\&\, \mathrm{Pf}(y, \mathrm{neg}(x))).$$

This reads '*y* is a proof of *x* and there is no shorter proof of the negation of *x*'. A moment's reflection reveals that, for a consistent formalism, this numeralwise expresses what Pf does. An even simpler device will make the derivability of "consistency" more blatant. Define Pf′ as

$$(4) \qquad \mathrm{Pf}(y, x) \,\&\, \neg \mathrm{Pf}(y, \bar{k})$$

where k is the Gödel number of '0 = 1'. If the formalism is consistent (4) numeralwise expresses what Pf(y, x) does. Consider the consistency schema (5).

$$(5) \qquad \neg \exists y \Phi(y, \bar{k})$$

Then, although the instance of (5) involving Pf is not derivable, the instance in which Pf′ replaces Φ, $\neg \exists y(\mathrm{Pf}(y, \bar{k}) \,\&\, \neg \mathrm{Pf}(y, \bar{k}))$, is a theorem of logic. What isn't derivable in the formalism is that Pf and Pf′ are coextensive and so numeralwise express the same relation.

In a sense, the G's constructed from such deviant Pf conditionally assert their own underivability. The condition is that the formalism is consistent. This condition the formalism cannot discharge. (These intuitive readings are discussed below.) The *prima facie* incorrectness of '*y* is a proof of *x* and *y* isn't a proof of 0 = 1' as expressing *y is a proof of x* is not misleading; a proper account of the Second Theorem must respect this fact.

V

The quarantining of the deviant predicates is accomplished in rigorous accounts of the Second Theorem that prove generalized versions of it. It does not follow as a matter of course that such accounts yield (3). ((3) If T is a sufficiently strong consistent formal system of arithmetic, then any sentence of T that says that T is consistent is not derivable in T.) A purely mathemati-

cal result about formalisms may, in producing an underivable formula for each formalism, clearly be a generalized Second Theorem because the result is a rigorous version of the sketch on page 340. But it remains to justify such a result's rejection of certain extensionally correct proof predicates.

Technical accounts of the Second Theorem may vary as to how manifest is the relation between their machinery and our semantic concerns. It is striking that they all bear some familiar hallmark of intensionality. What are some of these accounts?

The earliest treatment occurs in [Hilbert-Bernays, 1939], wherein three derivability conditions are enumerated and used to prove a rigorous version of the Second Theorem. Any proof predicate that satisfies the derivability conditions (*one* of which is essentially numeralwise expressibility) will suffice for the Second Theorem; the labor is in showing that any particular proof predicate does satisfy them. This approach is continued in the work of Löb (1955) and in recent work on modal systems, where $\Box$ is interpreted as *derivable in a formal system.*[12]

The Hilbert-Bernays derivability conditions were conditions on the formal proof relation, stricter than numeralwise expressibility, that were sufficient to guarantee an unequivocal Second Theorem in the following sense: Any proof predicate meeting the three conditions would satisfy the requirements of the proof of the Second Theorem; moreover, the proof relations constructed for familiar theories were seen to satisfy the conditions.

The work of Feferman (1960) and Jeroslow (1965) was stimulated by remarks of Kreisel (1965).[13] Feferman points out that merely "numerically correct" proof definitions are inadequate for certain results. Results for which they are adequate he calls "extensional", the rest "intentional". The deviant proof predicates lead to useful extensional results (e.g. Rosser) while others have no intrinsic interest (a provable "consistency" sentence). For Feferman the weakness of the Hilbert-Bernays approach is that verifying whether a particular predicate satisfies the conditions is laborious.

Feferman presents a large class of formal systems and proves the Second Theorem for them. The key to his approach is the notion of a formal system that he employs. The consistency sentence for any system is built from the proof predicate in some standard way (by a straight-forward transcription of any one of the equivalent definitions of consistency); the proof predicate is straight-forwardly transcribed from the *presentation* of the formal system. The trick is in obtaining a formal object to represent the presentation.

More precisely: If $\alpha(x)$ is a formula that numeralwise expresses the axioms of T, a proof predicate can be constructed in a standard way from α. The phrase 'a consistency sentence for a formal system' makes sense only if one individuates formal systems more narrowly than by their theorem sets. By narrowing the individuation it becomes possible to generalize various metamathematical results by conditions on the formulas α.

Since many α's numerically define the same set of axioms, different formal proof predicates will be defined for the same axioms; one for each α. Deviant α's are bizarre ways of presenting the axioms – bizarre enough to carry a trivial assurance of consistency (recall schema (5) with Φ replaced by Pf').

Feferman constructs a sentence $G\alpha$, for each formula α that numerates an extension of P, such that $G\alpha$ is underivable in the system corresponding to α. $G\alpha$ corresponds to the usual Gödel sentence. (5.6) is a simplified version of Feferman's generalization of the Second Theorem.

(5.6) Let α be a formula that numerates in Q the axioms of a consistent extension of P. Call this extension A. If α is an RE-formula $\vdash_A \mathrm{CON}_\alpha \rightarrow G$ and hence not $\vdash_A \mathrm{CON}_\alpha$.

(5.6) differs from a blindly extensional generalization of the Gödel result only in the restriction that α be what Feferman calls an RE-formula. While the precise definition of this does not concern us here what is crucial is Feferman's (5.9), which establishes that the purely extensional generalization, (5.6) without the restriction on α, is false.

(5.9) Let A be a certain kind of extension of P. There is an α^* numeralwise expressing A's axioms in A, such that $\vdash_A \mathrm{CON}_{\alpha^*}$.[14]

Not surprisingly, Feferman's α^* is a close relative of the deviant proof predicates described earlier. $\alpha^*(x) = \alpha(x) \,\&\, \forall z(z < x \rightarrow \mathrm{CON}_{\alpha|z}) \,\&\, \mathrm{StK}(x)$, where α numeralwise expresses A's axioms in Q, and Stk represents *is a sentence*. It can be read: x is an α-axiom and all the finitely axiomatized subsystems formed by the axioms shorter than x are consistent. A facile, but not essentially misleading, way of reading the corresponding 'consistency' sentence is: The largest consistent subsystem of A is consistent. 'The largest consistent subsystem of A' denotes A, given that A is consistent; but, of course, it is just this fact that is not given to A itself (cf. the end of the section IV).

In fact, then, the notation 'CON_P' is dangerously vague, unless one has a particular presentation of the axioms in mind. 'Its' formalization, 'CON_α',

makes this clear; particularly as there are (non-RE) α's that render CON_α derivable in P. The RE α's are not an *ad hoc* grouping; it can be argued (though not here) that the definition of RE-ness meshes perfectly with our intuitions about what formalisms are.

VI

The problem then is to rescue the plausibility of construing formulas as making metamathematical remarks, without falsifying (3). Taking any open sentence of arithmetic whose extension, in the standard interpretation, is the extension of *is a derivation of* (*via* some admissible Gödel numbering) as expressing the formal proof relation marks a commitment to an extensionalist semantics. The cavils of the earlier sections show that such a semantics would indeed fail to do justice to KURT. A semantics based on the narrower individuation of predicate expressions that, for example, Feferman's treatment mandates, will do justice to KURT[15]. It is worth emphasizing again that Gödel's Theorem XI cannot be proved in generality without somehow stigmatizing the deviant proof predicates.

The skeptical reader, wary of intensional semantics in general, may regard a coherent applied intensional semantics with suspicion. I will in what follows try to alleviate such suspicions.

It is the common meta-theoretic ground of many programs for semantic theory, and the centerpiece of Davidsonian approaches, that the truth-value assigned to a sentence depends only upon the semantic value of the parts. Of course, much controversy surrounds the proper elucidation of this gnomic remark. Intensionality is ascribed if the semantic value of a part needs be something other than the extension of that part. It is patent here that the notion of part is crucial to such ascription. Frege argued for such a duality of semantic values by finding contexts that produced a difference in truth-value despite co-extensiveness of parts. Russell's theory of descriptions can sometimes be used to avoid ascriptions of intensionality by a change in the ascription of parthood.

Consider the language of elementary proof theory. Evidently it contains the predicate 'is a derivation-in-T of _____', for each formalism T. (Everything that follows will apply as well to any synonymous predicate, whether couched in Japanese, German, or English.) Using such a predicate we can express consistency – 'There is no derivation-in-T of $\perp$' where $\perp$ is a favorite

contradiction of a theorem of T. Call this sentence Con(T). We are interested in the semantics of this piece of (technical) natural language. In particular, is the meaning of Con(T) given by the extension of its parts? Not if we wish to express (3) in this language. How did the technical details reveal this?

Con(T) is not a candidate for being underivable-in-T. We do want to show that it is *unprovable* in T by appropriately formalizing the language of elementary proof theory and producing a sentence of T, say CON_α, whose underivability permits the inference to the unprovability of Con(T). This is precisely the problem of page 340 – establihsing (3) on the basis of a certain technical fact. Let us look at this sort of inference from a wider point of view.

We say that it is provable in P that $2 < 3$ because a certain formula of P is derivable and the standard interpretation for P makes the appropriate link between the formula and the manner in which the standard model is described. Mates[16], for example, discusses the reasons for this last clause. He points out that in establishing instances of Tarski's schema T, the way in which the interpretation is described is utilized. The *same* interpretation, I, *given differently*, yields both

'La_1a_2' is true under I iff 2 is less than 3.

and

'La_1a_2' is true under I iff the only even prime is less than 3.

as consequences of the definition of truth in an interpretation. For purposes that exceed mere consideration of truth conditions in an extensional language, the non-identity of the two displayed sentences is vital. One such purpose, ubiquitous in logic texts, is judging whether a formal sentence is an adequate rendering of an English sentence; and this is relative to the way in which the interpretation is given.

The context we are considering (that $2 < 3$ is provable in P) is very like the translation context that Mates is concerned with. He points out that the truth-value of wffs is independent of the manner of specification of the interpretation, though the meaning is not. The Fregean move involves locating a context that is sentisitive to the manner of specification, but in the language being interpreted; this will produce a difference in truth-value, not just in meaning. The Second Theorem provides such a context; the proof of the Second Theorem requires significantly more than correct extension be true

of the provability predicate. The language of elementary proof theory, in as much as it contains (3), is intensional.

The skeptic may see the 'says that' idiom of (3) as begging the question, and look to (6) as expressing the content of a generalized Second Theorem.

(6) If T is a consistent formal system of sufficient strength, it is not provable in T that T is consistent.

It should be clear by now that this won't help. A cogent argument for (6) will have recourse to a proof of a generalized Second Theorem. Viewed as a semantics the Feferman treatment assigns *true* to (6) by virtue of (5.6). The key to the linkage is the occurrence of the formulas α as part of the formalized consistency sentence. The α's are assigned by the semantics as representatives of formal systems. We can think of them as fixing the reference for proper names of the formalism. (5.9) demonstrates that a (6) whose semantics treated co-extensive α's indiscriminately would be false.

The skeptic might now object to the implicit syntax, that is, the parts, and say that the superficial form of (6) is misleading.

(7) If T is a formal system of sufficient strength, the sentence CON_α is not derivable in T.

(7) might be offered. The skeptic goes on to say that an appropriate technical account, says Feferman's supplies the sentences CON_α. This account must, of course, bow to the requirements set by (5.9); this will force the skeptic to identify formal theories with RE α's.

Having been forced to the bifurcation (RE/non-RE) of the class of α's numeralwise expressing each set of theorems, the skeptic has sacrificed two things: 1) an explanation of the bifurcation and, more importantly for present purposes, 2) a reason for regarding (7) as a remark about consistency. The skeptic is unable to distinguish the First and Second Theorems in any interesting way (cf. Section III). For if (7) is to be about consistency, by way of CON_α being a consistency sentence for formalism T, then a semantics is required – one which (5.9) tells us makes the truth of (7) depend on something other than the extensions of the parts of CON_α.

If a semantic theory for the language of elementary proof theory allowed a non-RE α to be equivalent to some RE α then we would see that it had to be an incorrect theory. For then it would be part of the meaning of 'formalism' that a formalism be consistent. It is part of the charm of the Gödel

Second Incompleteness Theorem that it entails that such a semantic theory would not only be empirically false, but simply inconsistent.

In sum, recalling the remarks at the end of section II, (1) is to (2) as (5.6) is to (3). The language of elementary proof theory has a coherent intensional semantics that certifies this ratio. An extensionalist semantics is neither desirable nor necessary for all of mathematics.[17]

NOTES

[1] Resnik (1974) and Detlefson (1979) are recent exceptions that take account of the special features of the Second Theorem.

[2] See Boolos–Jeffrey (1974), Boolos (1978) and Monk (1976).

[3] (1) first appears in Tarski (1960).

[4] The *converse* implication supports a valid deductive argument from (1) to (2). Unfortunately the converse implication is false; weaker theories than Q are sufficiently strong. The suggested implication is true, but not sufficiently helpful. A complete analysis would bring out the contextual nature of 'sufficiently strong'. Alternatively, one could use the converse implication, replacing the predicate 'is an extension of Q' with 'has those features of Q that are relevant to the proof of G1', or some illuminating coextensive predicate.

[5] If (2) strikes one as mathematical, note that all that is really needed for my point is that passage from (1) to (2) is mediated by theses unsupported by mathematical evidence. (2) seems mathematical because the mediating theses seem definitional.

[6] 'Derivable', 'underivable', etc. will be used in connection with formulas of a formalism; 'provable', 'proof', etc., are reserved for the ordinary notions of unformalized mathematical practice. I occasionally violate this convention, for the sake of custom, in the context of discussing *is a proof of* and the 'proof' predicate. Strict speaking would demand *'is a derivation of'* and 'derivation predicate'. Note that the looser remark cited in the text offers little obvious guidance regarding a choice between 'underivable' and 'unprovable'.

[7] See Note 1, above.

[8] P is the well-known Peano arithmetic, so-called because of its name. Q is a well-studied theory in the language of arithmetic. It has finitely many axioms (seven, all simple and clearly true in the standard model), all recursive functions are representable in Q, and yet it is a rather weak subtheory of P (that addition is commutative is not a theorem of Q). There are even weaker theories that will suffice for the First Theorem; Q's virtue is its finite axiomatizability. See Tarski (1960) and Boolos (1978). For now 'CON(P)' is a simple term.

[9] Numeralwise expressibility is a three place relation among formal systems, relations among or properties of numbers, and predicates of formal systems. A formal predicate that numeralwise expresses a relation in an arithmetically correct formal system is thereby guaranteed to be extensionally correct with respect to that relation. Φ numeralwise expresses R, R an m-place relation, iff

(i) if $R(\vec{n})$, then $\vdash \Phi(\vec{\overline{n}})$

(ii) if $\neg R(\vec{n})$, then $\vdash \neg\Phi(\vec{\overline{n}})$,

where '$\bar{}$' denotes the function from numbers to their standard numerals in the formalism. All recursive relations are numeralwise expressible in Q and its extensions, including, of course. P. Φ is said to *numerate* R, in consistent formalisms, if $R(\vec{n}) \leftrightarrow \vdash \Phi(\vec{\overline{n}})$.

[10] Of course G implies that its formalism is consistent – provided that G really does say that G is underivable and, *a fortiori*, that something is underivable. Any formula that says that something is underivable, either existentially or using a canonical name, is a consistency sentence. That is, G and CON(P) are *not* importantly different provided that G really does say that G is underivable. Kleene (1950), p. 211 has a similar remark, but with the semantic proviso on G hidden in Kleene's use of 'intuitively' and 'intuitive'.

[11] A typical proof of the First Theorem enforces this extensional agreement by constraints on the open sentences used to build up G; in particular, the constraint that Pf numeralwise express *is a proof of*. This allows such a proof to be purely syntactic, in that it need not mention the standard interpretation. The use of standard numerals in the definition of numeralwise expressibility is, however, crucial.

[12] See Boolos (1978).

[13] The cleanest presentation of the technical facts is in Monk (1976), pp. 298–307. Feferman (1962), pp. 265–272, contains a useful précis of the technical material in Feferman (1960). In Monk (1976) the RE-formulas are not mentioned as such, but the remark on p. 304 concerning '$^{\cup}(g^{*}{}^{*}\Delta)$' does as well. Monk's text is particularly nice for its dovetailing of the Löb and Feferman material.

[14] For a finitely axiomatized theory there is a unique, principled, best choice for an α. For finitely axiomatized A, this α is written [A] and [A] = the obvious formalization of '$x = a_1 \vee x = a_2 \vee \cdots \vee x = x_n$' where $a_1, \ldots, a_n$ are the (Gödel numbers of the) axioms. *A is reflexive* just in case for each finite $F \subset A$, $\vdash \mathrm{CON}_{[F]}$. Feferman's G_α, CON_α, etc., are specific constructions defined over a broad class of canonically presented formalisms, schematic only in α. No theory with induction can be finitely axiomatized, no theory without induction can formalize formalisms – so the little fact here is of no help with consistency sentences.

[15] Jeroslow's approach, though intertranslatable with Feferman's, is more direct. Jeroslow avoids the standard encodings of the usual primitive recursive syntactic relations and functions, whereas Feferman presents a generalized theory of those functions and relations. Jeroslow specifically represents formal systems as Post Canonical Systems: "Formal logics are not usually understood as Post Canonical Systems, but there is a natural, uniform procedure for viewing them as such, provided that *all* the mechanical rules which constitute the formal logic are specified, even the inductive rules for generating the terms, formulas, etc. The idea here is that the predicates of proof theory are always inductively defined, and Post Canonical Systems are the language of inductive definitions *par excellence*." Post Canonical Systems thus formalize the presentations of formal systems given in logic books. As the Feferman approach is certified as semantically correct by arguing that the RE/non-RE distinction is principled, the Jeroslow account is certified by arguing for Jeroslow's Thesis: PCS's are the best representations of formal systems. The details of such argumentation I leave to another paper.

It is in the spirit of the origins of formal systems as an object of study that they be regarded as systems for generating syntactic objects, in categories, independently of their intended meaning. Note that, from this point of views, axiom schemata have no place, as such; non-finitely axiomatized theories are to be identified with "the finite number of rules which describe the generation of the infinite number of axioms" (Jeroslow, 1971). Not only is this accord with the view of formal systems as combinatorially secured producers of theorems, but also can be connected to the epistemological motives behind a Hilbert-style program. Kreisel (1965) adduces some additional conceptual grounds for Jeroslow's Thesis in pointing out that we often wish to distinguish formal systems by their rules, and not by their theorems or even their set of proofs. Typical contexts that require such a fine-grained distinction of theories are evidentiary ones. One formulation of a set of theorems may be evident (i.e., evidently true) and hence foundationally sound and another not. Moreover, the establishment of their (extensional) equivalence may not be evident.

[16] Mates (1972), pp. 75–78.

[17] An ancient ancestor of this paper existed in 1976. I have since benefited enormously from the comments of George Boolos, Harold Levin, Michael Resnik and most especially Mark Richard.

BIBLIOGRAPHY

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

Boolos, G. and R. Jeffrey: 1974, Computability and Logic (Cambridge University Press, Cambridge).

Detlefson, M: 1979, 'On interpreting Gödel's Second Incompleteness Theorem', Journal of Philosophical Logic 8, pp. 297–315.

Feferman, S.: 1960, "Arithmetization of Metamathematics in a General Setting", Fundamenta Mathematicae XLIX, pp. 35–92.

Feferman, S.: 1962, "Transfinite Recursive Progressions of Axiomatic Theories", JSL XXVII, pp. 259–316.

Hilbert, D. and Bernays, P.: 1939, Die Grundlagen der Mathematik, 2nd ed. (Springer-Verlag, Berlin).

Jerolsow, R.: 1971, 'On Gödel's consistency theorem', unpublished manuscript.

Jeroslow, R.: 1973, 'Redundancies in the Hilbert-Bernays derivability conditions for Gödel's Second Incompleteness Theorem', Journal of Symbolic Logic XXXVIII, pp. 359–367.

Kleene, S. C.: 1950, Introduction to Metamathematics (Van Nostrand, Princeton).

Kreisel, G.: 1965, 'Mathematical logic', in T. L. Saaty (ed.), Lectures on Modern Mathematics III (Wiley, New York).

Lob, M. H.: 1955, "Solution of a Problem of Leon Henkin", Journal of Symbolic Logic XX, pp. 115–118.

Mates, B.: 1972, Elementary Logic, 2nd ed. (Oxford University Press, New York).

Monk, J. D.: 1976, Mathematical Logic (Springer-Verlag, New York, Heidelberg, Berlin).

Resnik, M.: 1974, "The Philosophical Significance of Consistency Proofs", Journal of Philosophical Logic 3, pp. 133–147.

Tarski, A., A. Mostowski, and R. Robinson: 1960, Undecidable Theories (North-Holland, Amsterdam).

Department of Philosophy and Religion,
North Carolina State University,
Box 8103,
Raleigh, NC 27695-8103,
U.S.A.