EXPRESSING CONSISTENCY:

GODEL'S SECOND INCOMPLETENESS THEOREM AND

INTENSIONALITY IN METAMATHEMATICS

by

DAVID D. AUERBACH

B.S., City College of New York
1969

SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF

PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1978

Signature of Author.....
Department of Philosophy and Linguistics, May, 1978

Certified by.....
Thesis Supervisor

Accepted by.....
Chairman, Departmental Committee

IN MEMORY OF

OLA AND JOSEPH AUERBACH

and

HANNAH RUBIN

PREFACE

In the fall of 1971 I attended a course given by Burton Dreben. Professor Dreben asked the class to show the provability of a consistency sentence formed using Rosser's proof predicate. I did, and was charmed and intrigued. With only the vague idea that deep issues lurked in the shadows, I attempted to coax into greater illumination the glimmer of intensionality I thought I espied. George Boolos encouraged me in the belief that I had lit upon an interesting topic and kindly consented to supervise the dissertation. Later, in detailed criticisms of drafts, he was to transubstantiate much incoherence into coherence. The residual incoherence is mine.

My colleagues and friends Richard Nagel and Harold Levin were constantly available sources of valuable criticism, insight and food. Professor Levin in particular provided steady stimulation concerning matters both logical and philosophical. Fred Katz, in the course of many discussions, aided my thinking during the early drafts. To Richard Cartwright and Jerrold Katz I owe somewhat vaguer, but no less important intellectual debts. Professor Cartwright also prompted some specific clarifications of portions of the text; he was kind enough not to demand more.

R.G. Jeroslow generously sent me his extremely insightful unpublished work on encodings.

I would also like to acknowledge the importance of Adrian Piper, Terry Vance, and Randy and Leilani Carter to the writing of this dissertation. Were it not for their influence, I, and it, would be very different.

TABLE OF CONTENTS

Chapter 1

PRELIMINARY METHODOLOGICAL REMARKS

Mathematics and its recent offspring, metamathematics, are often admired for their precision and clarity. The study of other matters has frequently been modelled on the techniques of mathematics - and the failure of some such attempts is sometimes supposed to reflect ill on the subject matter. Nonetheless, this vaunted precision and clarity has not typically been carried over into philosophical discussions concerning mathematics; nor is there anything approaching unanimity regarding basic problems in the epistemology and ontology of mathematics. In view of the fact that metamathematics has been developed as a tool for the study of mathematics itself, one might hope that careful consideration of major metamathematical results would be philosophically helpful.

In fact, much has been written on the "implications" of various metamathematical results. Such writings often suffer from methodological defects. It is rare that a metamathematical result will entail an interesting philosophical thesis. This may happen -- when a sufficiently precise philosophical position entails the denial of a metamathematical result. And it may also occur when the philosophical

position involves mention of a specific piece of mathematics - the Hilbert program is a ready example of this.

What is left out of, or only implicit in, many arguments to "implications" of metamathematical results are theses connecting the result to the subject matter. The unique status of the Hilbert program is accounted for by mention of such a mathematical thesis being an explicit part of the program. The suppression of such theses sometimes obscures the fact that metamathematical theorems are mathematical results about mathematical objects. If the necessary connecting theses are supplied, viewed as previously suppressed premises, the resulting argument can appear question-begging - and hence of no help in convincing the non-believer[1].

The value to philosophy of technical considerations lies in part in making various intuitive concepts precise. Much of the value of metamathematical results consists, not just in answering certain foundational questions, but in giving them a precise sense. The connecting theses to which I refer are often claims to the effect that a precise explication of some concept is correct. (Webb's "Gödel's Theorem and Church's Thesis: A Prologue to Mechanism" contains a suggestive historical account of the simultaneous clarification of foundational concepts and technical development of those concepts.)

---

[1]If I understand Kreisel, a concern with such theses is a theme of his writings; Benacerraf also makes this point.

I will show that certain necessary connective theses associated with Gödel's celebrated Second Incompleteness Theorem have a somewhat surprising character. Before I uncover these theses and reveal their character, I am going to sketch a modest and (I hope) uncontroversial schematic of a philosophical treatment of Gödel's First Incompleteness Theorem. This will serve two purposes. First, the treatment will illuminate the above methodological remarks. Second, the simplicity of the treatment will contrast with the difficulties that accompany an attempted similar treatment of Gödel's Second Incompleteness Theorem; consideration of these occupy the bulk of this paper. The rigorous resolution of these difficulties eventuates in a novel conclusion concerning the nature of an adequate semantics for a certain portion of mathematics.

In 1930 Gödel proved that a certain formal system, which he called P, is either incomplete or 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 candidate for greater importance would be a theorem to the effect that a large and important class of formal systems shares the property of incompleteness with P. Although Gödel does so extend his result in that same paper, throughout this paper I will suppress references to primitive recursive extensions of P.

Such a theorem requires for its best expression the theory partly[2] discovered by Gödel in the 1931 paper: the theory of recursive relations. Furthermore, a refinement of Rosser's is needed to yield the Incompleteness Theorem we all know and love[3].

Let us call this result G1. What is G1? (1)

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

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. Justifying the 'important' of 'large and important class of formal systems' is another matter.

(1) is a provable mathematical result. (2)

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

---

[2]'Partly' both because Gödel discovered a part of the theory (primitive recursive functions not general recursive) and because he was partly responsible for the discovery, along with Herbrand, Kleene, Turing, et al.

[3]Not every Gödel sentence for a theory is undecidable in that theory, though every Rosser sentence is, in every consistent extension of P.

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 Incompleteness Theorem?

A partially satisfactory argument from (1) to (2) can be obtained. When (1.1) and (1.2)

(1.1) If a formal system is sufficiently strong it is an extension of Q.

(1.2) Every formal system is axiomatizable.

are joined with (1) as premises, then (1.9)

(1.9) Every sufficiently strong consistent formal system is incomplete.

follows, and a fortiori, (2) follows*.

---

*I have not said what a formal system is. For now it is enough that a formal system can be given by an axiomatization. This fails to yield individuation criteria; for immediate purposes we may have in mind the set of theorems generated by the axiomatization. I am presupposing that either notational variants don't count, or that the underlying syntax is abstract.

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 (commutativity of addition is not a theorem of Q). R is a subtheory of Q, and, though very

We have had to stipulate that (1.1) is true; this is unfortunate since the significance of 'sufficiently strong' is often taken to be all recursive functions are representable in[5]. Thus (1.3) and (1.4)

(1.3) If a formal system represents all recursive functions then so do all of its consistent extensions, and Q represents all recursive functions.

(1.4) A formal system is sufficiently strong if and only if it represents all recursive functions.

are not sufficient to yield (2); whereas it is false that only Q and its extensions represent all recursive functions.

---

weak, has infinitely many axioms; all recursive functions are representable in R. P is the famous Peano arithmetic.

The talk in the preceding paragraph nominally opts for a narrower individuation of entities such as P, Q, and R than that yielded by same theorem set. For, I talk of R as having infinitely many axioms. I shall ultimately be arguing that for many important contexts of technical proof theory even finer individuation is needed. However, in the current context this need does not arise, and we may talk of axiomatizations of theories, if need be. In the case at hand the matter is even simpler, since R (as a theorem set) has no finite axiomatization; nor does P. P, Q, and R are fully characterized and their important properties revealed in [TaMoRo].

[5] A more plausible alternative to (1.1) would be to regard 'sufficiently strong' as having indexical properties. So its occurence in (2) is to be explicated by (1.1) - though its meaning in general is more like: is an extension of an appropriate formal system.

R, for example, is weaker than Q and suffices. Another undesirable feature of this argument is that 'of arithmetic' is only smuggled in *via* an *a fortiori* clause when what is wanted is the observation that Q is "of arithmetic" and all its extensions (in the language of arithmetic) are "of arithmetic"[6].

An easy, albeit ad hoc, modification replaces (2) by (3),

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

plus the remark that there is no formal system that is interesting, represents all recursive functions, is "of arithmetic", and not an extension of Q. 'Interesting' here is a term of art.

(1.9) carries no presupposition that there are formal systems of arithmetic. The thesis that Q and its extensions (in the language of arithmetic) are formal systems of arithmetic is needed to make (2), and not (1.9) the appropriate expression of an important fact. These days we briefly say that a theory is "of arithmetic" if its sentences are true in the standard model. This latter claim is a substantial one itself. The burden of justifying the

---

[6]A slightly different approach to (1) and (2) can be found on page 182 of [B&J]. I take (3) to correspond to [B&J]'s remark that Q is a rather weak theory.

claim that Q is "of arithmetic" has been shifted to the justification of the claim that being true in the standard model is sufficient for a theory to be "of arithmetic". Dedekind's isomorphism theorem and his existence proof for infinite models both represent early attempts to connect the natural numbers with the standard model (cf. [Webb]). Connective theses often assert that a particular precise formalization is an adequate formalization of some notion. Church's Thesis is the paradigm of this ilk[7]. Other examples include claims that a formal semantics corresponds in a certain way to an intended interpretation (first-order semantics and logical validity (cf. [Kreisel]); Kripke semantics and Leibnizian possibility).

Nor need the intended interpretation be even partially articulated prior to the creation of the formalization. Often the value of formalization is in retroactively articulating, not just clarifying, intuitive conceptions. Here I am thinking of Z-F and the iterative conception of sets (see [Boo2os]). Finally an example central to twentieth century logical investigations is the claim that set theory ade-

---

[7]As Webb stresses, Dedekind created the early paradigm. The analogy is: <u>of type</u> ω is to the standard model as being mechanical computable is to Turing computability. The analogy runs deep, though Dedekind's thesis has had a history of more immediate acceptance. I am glossing over the fact that both Turing's and Dedekind's Thesis, when subject to careful and detailed analysis, can be prised into non-equivalent versions.

quately formalizes (in yet another use of 'formalizes') mathematical practice. (see [MONK]).

These connective theses are the additional premises needed to produce a valid argument from a mathematical theorem to a philosophical claim. Church's Thesis is often used to warrant the claim that (2) is about a large and <u>important</u> class of formal systems. 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.

Chapter 2

PRELIMINARY CAVILS

What would a treatment of Gödel's Second Incompletness Theorem be like that was analagous to my modest sketch of a philosophical account of the First Incompleteness Theorem? First, a statement of what Gödel proved and an up-to-date generalization of it (concerning a large and important class of formal systems); and an argument from that generalization to an analogue of (2). Of major interest to us will be the necessary additional premises. As we shall see, such a project will not work out as neatly as did the account of the First Theorem. Even for the purposes of this modest goal of producing a parallel account, and eschewing direct concern with deep philosophical "implications" of it, we shall find the Second Theorem to be rather more recalcitrant.

An early reference to the Second Theorem is to be found in Gödel's 1931 paper, in which a proof of what he calls Theorem XI is sketched. As a full proof is exceedingly tedious (involving a proof that the first result can be recreated in the formal system), Gödel is content to simply point out that it is clear that such a proof can be given. What does Gödel's proof of Theorem XI show?

Briefly put, an unprovable formula is exhibited, different from 17Gen r, the one exhibited in the proof of his First Incompleteness Theorem. In a footnote to his statement of XI Gödel remarks that the unprovable formula of that theorem is not just any formula that is built in the described way from a predicate that numeralwise expresses is a proof of. 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 unprovable (in P). That is, he shows that there is a proof that the consistency of P implies that a certain formula is unprovable. Corresponding to this proof, there is a derivation in P of a conditional 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 godel sentence is not provable. By the construction of the First Theorem this consequent is the godel sentence itself. Hence the conditional whose antecedent is the consistency sentence and whose consequent is the godel sentence is a theorem of P. And since modus ponens is a rule of inference of P, the formalized statement of consistency cannot be provable if P is consistent.

The above is a rough sketch of the idea of the proof. A very detailed proof would involve constructing (or, at least, showing how to construct) the crucial formal deriva-

tion. The recalcitrance of the Second Theorem, which I mentioned at the beginning of this section, centers on the notion of formalization, which is used in the above sketch. I will explain in what follows how it could be that formalization is responsible for what I have called the recalcitrance of the Second Theorem. However, the semantic flavor of this theorem can be immediately appreciated when we notice that what is unprovable is a formula of P and that this formula is said to say that P is consistent. We shall see that, unlike some of the informal intuitive descriptions of the First Theorem that are often given, this apparently semantic characterization is unavoidable in view of certain purely technical considerations. In fact, the contrast appears locally in the conditional itself; that the consequent says that some formula is unprovable can be replaced by much weaker constraints - namely that the proof predicate used numeralwise express the proof relation. Stronger semantic constraints are necessary for the antecedent. These matters are taken up in detail in the next section.

But before we advert to these more technical considerations, certain features of the Second Theorem can be examined. Gödel gave (or sketched) a proof of (4)[8],

---

[8]For ease of presentation I am continuing to suppress reference to primitive recursive extensions of P; thus 'WID' and not 'WID(K)'.

(4) The formula that translates WID is not provable in P.

which is clear enough, provided that we understand the definite description that occurs in it. [Let Hecuba =df the denotation of this description. Then (4')

(4') Hecuba is not provable in P.

is an even sparser expression of what Gödel proved.] Understanding the description involves understanding the meaning of 'translates' in this context. What should it mean?

As far as (4) goes it seems to (and (4') definitely does) state no advance over the First Theorem. What is the formula that translates WID (who is Hecuba?) and what makes it more special than the unprovable formula constructed for the First Theorem?

WID itself was defined as in (5).

(5) WID = '$\exists x$(Form(x) & $\neg$Bew(x))'[9]

Form and Bew are predicates of numbers, such that, under the given Gödel numbering, Form holds of n just in case n is the godel number of a formula and Bew holds of n just in case

[9]WID is simply the arithmetic statement given. It is well to keep in mind that WID is a remark about numbers, couched in "logician's English."

the formula whose godel number is n is provable. So WID holds just in case there is an unprovable formula of P - a standard characterization of consistency. We may say that WID says that P is consistent[10], provided that Form expresses formulahood and Bew expresses provability.

Even so, WID is certainly not a formula of P, and hence it isn't what's not provable in P, as (4) makes clear.

We say that, for example, 2 + 2 = 4 is provable in P because a certain formula of P is provable and the standard interpretation for P makes the appropriate link between the formula and the manner in which the standard model is described. As we shall see, such a truth-defintional semantics will not work for the Second Theorem. [In this context, i.e., considering whether 2 + 2 = 4 is provable in P, I refer to the "manner in which the standard model is described"; in this case, the standard manner. This is not

---

[10] 'P is consistent' says that P is consistent, while WID makes a remark about numbers. This is not where problems lie. Suppose that, bureaucratically, we were to "identify" each person with his/her social security number. Then relations among people would correspond in a natural way to relations among numbers - and we might even go homophonic on relation names and utter the likes of '122368871 loves 111389411'. This may be perverse but neither confused nor astounding. The homophonic ploy is useful for pointing out why such a harmless isomorphic encoding is useful. If we also went homophonic on numbers - called 122368871 by the name of the person identified with it - it would be hard to imagine a point to the enterprise. The point of such an enterprise involves the utilization of a favored vocabulary and not a favored ontology. The favored vocablulary may introduce entailment relations based on meaning. (Cf. p. 106ff)

the intensional point concerning mathematics that I am eventually getting at; it is simply the point that [Mates], for example, covers on pp. 75-78.

Mates 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

> '$L^2a_1\ a_2$' is true under I, if and only if 2 is less than 3.

and

> '$L^2a_1\ a_2$' is true under I if and only if 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, 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 translation of an English sentence; and this is relative to the way in which the interpretation is given. Mates is thus led to standardize the presentation of certain interpretations, so as to render less indeterminate answers to questions of adequate translation. I say 'less indeterminate' because Mates merely shows the inadequacy of a possible account of <u>adequate translate</u>, not the adequacy of its replacement.

The context we are concerned with (consideration of whether 2 + 2 = 4 is provable in <u>P</u>) is very like the trans-

lation, or regimentation, context that Mates is concerned with. He points out, however, that the truth-value of wffs is independent of the manner of specificatcn of the interpretation (though the meaning is not). The Fregean move that I am aiming for involves locating the intensional context, not in the metalanguage, but in the language being interpreted, thereby yielding a difference in truth-value. So, while P qua arithmetic may have a truth-definitional semantics à la Mates, P qua proof theory does not. Section 10 is devoted to an expansion and explanation of these bracketed remarks.]

What about the formula that translates WID; can it be said to say that P is consistent? Whether it can clearly depends on the notion of translation involved. Let 'A' range over arithmetical sentences of English, 'α' over sentences of P and let the substituends for 'p' be sentences of English. Then (6)

(6) If α translates A then if A says that p, α says that p.

seems a reasonable attempt at an adequacy condition on an account of a formal notion of translation into P; Of course we will want to generalize uniformly to translation-into-the-language-of-T_A, where T_A is any suitable formal system. The project becomes manageable if the range of 'A' is restricted to a small enough class. I will continue to

concentrate on the consistency sentence and related syntactic remarks.

Although none of this concern with translation was involved in our treatment of the First Theorem, the suspicion might arise that the concern with formalization, translation, and "says that" is an artifact of my presentation. Furthermore, there is a notion of translation, viz. numeralwise expressibility, utilized in the proof of the First Theorem; it might be suspected that this nction would serve here.

Numeralwise expressibility is a relatien among certain number theoretic relations, open sentences, and formal systems. The relations is the godel number of a proof in P of and is a formula of P are amongst those "certain" relations. Let Pf(y,x) be an open sentence of P that numeralwise expresses the proof relation and Fm(x) an open sentence of P that numeralwise expresses sentencehood. Then, given the availability of truth functions and quantifiers in P, we construct the sentence of P, $\exists x(Fm(x) \& \neg \exists y Pf(y,x))$, thereby mimicking , in quantificational structure, our definition of consistency (which was that there exists a formula that is not provable.). This sentence is our candidate for a formal consistency sentence. Moreover, our construction will clearly generalize to many formal systems. Let us see if it actually produces a viable candidate.

Taking the foregoing as containing an implicit account of translation, we are led, in the light of (6), to consider (7) as a putative statement of the Second Theorem.

(7) The formula that "says that" P is consistent is not provable in P.

(7) has some problems. More than one (infinitely many) open sentences numeralwise express a given relation[11]. Our construction did not actually produce a unique consistency sentence. Should 'The' in (7) be 'Some' or 'Any'? A desire to obtain something recognizably a version of the Second Theorem would lead to choosing 'Any'. However, another condition on this treatment is that we obtain a true expression of the result - and 'Any' will not give us that on our present notion of translation. For consider (8).

---

[11]The technically oriented will recognize that I am sliding over some buried distinctions here. For a fixed godel numbering and beta function, infinitely many open sentences numeralwise express a given relation. A change in the godel numbering or beta function will in general change the class of sentences representing a given relation - thought of as a syntactic relation. The godel numbering and beta function do not touch the "pure" relation of numeralwise expressibility. As is well-known, given an effective coding into arithmetic relations, all the relevant theorems are preserved. Given the invariance results I will usually suppress reference to the relativity to the godel encoding. All our concerns will be post-coding. (Although cf. p. 108f).

(8) If T is a formal system with property φ, then any formula of T that says that T is consistent is not provable in T. (φ standing in for some explicans of 'being consistent and sufficiently strong'.)

(8) is an analog of (2), but is not forthcoming on our current account of translation, due to our stricture about falsehood. An appropriate place to look for help would be in a proof of the Second Theorem; not the one Gödel gave, but a proof that is detailed and general. Before seeing what notion of translation such a proof would use or presuppose, let us see what makes (8) false on our current account.

Briefly, it is simply that some sentences, constructed as above, are provable. Since we have good reason (as we shall see) for thinking Gödel's Second Incompleteness Theorem true, I conclude that they are not in fact consistency sentences. Fortunately for this project the provable "consistency" sentences can be seen to be defective on grounds other than a desire to make (8) true. Were we not to admit the reasoning to their semantic deviance we would have no non-circular argument for (8). The establishment of (8) is after all a monumental intellectual achievement and not simply self-evident. We cannot rationally hope to argue for (8) on the basis of a firm mathematical theorem and simultaneously support the additional premises needed by the

condition that (8) is true. Our desire and our belief that (8) be true may motivate, but not certify, our search for rational grounds for the belief. I shall eventually exhibit and analyse these grounds - that is, the theories that supply the link from the mathematical result to (8) - and argue that they are semantic theories of a certain sort.

Those with philosophical scruples that mitigate against the non-extensionalist cast present in such theories are left without (8) (and without a host of interesting foundational programs). Historically, this fact is of interest insofar as referential semantics for mathematics has been held up to natural language semanticists as something they would do well to imitate. I will show that this is a bad model even for that portion of natural language that is mathematical. My strategy will be to uncover in the technical literature concerning the Second Theorem, rigorous theories of proof predicates and consistency sentences. I will show that these theories will support (8), if interpreted as semantic theories for the language of proof theory. These theories are then my candidates for connective theses appropriate to Gödel's Second Theorem. I shall argue for their correctness as semantic theories, and, finally, show them to be intensional semantic theories.

This latter fact can be (psychologically) obscured by the incestuous nature of the theories involved. In semantic theories the objects that are assigned to syntactic entities

can be of many kinds; tables, chairs, numbers, people, relations among tables, chairs, numbers, people, sets of tables, chairs, numbers, people, sets of sets of tables, chairs, numbers, people, etc. Thus we often distinguish syntactic from semantic accounts, not formally, but according to the nature of the entities involved. If non-linguistic entities are involved, then we have a semantics. If only linguistic (i.e. syntactic) entities, then syntax. This latter conditional is false. The counter-example is just the case I shall be concerned with - the objects of the interpretation, the objects assigned by the semantics to syntactic entities, are themselves syntactic entities[12]. These matters are pursued in detail in secticns 6 and 10.

In the next section I expose the defects of the suspect sentences and counter the suspicion mentioned above - that the difficulties are an artifact of my presentation.

---

[12]Clearly I think that intensionality is a linguistic matter and that talk of intensicnal objects, when coherent, is implicitly parisitic on a linguistic noticn. My thesis, in the intensional object argot, is that formalisms are intensional objects.

Chapter 3

FURTHER CAVILS

Gödel's First Incompleteness Theorem predicates a simple syntactic property of a large class of formal systems. This property, incompleteness, is simple in at least the the following respect - it is definable in terms that do not invoke anything akin to a translation relation: T is incomplete if and only if there is a sentence of T such that neither it nor its formal denial is provable in T.

Gödel's Second Incompleteness Theorem predicates a certain more complex property of a large class of formal systems. Call this property Z. T has Z if and only if there is a sentence of T, z, such that z is not provable in T and z says that T is consistent. The Second Theorem says that all sufficiently strong consistent T have Z. The last clause in Z's characterization is not well defined. Note first that the variable 'T' occurs after 'says that' in the characterization of Z. As we shall see, Z shares certain features with other problematic properties. A purported open sentence that purports to stand for Z may not pick out a property; in the sense that it has been argued that 'I remember x', 'George is thinking about x', 'Edna believes that x is bald', 'x is so-called because of its size',etc., do not - and hence aren't even open sentences.

Someone might alertly observe, however, that although the First Theorem, as I stated it, doesn't involve a formal system making a remark about itself, every schoolchild knows that the proof of the First Theorem produces a witness to the formal system's incompleteness. This witness is a formula such that neither it nor its formal denial is provable; and, the story goes on, it achieves such witness status by being a formula that does make a syntactic remark - namely that it, itself, is not provable. So, the argument goes, either the First and Second Theorems have the same troubles about formulas expressing propositions or neither does; and my remarks of the last section and the preceding paragraph are red herrings.

Although this strawperson argument is invalid, it is instructive. It points up that the difference between the First and Second Theorems is the difference between the syntactic form of the open sentences that purport to pick out the properties that those theorems attribute to formal systems. In the present case, what this comes to is that it is no necessary part of the proof of the First Theorem that the formal sentence say of itself that it is unprovable. That it may seem to be one is an artifact of certain informal, motivating, semantic accounts of the First Theorem. Certain <u>entailments</u> of the proposition that the godel sentence says of itself that it is unprovable are used and these are sufficiently captured by the relation of numeralwise

expressibility[13]. This is clear from an examination of the proof, or, more revealingly, by consideration of the Rosser improvement - where, as we shall shortly see, the undecidable formula does <u>not</u> say that it is not provable. On the other hand, we shall see that it is a necessary part of a proof of the Second Theorem, and not a mere artifact of my preceding informal account, that the consistency sentence for T "say" that T is consistent. That is, cavils aside, the technical considerations <u>insure</u> that this is the case.

To bring these general considerations down to cases (finally) requires some technical machinery. Numeralwise expressibility, hereafter n.e., is a three-place relation among relations or properties, formulas, and formal systems.

(9) If R is an arithmetic relation or property[14], r a formula of T, T a formal system, then r <u>numeralwise expresses</u> R <u>in</u> T if, for any n-tuple of numbers, $\langle m_1, \ldots, m_n \rangle$

i. $R\langle m_1, \ldots, m_n \rangle \rightarrow \vdash_T r\langle \bar{m}_1, \ldots, \bar{m}_n \rangle$

ii. not $R\langle m_1, \ldots, m_n \rangle \rightarrow \vdash_T \neg r\langle \bar{m}_1, \ldots, \bar{m}_n \rangle$

---

[13] I have the impression that [Mostowski] deserves credit for first clearly sorting out the welter of syntactic and semantic theorems. The thesis I am presenting is ultimately to the effect that, contra the spirit of Mostowski's hasty appendix, the Second Theorem is intrinsically semantical.

[14] By 'arithmetic' is meant of or among numbers. The function denoted by '$\neg$' has to be effectively given, and T is presupposed to have denial explicitly available. $\bar{n}$ is the standard numeral of T for n.

(9) is the definition. It is just the recursive relations that are n.e. in P.

The culmination of Gödel's famous series of definitions is the arithmetic relation which we will write 'yBx'. yBx just in case y is the godel number of a formal proof in P whose last line has godel number x[15].

Since by Church's Thesis yBx is recursive, there is a formula of P that numeralwise expresses yBx. We have already called (some such) it 'Pf(y,x)'. So, whenever nBm (which I'll read 'n is a proof of m'), the result of substituting the numeral of P for n and m into Pf is provable in P. And if not nBm then $\vdash_T \neg Pf(n,m)$.

A crucial arithmetic function, also recursive, is the famous Gödel diagonal substitution function. S(x) is the godel number of the formula that results from substituting the numeral for x into the formula with godel number x. Let Pd(y,x) be any formula that n.e. yBS(x). Let q be the godel number of $\forall y \neg Pd(y,x)$. Then $\forall y \neg Pd(y,q)$ is a Godel sentence. The plausibility of regarding this sentence as saying that it is not provable arises from considering the standard interpretation and taking Pf as expressing the proof relation. To what extent is this merely a pun?

---

[15]Gödel didn't define yBx this way; it was built up from simpler relations by certain constructions, in order to achieve the formal result. Our imminent invocation of weak Church's Thesis is but a dispensible short-cut. We shall see that for the Second Theorem the construction is more relevant.

The meta-theorems that warrant the inferences needed in the First Theorem that intuitively would flow from a sentence that said that it is not provable are forthcoming just on the hypothesis of numeralwise expressibility. (These meta-theorems are of the form: If $\vec{n}$ are the godel numbers of syntactic objects standing in relation R, then $r(\vec{n})$ is a theorem of T, and conversely. One is thus enabled to argue from syntactic fact to facts of derivability and conversely.)

The formal argument given by Gödel in the proof of the First Incompleteness Theorem does not parallel the intuitive one that occurs from contemplating 'I am not provable'. The intuitive argument utilizes the notion of truth. The proof of the First Theorem avoids this by "pulling the metalanguage down into the object language"[16]- but only a constructive fragment. It does not have to (nor could it) pull truth down.

To put it less aphoristically, Gödel showed that recursive syntactic relations could be represented in formal arithmetic in the aforementioned weak, numeralwise bi-conditional way. Tarski's Theorem is to the effect that satisfaction (and hence truth) are not such relations. Thus some talk of proof could be reproduced in formal arithmetic -

---

[16]I first heard this phrase from Burton Dreben, from whom I also first heard of provable "consistency" sentences.

enough so as to yield an undecidable sentence. Talk of truth, insofar as it would require a truth predicate, is impossible for formal arithmetic.

Let us consider the thesis that if r numeralwise expresses R then then $r(\vec{n})$ says that $R(\vec{n})$. Call this the n.e. thesis. Were even this true, in order for either the consistency sentence or the godel sentence to be said to express anything, the thesis would have to be extended in the following manner. Stipulate that the truth-functional connectives and the quantifiers have their usual meaning. I take it that the usual meaning of the quantifiers involves a specification of the domain of quantification. The formal system, uninterpreted, is not capable of such specification - in fact this is a consequence of the First Theorem, in the following sense. The First Theorem shows that there must exist non-standard models of formal arithmetic - models non-isomorphic to the standard model. In this sense the formal system does not rule out non-numbers from the domain of quantification and so doesn't "fix" the domain. The relevant intuition concerning the godel sentence is that it is false in some non-standard models - there are non-standard (godel numbers of) proofs of it. So the godel sentence, given the n.e. thesis, says that it is not provable only relative to the standard interpretation (given in the standard manner, cf. p 19f).

A thesis fully warranting a "says that"-claim for the godel sentence will represent an ideological increase even over our expanded n.e. thesis. Although there are seeming (first-order extensional) semantic notions used there, they are all replaceable by explicit constructive definitions. Thus, for example, standard numeral for n needs no semantic apparatus for its definition - '0' preceded by n 'f''s (or whatever). Indeed, as indicated above, even is a formal proof of can be so dealt with. As all the needed notions are recursive, they are in fact definable in P.

In any case our project is more modest. It is to give an account of a representation or translation relation sufficient to justify (8). In doing this we certainly do not seek to expand the class of relations "represented" but to strengthen the representation relation, to carve up the class more finely. However, only a small subset of the recursive relations and statements are at issue. For what we have seen so far is that the First Theorem does not require a real translation relation; that if we want the godel sentence to say what is often claimed it says a little first-order extensional semantics seemed plausible; and that in contrast, (8) explicitly refers to a sentential formula stating that T is consistent. If the Second Theorem is in fact to be straightforwardly about consistency, then explicating 'a sentential formula of T stating that T is consistent', for variable T, is of some import. And while a par-

ticular proof of an instance, i.e. about a particular formal system, may exhibit a particular formula that is unprovable, it even then has to be justified as being a formula that says that that system is consistent.

The use of the standard model, though it explains why we say the godel sentence and its ilk say what we say they say, and is not an ad hoc technical fix, is simply not satisfactory. We want an explication of when a sentence of a formal system, T, expresses a particular proposition - that T is consistent. The expanded n.e. account is methodologically adequate but doesn't cover the cases. That is, there is nothing incoherent in this account; it doesn't violate any methodological canons. It is just empirically inadequate - it makes the wrong predictions. In particular, as we shall see, it doesn't distinguish sentences that are consistency sentences from some that aren't.

Before coming to the bare-boned technical data, it should be noted that we can envision, in advance, one kind of solution. If we rule out the use of an explicit semantic apparatus, the remaining means of describing the expressive abilities of formal systems is in terms of syntactic structure and of what formalisms can prove. Numeralwise expressibility is an instance of such a description. So we might expect further conditions to be of this sort; i.e. conditions on derivability.

Chapter 4

DEVIANCE

Let us suppose that we have fixed on a few standard means of building formal consistency sentences from a formal proof relation. (This reduction of the problem can be justified.) Given a formal proof relation we will then have a stock of consistency sentences made up of the standard transcriptions and their logical equivalents. What is required of the proof relation that goes into the consistency sentences?

In what preceded I gave reasons for believing that there is no good reason to believe that numeralwise expressibility will pick the appropriate proof predicate. Even if numeralwise expressibility was sufficiently strong to yield a coherent technical result, this would hardly establish (8). The technical solutions to the problem of generalizing the Second Theorem, which suffice to prove the technical result, need to be examined as to why they are solutions of the right sort; i.e. why they support (8).

For the First Theorem there was no requirement that the 'express' in 'numeralwise express' be anything but a pun. For the Second Theorem I argued that the seeming logical form of (8) required that 'expressibility' be taken more literally.

If (8) is false we are, of course, done. In this section I content myself with showing that requiring only numeralwise expressiblity of the proof relation entails that (8) is false. Later, in section 10, I shall reveal that there are coherent accounts that certify (8).

Rosser exploited the fact that the only property of the formal provability relation needed in the First Theorem is that it numeralwise express $\underline{yBx}$. By constructing a new open sentence that n.e. $\underline{yBx}$ but had special properties as well, Rosser was able to improve the Godel result. Let Pf be a formal proof relation as Godel would define it; then define Pf' as in (10).

$$(10)\quad Pf'(y,x) \equiv Pf(y,x) \;\&\; (\forall z_1 \leq y)(\forall z_2 \leq y)(\forall w \leq y) \neg[\, Pf(z_1, w) \;\&\; Pf(z_2, ng(w))\,]$$

Pf' like Pf n.e. $\underline{yBx}$. $(nBm \rightarrow \vdash Pf(\bar{n},\bar{m}))$. Suppose the formalism consistent. Then nothing is a proof of a sentence and its denial. $\neg nBm \rightarrow \vdash \neg Pf(\bar{n},\bar{m}) \rightarrow \vdash \neg Pf'(\bar{n},\bar{m})$.

However, (19)

$$(19)\quad \forall x\, \forall y_1\, \forall y_2 \neg[\, \varphi(y_1, x) \;\&\; \varphi(y_2, ng(x))\,]$$

becomes an unprovable consistency sentence when φ is replaced by Pf; but when φ is replaced by Pf' the result is provable. In fact it is trivially provable - consistency is built in.

A similar result is more simply achieved using the Rosser form (20).

(20) $Pf(y,x) \;\&\; \forall z \leq y\ (\neg Pf(z, ng(x)))$

An even simpler deviant expression is yielded by (21);

(21) $Pf(y,x) \;\&\; \neg Pf(y,k)$

where k is the godel number of '0=1'. If T is consistent $Pf(y,x) \;\&\; \neg Pf(y,\bar{k})$ n.e. what $Pf(y,x)$ does. Consider the consistency schema (22).

(22) $\neg\, \exists y\ \varphi(y,\bar{k})$

Then, although with non-deviant Pf in for φ (22) becomes not provable, (23)

(23) $\neg\, \exists y\ (Pf(y,\bar{k}) \;\&\; \neg\, Pf(y,\bar{k}))$

is.

What isn't provable in T is that (21) and Pf do numeral-wise express the same relation; i.e. not $\vdash_T Pf(y,x) \longleftrightarrow [Pf(y,x) \;\&\; \neg Pf(y,\bar{k})]$.

In each of these cases it is intuitively clear that the deviant predicates are in fact deviant. Thus arises the aforemen-

tioned empirical inadequacy of the expanded n.e. thesis. However, not all nonstandard consistency sentences are of this type. Consider the following line of argument adapted from Kleene.

The Gödel sentence itself is a consistency sentence. Why? Any sentence that says a formula is unprovable is a consistency sentence, and the Gödel sentence says that it is not provable. Or, since its existential generalization is a consistency sentence, then it at least entails a consistency sentence; and vice versa. And, provably so - $\vdash_P CON \leftrightarrow G$[17].

I don't think that $\vdash CON \leftrightarrow G$ is sufficient to show G a consistency sentence. After all, not all theorems are synonymous. The rest of Kleene's argument is worth unpacking. No consistency sentence of a formal system literally transcribes the form of the English sentence 'P is consistent' - no consistency sentence is of the form Fa. They do transcribe any one of a number of definitional expansions of such sentences, with the reference to P, or whatever formal system, being implicit in the proof predicate used. If a

---

[17]On p. 211 of Introduction to Metamathematics Kleene says: "Intuitively [the Gödel-Rosser undecidable formula] itself expresses an equivalent [to consistency], via the long intuitive proof of Gödel's theorem. For by [G1] if the system is consistent, the G-R formula is unprovable and... if G-R is unprovable, the system is consistent." That G-R is the appropriate arithmetization of G-R is unprovable plays a crucial role here - hence the reference to the "long intuitive proof". Not any unprovable formula will do.

general theory of proof predicates (i.e. an account of proof predicate for T, for variable T) is available then one can have an explicit variable over formal systems. The various syntactic definitions of consistency have a common character. Not only are they all trivially equivalent (provably so in any system we will be interested in), but they all assert the unprovability of something. Some assert the unprovability of $\perp$, some of '0=1', some simply say that there is a sentence that is unprovable. Any sentence that says that some unprovable formula is unprovable is a consistency sentence. The Gödel sentence, G, is of this form. If one replaces the term for G in G by the godel numeral for $\perp$ the result is a more usual (and equally unprovable consistency sentence)[18].

This observation that G has some of the necessary properties of consistency sentences depends on the form of G,

---

[18]There are some issues of propositional identity that arise here. It seems plausible to say that 'Edna is not here' and 'Someone is not here' both say that someone is absent. And 'George is not here' is equally an "absence" sentence. It would, however, be misleading, with respect to natural language, to say that all these sentences express the absence proposition. Of course for natural languages we want our semantic theory to respect all semantic facts. For the case in hand we will deliberately be working in an impoverished semantic theory. Moreover, the following fact about formalisms is relevant: All the consistency definitions examined so far are provably equivalent. (For the ones that assert the unprovability of a particular sentence, the inference from a pure consistency sentence, i.e., one with no terms referring to a particular sentence, to them requires the particular sentence to have certain properties.)

and does not proceed from the observation that there is an unprovable formula, namely G. No mere example of unprovability, say 0=1, is a consistency sentence. The proof of the First Theorem does not establish G as a consistency sentence. First, as we have seen, it does not use the fact (when it is a fact) that G says "I am not provable." Second, the Rosser improvement highlights this by producing an undecidable sentence that does not express its own unprovability - and whose corresponding consistency sentence is provable.

So, a non-deviantly constructed G is not a deviant consistency sentence. There is, though, another kind of deviant consistency sentence. It is clear from the proof of the First Theorem that each instance of G is provable. That is, for every n $\vdash_P \neg Pd(\bar{n}, \bar{q})$. It is one lesson of the First Theorem that quantifiers cannot always pass through turnstiles. One can, however, formulate the proposition that every substitution instance (using standard numerals) of $\neg Pd(y, \bar{q})$ is provable. Let p be the Godel number of $\neg Pd(y, \bar{q})$; then (24)

(24) $\vdash_P \forall y$ Bew Sub $(y, \bar{2}, \bar{p})$

holds. (24) should not be confused with the remark that one can prove that every instance of G holds. One can, but <u>P</u> can't.

In [Webb], Webb asserts that a substitutional version of consistency "is a nice provable expression of consistency after all." Let $k = \ulcorner 0 = 1 \urcorner$ and let $Pf(x,y)$ be normal, and $A(x)$ abbreviate $\neg P(x,\bar{k})$, then (à la [Webb] p. 30) we get

$$(25) \qquad \vdash_P \forall n \; Bew(S(\ulcorner \bar{A}(x) \urcorner, \ulcorner \bar{n} \urcorner))$$

where S is the substitution function term. (In fairness, Webb doesn't say that the substitution version "says" P is consistent - he says its assertive force is less than that of CON(P).)

Do (24) or (25) exhibit a provable consistency sentence[19]? I argued above that if a sentence said that some formula was unprovable it was a consistency sentence. Conversely, every consistency sentence must say that. The sentences in (24) and (25) neither say nor imply that. They say that lots of things are provable; in point of necessary fact all those sentences being provable cannot be used to

---

[19]The following intuition may help in seeing that (24) is true. $Sub(i, j, k)$ is the (godel number of the) formula with $\bar{i}$ replacing x_j in k. $Sub(x_i, x_j, x_k)$ is a term for $Sub(i, j, k)$. So $Sub(y, \bar{2}, \bar{p})$ denotes the result of substituting the standard numeral for y in p for 'y'. The fact of the use of standard numerals lops off, intuitively speaking, the non-standard interpret- ation of the quantifier $\forall y$. That no particular standard proof is a proof of q is provable. Similarly for (25), the "numeralization" of the quantification over proofs omits non-standard "proofs". P cannot prove that its proofs are none of them non-standard.

prove consistency. They simply don't have the right form. That neither says that something is not provable is patent - starting as they do with a universal quantifier. That they don't entail that something is not provable follows from the facts that they are provable, that modus ponens is a rule of P, and the Second Theorem.

Chapter 5

FEFERMAN

These deviant proof predicates and consistency sentences indicate that what is needed is a precise definition that would pick out the (or only) formulas that correctly express _is a proof of_. Viewed as a technical problem, that of producing a generalization of the Second Theorem, this need has not been ignored - and Kreisel has emphasized its foundational aspects. As a technical problem the earliest treatment occurs in [H-B] where three derivability conditions are enumerated and used to prove a rigorous version of the Second Theorem. Any proof predicate that satisfies the derivability conditions will suffice for the Second Theorem[20]; the labor is in showing that any particular proof predicate does satisfy them. This approach is continued in the work of Löb ([Löb]) and in important recent work in modal systems (where $\Box$ is interpreted as _provable in a formal system_).

Between 1939 and 1953 very little happened on the technical front. The Hilbert-Bernays derivability conditions

---

[20]In light of my previous remarks it should be emphasized that the Second Theorem here is not justifiedly about consistency - not until the derivability conditions are semantically justified.

were conditions on the formal proof relation, in addition to numeralwise expressibility, that were sufficient to guarantee an unequivocal Second Theorem in the following sense: Any proof relation meeting the three conditions would satisfy the requirements of the proof of the Second Theorem and, moreover, the proof relations constructed for familar theories were seen to satisfy the conditions.

Feferman's 1957 dissertation and 1960 paper ([Fef]) and Jeroslow's papers ([Jer1], [Jer2], [Jer3]) were stimulated by remarks of Kreisel ([Kreisel]). Feferman points out that "numerically correct" proof definitions are inadequate for certain results. Results for which they are adequate he calls 'extensional', the rest 'intensional'. The deviant proof predicates are intensionally incorrect. Some intensionally incorrect predicates lead to useful extensional results (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. Despite the value of his technical work, this is all Feferman has to say about its philosophical import.

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 tran-

scription of any one of the equivalent definitions of consistency); the proof predicate in turn is straight-forwardly transcribed from the presentation of the formal system. The trick is obtaining a formal object to represent the presentation (no set theory allowed).

More precisely: If t(x) is a formula that n.e. the axioms of T, a proof predicate can be constructed "in a standard way" from t. Does the phrase 'a consistency sentence for a formal system' make sense? Only if one individuates formal systems more narrowly than by their axioms (and certainly more narrowly than by their theorem sets) - viz. by the formulas t(x). For the equivalence classes of such t's obtained by the same extension as relation are incorrect for the technical contexts in which they occur[21]. By narrowing the individuation it is then possible to generalize various metamathematical results by conditions on the formulas t.

Feferman defines a proof relation given a formula t that numerically defines the set of axioms. Since many t's numerically define the same set of axioms, for the same axioms different formal proof relations will be defined; one

---

[21] For, utilizing same extension as is equivalent to utilizing numeralwise expressibility as sufficient to characterize a formal proof predicate. We saw in section 4. that the existence of deviant, but extensionally correct, formal proof predicates would make generalizing about formalisms difficult in the absence of finer discriminations.

for each t. So either there is no such thing as the formal proof relation for a theory or a theory is not its set of theorems or even its set of axioms plus fixed deductive apparatus[22]. The proof relation has to be "taken in intension." If we talk according to the first alternative we can say that the identity of the proof relation is sensitive to the mode of presentation of the theory. Deviant proof relations are bizarre ways of giving the axioms; bizarre enough so as to carry a trivial assurance of consistency. ("Tell me how old you are and make no mistake." "I'm as old as I am." The situation is also reminiscent of Frege's problem with the meaning of proper names: If a proper name means what a particular definite description means, then certain non-analytic propositions end up as analytic. See p. 73.)

The formal proof relation is obtained from a t by mimicking the construction of the real proof relation, generalized over uniformly presented formalisms. The idea is to reflect the notion of logical derivability from axioms by copying its usual definition step by step in formalized syntax. The syntax of a formal system is developed in number theory, relativized to an unspecified non-logical vocabulary

---

[22]A third alternative: There is a unique special t for a theory, which can be used to define the formal system. This turns out to be plausible only for certain theories, for certain purposes. In particular, a finitely axiomatized theory has a unique (up to λ-equivalence) preferrable t. See p. 49.

K. Thus 'FmK' denotes the formulas of K, 'TmK' the terms, 'StK' the set of sentences in K. The usual syntactic notions, such as substitution, are characterized as number theoretic functions in the usual way - yielding, for example, '$Sb^{a}_{\beta}\ \gamma$', which denotes the substitution of term β for a in γ. Conventional abbreviations are introduced to restore standard notation; '$\gamma(\beta_0,\ldots,\beta_n)$' for '$Sb\binom{v_0 \cdots v_n}{\beta_0 \cdots \beta_n}$'. Finally, as the logical base a primitive recursive set AxK is added. A system or axiom system is a pair $\langle A,K\rangle$, where K is a subset of Const (the set of non-logical constants) and $A \subseteq StK$. $\underline{A} = \langle A,K\rangle$. $L_K = \langle O,K\rangle$. Let $A/n = \{\varphi \in A: \varphi \leq n\}$ and $\underline{A}/n = \langle A/n,K\rangle$. The proof relation $Prf_{\underline{A}}$ is the relation such that "for any φ, ψ, $Prf_{\underline{A}}[\varphi,\psi]$ iff $\psi \in Sq$ and $\varphi = (\psi)_{L(\psi)-1}$ and for each $i < L(\psi)$, $(\psi)_i \in FmK$ and either

i) $(\psi)_i \in AxK$

ii) $(\psi)_i \in A$

iii) for some $j,k < i$ $(\psi)_k = (\psi)_j \rightarrow (\psi)_i$.

...$Pr_{\underline{A}} = \{\varphi: \exists\psi Prf_{\underline{A}}[\varphi,\psi]\}$. $\vdash_{\underline{A}}\varphi \equiv \varphi \in Pr_{\underline{A}}$."

With this equipment the usual metamathematical theorems are forthcoming. Feferman's (2.2) is the Deduction Theorem and (2.3) is the following finite deducibility result:

(F2.3) For any $\varphi \in FmK$, $\varphi \in Pr\underline{A}$ iff $\exists n\varphi \in Pr\underline{A}/n$

Furthermore, $A \in CL$ iff $Pr_{\underline{A}} \subseteq A$ and $A \in CON$ iff for $\varphi \in FmK$, not $\vdash_{\underline{A}}\varphi$ or not $\vdash_{\underline{A}}\neg\varphi$. An easy result is (2.6): $\underline{A} \in CON$ iff $\forall n\ \underline{A}/n \in CON$.

Formal arithmetic is introduced via K_0 whose symbols we write to remind us that we will use them for arithmetic; e.g. $\delta + \epsilon = f_{1,0}[\delta, \epsilon]$. Numerals are introduced by: $\bar{0} = c_0$, $(\overline{n+1}) = (\bar{n})'$. Q and $\underline{P}$ are singled out as interesting K_0-theories. Q is the well-known Q, finitely axiomatized, and $\underline{P}$ is $\underline{P}$, so-called because of its name. Feferman's (3.1) establishes the adequacy of Q for the intended interpretation of +, ., =, etc.

(F3.1) (i) $\vdash_Q \bar{n} + \bar{m} = \overline{n + m}$ and $\vdash_Q \bar{n} \cdot \bar{m} = \overline{n \cdot m}$

(ii) $\vdash_Q \bar{n} \neq \bar{m}$ if $n \neq m$

(iii) $\vdash_Q x \leq \bar{n} \leftrightarrow x = \bar{0} \vee x = \bar{1} \vee \ldots x = \bar{n}$

(iv) $\vdash_Q x \leq \bar{n} \vee \bar{n} \leq x$

Feferman's term for numeralwise expressibilty is 'bi-numerate'. A p.r. extension of $\underline{P}$, $\langle P', K \rangle$ essentially consists of additonal function symbols - to go on the left side of primitive recursive defining equations (extending K_0 to K) - the defining equations (extending P to P'), and the extension of the induction schema to cover FmK. The desired results are forthcoming as (F3.4): For each p.r. function there is a p.r. extension of $\underline{P}$ with a term that numerates it in $\underline{P}'$ and every such formula numerates a p.r. function. For each p.r. relation there is a p.r. extension with a term that binumerates it, etc. By elimination techniques (bi-)numeration in $\underline{P}'$ can be replaced by (bi-)numeration in $\underline{P}$.

Two primitive recursive classes of formulas of $\underline{P}$ are defined. The superscript indicates the elimination mapping.

(F3.6) Df. $\varphi \in FmK$

i) φ is PR-formula if $\exists \underline{P}'$, $\underline{P}'$ a p.r. extension of $\underline{P}$, and a term t of $\underline{P}'$, such that $\varphi = (t = 0)^{(\underline{P}')}$.

ii) φ is an RE-formula if for some PR-formula ψ, $\varphi = \exists \ldots \exists \psi$.

As one might suspect these classes are closed under disjunction, conjunction and bounded quantification (and denial in the PR case). If φ_1, φ_2 are PR-(RE-)formulas we can effectively find PR-(RE-)formulas φ and ψ such that $\vdash_{\underline{P}} \varphi \leftrightarrow \varphi_1 \vee \varphi_2$ and $\vdash_{\underline{P}} \psi \leftrightarrow \varphi_1 \wedge \varphi_2$. These results culminate in (F3.11)

(F3.11) (i) If φ a PR-formula and $Fv(\varphi) = \{v_0, \ldots, v_n\}$, then φ is a bi-numeration in Q of an (n+1)-ary primitive recursive relation R; further to each such R corresponds a PR-formula φ which bi-numerates it in Q. (ii) If φ is an RE-formula and $Fv(\varphi) = \{v_0, \ldots, v_n\}$, then φ is a numeration in Q of an (n+1)-ary recursively enumerable relation R; further, to each such R corresponds an RE-formula φ, which numerates it in Q. (iii) ...(i), (ii) apply to any recursively axiomatizable consistent extension of Q.

Feferman then specifies a particular p.r. extension M of $\underline{P}$. The metatheorems justify the following notational device. With certain p.r. functions there is associated a function symbol of M; denote this function symbol by the ordinary non-uniform, mathematical notation for the p.r. function, with a dot added. Mutatis mutandis for relations. Thus '$\mu \cdot^{\omega}$' denotes a term of M that represents exponentiation. Arithmetized versions of metatheorems can be written by taking the metatheorems, splattering them with ink and placing a turnstile in front.

The only difficulty is with propositions involving provability. AxK, the set of logical axioms, was explicitly given; but in the definition of $\mathrm{Prf}_{\underline{A}}$ membership in A is mentioned. Feferman's crucial method for dealing with this is contained in defintion (F4.1), which defines a formula Prf_{α}. Here 'α' denotes not an axiom system A, but a formula in one variable, x. Feferman asserts that if $\alpha(x)$ expresses that x belongs to A then $\mathrm{Prf}_{\alpha}(x,y)$ will express that y is a proof from $\langle A,K\rangle$ of x. Interpretation aside, (F4.1) is the dotted version of the definition of $\mathrm{Prf}_{\underline{A}}$, with α standing in for A[23].

---

[23] (4.1) α a formula of M, u,v,w not free in α and distinct from x, y, z. $\mathrm{Prf}_{\alpha} = (\mathrm{Sq}(y) \wedge \mathrm{Lh}(y) \neq 0 \wedge \forall u\{u<\mathrm{Lh}(y) \rightarrow \mathrm{Fml}_K((y)_u) \wedge [\mathrm{Ax}_K((y)_u) \vee \alpha((y)_u) \vee \exists v \exists w(v<u \wedge w<u \wedge (y)_u = (y)_w \rightarrow (y)_u)]\} \wedge x = (y)_{\mathrm{Lh}(y) \dot{-} 1})^{(M)}$

$\dot{Prf}_\alpha$ is not just any formula that bi-numerates $Prf_{\underline{A}}$. Two things insure this. The dot notation is defined in terms of an effective procedure yielding a formula (this is the purpose of using the well-behaved PR- and RE- formulas) and $\dot{Prf}_\alpha$ depends on α. If $Fv(\alpha) = \{x\}$ and $\alpha' = \alpha(x) \wedge x \leq z$, then $\dot{Prf}_{\alpha'}$ is denoted by '$\dot{Prf}_{\alpha|z}$'. In the second definition numbered '(4.2)' Feferman defines a particular formula α to go with A, when A is finite ($A = \{k_1, \ldots, k_n\}$): $[A] \equiv x = k_1 \vee x = k_2 \vee \ldots x = k_n$. (cf. note on p. 44). (F4.3) gives the expressions for provability. That these constructions are "extensionally" correct is the content of (F4.4).

(F4.4) Let $\alpha \in FmK$, $Fv(\alpha) = \{x\}$. Let $\underline{A} = \langle A, K\rangle$ be ... arbitrary ... system and $\underline{S} = \langle S, K'\rangle$ a theory with $Q \subseteq S$.

(i) If α numerates (bi-numerates) A in $\underline{S}$ then $\dot{Prf}_\alpha$ numerates (bi-numerates) $Prf_{\underline{A}}$ in $\underline{S}$.

(ii) If α numerates A in $\underline{S}$ and $\underline{S}$ ω-consistent, then $\dot{Pr}_\alpha$ numerates $Pr_{\underline{A}}$ in $\underline{S}$.

As with any arithmetization of a relation, one is interested in $\dot{Prf}_\alpha$ and $\dot{Pr}_\alpha$'s adequacy. We already know that $\dot{Pr}_\alpha$ is not going to behave as pleasantly as $x \dot{+} y$, but certain theorems, reflecting elementary truths about provability in A, are forthcoming with α merely occuring schematically. Thus,

(F4.6) Let α be a formula of $\underline{M}$, $Fv(\alpha) \subseteq \{x,z\}$

i) $\vdash_{\underline{M}} \dot{Pr}_{\alpha}(x) \rightarrow \dot{Fm}K(x)$

ii) $\vdash_{\underline{M}} \dot{A}x\dot{K}(x) \rightarrow \dot{Pr}_{\alpha}(x)$

iii) $\vdash_{\underline{M}} \alpha(x) \wedge \dot{Fm}K(x) \rightarrow \dot{Pr}_{\alpha}(x)$

iv) $\vdash_{\underline{M}} \dot{Pr}_{\alpha}(x) \wedge \dot{Pr}_{\alpha}(x \dot{\rightarrow} y) \rightarrow \dot{Pr}_{\alpha}(y)$

v) If ψ is any formula of M,

$\vdash_{\underline{M}} \forall x[(\dot{A}x\dot{K}(x) \rightarrow \psi(x)) \wedge$

$(\alpha(x) \wedge \dot{Fm}K(x) \rightarrow \psi(x))] \wedge$

$\forall x \forall y[\dot{Fm}K(x) \wedge \dot{Fm}K(y) \wedge \psi(x) \wedge \psi(x \dot{\rightarrow} y) \rightarrow \psi(y)] \rightarrow$

$\forall y(\dot{Pr}_{\alpha}(x) \rightarrow \psi(x))$

These are elementary consequences of the fact that M can "follow" an inductive definition, so that i)-v) are verifications that the inductive definition was captured. Feferman points out that the first and second conditions of Hilbert-Bernays follow from (F4.4) and (F4.6) and are thus independent of the choice of α.

Feferman then proves a batch of theorems to the effect that various theorems have provable arithmetized versions. These are still dependent only on the inductive nature of the definition of $\dot{Prf}_{\alpha}$ and its extensional correctness. The theorems are proved by "following" an explicit constructive proof of the original theorem. Feferman's dot notation makes this procedure moderately easy to comprehend. In particular the following is proved:

(F4.10) i) For any particular $\varphi \in FmK$ $\vdash_{\underline{M}} Con_{\alpha} \longleftrightarrow \neg Pr_{\alpha}(\overline{\varphi} \wedge \neg \overline{\varphi})$

ii) $\vdash_{\underline{M}} Con_{\alpha} \longleftrightarrow \forall z Con_{\alpha|z}$

iii) $\vdash_{\underline{M}} \forall x(\beta(x) \wedge FmK(x) \rightarrow \alpha(x)) \rightarrow (Con_{\alpha} \rightarrow Con_{\beta})$

A diagonal lemma is proved, and for each $\alpha \in FmK_0$, γ_{α} is the sentence constructed à la the lemma such that $\vdash_Q \gamma_{\alpha} \longleftrightarrow \neg Pr_{\alpha}(\gamma_{\alpha})$. The underivability of γ_{α} is shown for arbitrary α numerating arbitrary extensions of Q, provided the extensions are recursively enumerable. Restrictions on the form of α are needed for the Second Thecrem. For this purpose, two important facts are supplied by Th. 5.4 and Cor 5.5. Theorem 5.4 is the formalized version of (F3.10), which says that any true Bounded Prenex Formula is provable in Q. (F5.4) says that this itself is provable in M. Cor. 5.5 says that if φ is Q-provably equivalent to some $\psi \in$ BPF, then $\vdash_{\underline{P}} \varphi \rightarrow Pr_{[Q]}(\overline{\varphi})$. The stronger condition, $\vdash_{\underline{M}} \varphi \rightarrow Pr_{\alpha}(\overline{\varphi})$, is obtainable from $\vdash_{\underline{M}} [Q](x) \rightarrow Pr_{\alpha}(x)$, by (F4.7) i). (Cf. p. 286 [H-B] v. II).

This is enough equipment to prove Theorem 5.6, a version of Gödel's Second Incompleteness Theorem; and represents in Feferman's context the all-important third derivability condition of [H-B].

(F5.6) Let $\underline{A} = \langle A, K \rangle$ be a consistent axiom system with $\underline{P} \subseteq \underline{A}$. Suppose α is an RE-formula which numerates A in S, where $Q \subseteq \underline{S} \subseteq \underline{A}$. Then

$$\vdash_{\underline{A}} \dot{Con}_{\alpha} \leftrightarrow \gamma_{\alpha} \quad \text{and hence not} \quad \vdash_{\underline{A}} \dot{Con}_{\alpha}.$$

The proof is instructive as the use of (F5.5) is made explicit. γ_{α} is equivalent in Q to $\dot{Pr}_{\alpha}(\overline{\gamma}_{\alpha})$. Since every RE-formula is logically equivalent to a formula, effectively found, in bounded prenex form, (F5.5) is applicable and we get $\vdash_{\underline{P}} \neg\gamma_{\alpha} \rightarrow \dot{Pr}_{[Q]}(\overline{\neg\gamma}_{\alpha})$. We also have, since Q is finite

$\vdash_{\underline{S}} [\dot{Q}](x) \rightarrow \dot{Pr}_{\alpha}(x)$

and hence (by (F4.7i and ii)

$\vdash_{\underline{A}} \neg\gamma_{\alpha} \rightarrow \dot{Pr}_{\alpha}(\overline{\neg\gamma}_{\alpha})$

and

$\vdash_{\underline{A}} \dot{Con}_{\alpha} \wedge \neg\gamma_{\alpha} \rightarrow \dot{Pr}_{\alpha}(\overline{\gamma}_{\alpha})$

$\vdash_{\underline{A}} \dot{Con}_{\alpha} \wedge \neg\gamma_{\alpha} \rightarrow \gamma_{\alpha}$

$\vdash_{\underline{A}} \dot{Con}_{\alpha} \rightarrow \gamma_{\alpha}$

(F5.9) reports a negative result. $\underline{A}$ is reflexive just in case for each finite $F \subseteq A$, $\vdash_{\underline{A}} \dot{Con}_{[F]}$; alternatively, for each n $\vdash_{\underline{A}} \dot{Con}_{[A|n]}$.

(F5.9) Let $\underline{A} = \langle A, K \rangle$ be a consistent, reflexive axiom system, $P \subseteq A$, A recursive. There is an α^* bi-numerating A in $\underline{A}$ such that $\vdash_P \dot{Con}_{\alpha^*}$.

Not surpisingly Feferman's α^* is related to our deviant proof predicates. $\alpha^*(x) = \alpha(x) \wedge \forall z (z \leq x \rightarrow \dot{Con}_{\alpha|z} \wedge \dot{St}K(x))$, where α is a bi-numeration of A in Q.

If we restrict attention to RE-formulas we might be tempted to find a natural consistency sentence constructed from them. Since the RE-formulas numerating an axiom set are not equivalent, the field has to be pared down. For finitely axiomatized (or axiomatizable) theories this goal can be met. Let $\alpha \leq_{\underline{B}} \alpha'$ iff $\vdash_{\underline{B}} \mathrm{Con}_{\alpha'} \rightarrow \mathrm{Con}_{\alpha}$. If A is finite [A] is minimal in $\leq_{\underline{A}}$. (Of course one first applies the restrictions of (F5.6) so that $\neg \vdash_{\underline{B}} \mathrm{Con}_{\alpha}$, since if $\vdash_{\underline{B}} \mathrm{Con}_{\alpha}$, α would be minimal. For non-finitely axiomatizable systems (F7.4) tells us that a similar solution is not available.

(F7.4) Theorem: Suppose that $\underline{A} = \langle A, K \rangle$ is a consistent reflexive axiom system with $\underline{P} \subseteq \underline{A}$. Then with each α which is a PR-formula numerating A in $\underline{P}$ we can effectively associate a PR-formula α' numerating A in $\underline{P}$ for which $\alpha' < \alpha$. Under the assumption that $\underline{P}$ is w-consistent, the above holds true with "RE" instead of "PR".

Chapter 6

JEROSLOW

Although Feferman later ([Fef2] p. 129) recants his use of the term 'intensional', I think it is apt. And it is apt for just the considerations Feferman has in mind in abandoning it: "To avoid confusion with the philosophical problem of intensions it seems preferable to use other terminology...." As sections 2. and 3. show, there are ample reasons for embracing the "confusion". In later sections I will show that the term 'intensional' is apt and that there is nothing intrinsically confused about so regarding Feferman's (and others') accounts of the Second Theorem. Brilliant as it is, Feferman's approach to the Second Theorem and related matters is not the only revelatory one.

Jeroslow's approach, though intertranslatable with Feferman's, is more direct than his. It avoids the standard encodings of the usual p.r. syntactic relations and functions, whereas Feferman presents a generalized theory of those relations and functions.

Jeroslow specifically identifies formal systems with Post canonical systems: "Formal logics are not usually understood as Post Canoncial 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." ([Jer2]) Post Canonical Systems thus formalize the "presentations" of formal systems given in logic books.

In both the Feferman and Jeroslow accounts the Second Theorem is proved; as I mentioned, the accounts are in some sense intertranslatable. But Jeroslow's account has some philosophical virtues that Feferman's approach doesn't; Jeroslow's treatment makes clearer the rationale for restrictions on the class of admissible proof predicates. Loosely speaking (in the terminology of Chomsky, et al), they are equal in descriptive, but not explanatory adequacy, in the following way.

One virtue of the Jeroslow treatment hinges on the identification of formal systems with Post Canonical Systems. This identification, plus the realization that there is no theory more appropriate than concatenation theory in which to describe Post Canonical Systems, form the core of Jeroslow's approach; the rest is "follow your nose."

The justification of the approach consists largely in justifying this thesis (Jeroslow's Thesis): Formal systems are Post Canoncial Systems. Section 10 will take up what

justifying the Jeroslow (and Feferman and modal) approach comes to in this context. It suffices for now to remark that it will be their adequacy as semantic theories that meet the cavils of sections 2 and 3, that will be justified. But what of the thesis internal to Jeroslow's treatment - Jeroslow's Thesis?

Three paragraphs ago I quoted a remark of Jeroslow's in support of what we are calling Jeroslow's Thesis. Jeroslow also remarks ([Jer2] p. 6): "It should be evident at this point that the representation of a formal logic as a PCS is so straightforward, involving as it does, merely rewriting in the PCS format the usual definition of the logic, that, whatever prior ideas one may have held regarding the logic, the PCS can be understood as the object to which those ideas pertain." Thus the identification cf formal systems with PCSs is supported by the identification, in the "forensic" sense of 'identification', of formal systems as PCSs.

Jerslow is asserting that the PCS thesis is supportable on conceptual grounds prior to seeing how well the theory based on it works. We shall see that there are many post-theoretic justifications for Jeroslow's Thesis (in the sense that it makes possible generalizations of the Second Theorem) but it is well to have these pre-theoretic reasons as well.

Kreisel adduces some ([Kreisel] p. 154) 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 evidential 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. Kreisel's example, appropriately enough, involves a standard and a deviant proof predicate. As we know, the deviant proof-predicate insures consistency but the proof of equivalence cannot be carried out in any system as weak as the ones considered[24].

Now representing formal systems by PCSs provides an individuation fine-grained enough to serve in evidential contexts. To see this we will follow Jercslow in defining what a PCS is and see how our usual presentations of formal systems go over in PCSs. This will bring us to the point where Jeroslow asserted that "it should be evident" that the PCS thesis is true. Once we have reached that point, I think it will be evident that Jeroslow's claims are correct. It is important to remember that in the usual rigorous presentations of formal systems many syntactic notions are defined inductively. Note that it is in the spirit of the

---

[24]Kreisel goes on to present various reasons why, in doing proof theory, one should, as it were, be at the level of rules. Every such consideration supports the PCS thesis insofar as PCSs represent formal systems at "the level of rules".

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. Finally, note that, from this point of view, 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." ([Jer1] p. 4). Not only is this in accord with the view of formal systems as combinatorially secured producers of theorems, but also connects to the epistemological motives behind a Hilbert-style program[25]. Recursiveness is only finitude once removed. PCSs just are the required finite sets of "rules" for producing the theorems.

What is a PCS? A PCS consists of an finite alphabet An = {a1,...an}, a set of strings of An (designated 'axioms') and a finite number of production rules written in the language A ∪ S, where S is a set of symbols called string variables (A ∩ S = 0). A production rule is notated: (+) $\mu_1, \ldots, \mu_r \rightarrow \mu$, where $\mu_1, \ldots, \mu_r, \mu$ are words in A ∪ S, and rule (+) is interpreted: If upon some fixed substitution of words in A for string variables, $\mu_1, \ldots, \mu_r$ turn into words already produced, then, under the same substitution, μ is produced. The axioms are considered produced. The theorems of a PCS are all axioms and all strings

---

[25]Such epistemological scruples have surfaced in recent times in the work of Quine and, more explicitly, Davidson.

obtained by applying the rules any finite number of times. These are all the theorems.

As Jeroslow remarks and shows, PCSs are the language of inductive definitions par excellence. We shall stipulate that mnemonic abbreviations such as 'Vbl', 'Tm', etc. are to count as single symbols.

Axioms: N1

Tmf&1&1

Tmf&1&11(v1)

Production rules: $N\alpha \rightarrow N\alpha 1$

$N\alpha \rightarrow Vbl\nu\alpha$

$Tmf\&\alpha\&\beta \rightarrow Tmf\&\alpha 1\&\beta$

$Tmf\&\alpha\&\beta(\gamma) \rightarrow Tmf\&\alpha\&\beta 1(\gamma, \nu 1)$

$Vbl\alpha \rightarrow Tm\,\alpha$

$Tm\delta, Tm\theta, Tm\gamma, \delta, \epsilon \rightarrow Tm\gamma, \theta, \epsilon$

$Tm\delta, Tm\theta, Tm\gamma, \delta) \rightarrow Tm\gamma, \theta)$

This PCS (which is given on p. 4 of [Jer2]) represents the characterization of terms in a formal system. With a little attention to detail, as regards, for instance, the necessity of the last rule, the ordinary inductive specifications in standard logic texts can be translated to a PCS. As Jeroslow points out some metanotions get a little hairy - free occurence of a variable in a formula being particularly hirsute. To every PCS, F, there corresponds a PCS, Bw(F), which generates all valid production sequences of F.

EXPRESSING CONSISTENCY: GODEL'S SECOND INCOMPLETENESS

Having given a uniform method of presenting formal systems, via Post Canonical Systems, Jeroslow builds a theory to directly "encode" the syntax of PCSs. There are a number of ways to do this: perhaps the clearest is in [Jer2]. A formal theory T is characterised that meets the following conditions:

P1: $\vdash \theta$ and $\vdash \varphi$ iff $\vdash \theta \wedge \varphi$

P2: If $\vdash \theta(x)$ then $\theta(t)$ for every term t.

P3: If $\vdash \theta$ and $\vdash \theta \supset \varphi$ then $\vdash \varphi$.

P4: $\vdash \theta(x)$ iff $\vdash \forall x\ \theta(x)$.

Furthermore, there are a function symbol * (for concatenation), distinguished constant λ (for the empty string), and axioms such that various facts of concatenation theory are provable. The notion of an atom is defined: $\forall u \forall v(x = u * v \wedge v \neq \lambda \supset v = x) \wedge x \neq \lambda \equiv At(x)$.

A mapping from an arbitrary PCS to T is defined by a choice of closed terms of T, provably distinct atoms, to serve as the images ("names") of the letters in the alphabet, An, of PCS F. This induces a map from the words of F (i.e. strings of letters from An or string variables) to terms of T; each ai in An is replaced by the appropriate term (provable atom), and a string variable by a free variable of T (same for same, different for different), placing an * between symbols. This map is denoted by '-'. For

example, $\mathfrak{a}\mathfrak{c}\mathfrak{e} = \mathfrak{a} * \mathfrak{c} * \mathfrak{e}$ and $\mathfrak{b}\mathfrak{a}\mathfrak{c}\mathfrak{B}\mathfrak{a}\mathfrak{e}\mathfrak{g} = \mathfrak{b} * \mathfrak{x} * \mathfrak{c} * \mathfrak{y} * \mathfrak{x} * \mathfrak{e} * \mathfrak{g}$.

For θ a formula of T, ExtF(θ) says that $\{x: \theta(x)\}$ extends the set of theorems of F. ExtF(θ) is easily definable in T - in fact by a universal formula. $\theta(x)$ is a K-description ('K' for 'Kreisel') of PCS F if

(α): $\vdash \theta(\overline{\mu}_i)$ for $i = 1,\ldots,s$, where the μ_i are the axioms of F.

(β): $\vdash \theta(w_1) \wedge \ldots \wedge \theta(\overline{w}_n) \supset \theta(\overline{w})$ for each production of F.

(γ): If (α) and (β) hold for φ, then $\vdash \theta(x) \supset \varphi(x)$.

Trivially any two K-descriptions are provably equaivalent. Instead of (γ) one can have ExtF(φ) $\supset$ ($\theta(x) \supset \varphi(x)$). An important immediate fact is that <u>if</u> θ <u>is a K-description of F and</u> μ <u>a theorem of F, then</u> $\vdash \theta(\overline{\mu})$; the proof uses P1-P4. (This is Jeroslow's Proposition 4.) It is also immediate that if θ is a K-description of F in T then it is a K-description of F in any theory T' in the language of T that extends T.

Jeroslow then introduces a specific extension of T, which he dubs the quantifier theory of concatenation; I will call it QT. QT consists of the following axioms:

QT1) $x * \lambda = \lambda * x = x \;\&\; (x * y = x \vee y * x = x \supset y = \lambda)$
$\&\; (x * y) = \lambda \supset x = \lambda \;\&\; y = \lambda$).

QT2) $(x * y) * z = x * (y * z)$

QT3) $x * y = x * z \supset y = z \;\&\; (u * v = t * s \supset \exists y(v = y * s \vee s = y * v))$.

QT4) $x * y = u * v \;\&\; At(y) \;\&\; At(v) \supset x = u \;\&\; y = v$.

QT5) $x \neq \lambda \supset \exists u \exists v \; (x = u * v \;\&\; At(v))$

(QT1-QT5 were the axioms of T.) For every formula $\varphi(x)$

QT7) $\varphi(\lambda) \;\&\; \forall x(\varphi(x) \supset \forall v(At(v) \supset \varphi(x * v))) \supset \forall x \varphi(x)$

h is a unary function symbol.

QT8) $At(x) \vee x = \lambda \supset At(h(x))$

QT9) $At(x) \;\&\; At(y) \supset (h(x) = h(y) \rightarrowtail x = y)$

For any n one can obtain n provably distinct atoms by $a_1 = h(\lambda)$, $a_{i+1} = h(a_i)$ $1 \leq i \leq n$

$(\exists x \subseteq y) \; \varphi(x) \equiv \exists u \exists v \exists x (y = u*x*v \wedge \varphi(x))$

$(\forall x \subseteq y) \; \varphi(x) \equiv \forall u \forall v \forall x (y = u*x*v \supset \varphi(x))$

$\Theta_n(x) \equiv (\forall y \subseteq x)(At(y) \supset y = a_1 \vee \ldots \vee y = a_n) \wedge x \neq \lambda$

Jeroslow then uses An to denote either the alphabet {a1,...,an} or the PCS whose axioms are those letters and whose production rules are $\alpha \rightarrow \alpha a_i$, for i = 1,...,n. (QT7) is used to prove that $\Theta_n(x)$ is a K-description of An in QT.

Jeroslow proceeds to construct a natural definition of an arbitrary PCS in QT. Since QT is the appropriate language for formalizing directly descriptions of PCSs, this is relatively straightforward, though cumbersome.

First with each PCS F is associated a PCS Bw(F) as described above. The only detail of importance is that 'D' is used as a delimiter in the sequences representing derivations. The following is the definition of a natural description of Bw(F) ([Jer2], p. 19)

> We suppose that F is in the alphabet $A_n = \{a_1, \ldots, a_n\}$; let a_{n+1} be called D ... and let a_{n+2} be called the letter A ....
>
> At this point in the discussion, we desire to demonstrate a method for obtaining ... a K-description of F.... Our route toward this end involves first obtaining a natural description in the free variable system of a PCS F' associated with F; F' is in fact a PCS which generates those strings which can be understood as proofs in F.
>
> ... the axioms of F' are the strings $DD\mu_i DD$ for $i = 1, \ldots, s$ and the rules of F' are the following: $\alpha \longrightarrow \alpha\mu_i DD$ for $i = 1, \ldots, s$ ... plus the productions
>
> $A_n\omega_1, \ldots, A_n\omega_n, \alpha_1 D\omega_1 D\alpha_2 D\omega_2 D\alpha_3 \ldots D\omega_r D\alpha_{r+1} \longrightarrow \alpha_1 D\omega_1 D\alpha_2 D\omega_2 D\alpha_3 \ldots D\omega_r D\alpha_{r+1}\omega DD$
>
> for every production (+) of F, finally plus those productions which insure that $A_n W$ is generated precisely if W is a word in the alphabet A ....
>
> Given a finite set of q terms $t_1, \ldots, t_q$, let $(\forall i \leq q)\, \mathcal{A}(t_i)$ abbreviate the disjunction $\mathcal{A}(t_1) \vee \ldots \vee \mathcal{A}(t_q)$. Then a natural description F'(x) of the theorems of F' is given by the disjunction of $(\exists y \subseteq x)(x = \bar{A}_n * y \wedge \Theta_n(y))$ with the conjunction of the two formulae,
>
> $$\mathcal{A}_{n+1}(x) \wedge (\forall y,u,v \subseteq x)\left(\begin{array}{l} x = u * \bar{D} * y * \bar{D} * v \wedge \mathcal{A}_n(y) \supset (u = \bar{D} \vee \\ (\exists t,s \subseteq u)(u = t * s * \bar{D} \wedge (\exists i \leq n)(s = \bar{a}_i))) \\ \wedge (v = \bar{D} \vee (\exists t,s \subseteq v)(v = \bar{D} * s * t \wedge (\exists i \leq n)(s = a_i))) \end{array}\right)$$
>
> where $(\forall y,u,z, \subseteq x)$ abbreviates $(\forall y \subseteq x)\ (\forall u \subseteq x)\ (\forall z \subseteq x) \ldots$, together with
>
> $$(\forall y,z,w,t \subseteq x)\left(\begin{array}{l} x = y * z \wedge y = t * \overline{DD} * w * \bar{D} \wedge \mathcal{A}_n(w) \\ \supset ((\exists i \leq s)(w = \bar{\mu}_i) \vee Der(t,w)) \end{array}\right)$$
>
> where Der is the formula $(\exists j \leq \rho)(\exists w_1, \ldots w_{r(j)} \subseteq t)(\exists u_1 \ldots u_{r(j)} \subseteq t)\, Der'$

and Der' is $Q_n(w_1) \wedge \ldots \wedge Q_n(w_{r(j)}) \wedge t = u_1 * \bar{D} * w_1 * \bar{D} * \ldots * u_{r(j)} * \bar{D} * w_{r(j)} * \bar{D} * u_{r(j)+1} \wedge Csq((+)_j, w_1, \ldots, w_{r(j)}, w)$

(and the j-th rule of F has r(j) premises).

It is proved that this natural description is a K-description of Bw(F) in QT. The induction schema, (QT7), is needed only in proving (γ) for the natural description. Actually, this much can be proved in a quantifier-free version of QT. Finally, let $\underline{F}(x) = (\exists y)(\exists u \subseteq y)(\underline{F}'(y) \;\&\; y = u * \bar{D} * x * \bar{D} * \bar{D}) \;\&\; \Theta_n(x))$, which is a K-description of F in QT. Moreover if $G'(x)$ is any K-description of Bw(F) and $G(x)$ is any K-description of F, then $\vdash_T \exists y(\exists u \subseteq y)(G'(y) \wedge y = u * \bar{D} * x * \bar{D} * \bar{D}) \wedge \Theta_n(x)) \leftrightarrow G(x)$. And if $G'(x)$ is a K-description of Bw(F) in T, and μ a word in An, then $\vdash G'(\bar{\mu})$ iff μ a theorem of F' and $\vdash \neg G'(\bar{\mu})$ iff μ is not a theorem.

As Jeroslow remarks, arithmetic ingenuity, i.e. use of the beta-function, is needed only in finding K-descriptions in theories which do not contain QT explicitly - e.g., $\underline{P}$. In particular, it is needed to define * in $\underline{P}$ (that is to produce a conservative extension cf $\underline{P}$ containing QT. Jeroslow notes that the conservative extension is much stronger than QT. He uses his quantifier-free version to extend arithmetization to very weak theories (weaker than RA).)

A logic is K-complete if it contains a K-description of every PCS (Jeroslow shows on p. 29 that various prima facie weakenings of K-completeness are equivalent to K-complete-

ness). Jeroslow proves that no finitely axiomatizable subtheory of P can be K-complete. "In particular, that subtheory of P which contains only the existence of the concatenation * and [QT1-QT5] ..., plus any other finite number of axioms, is not K-complete. Indeed, this latter subtheory contains the natural descriptions, but these cannot be shown to be the minimal solution to the inductive clauses of the relevant PCS. To be sure, the usual encodings are "pointwise" correct, and hence can be used to obtain the First Incompleteness Theorem. But they can hardly be said to "express" in the logic J the PCS which they do in fact describe."

The technical content of this remark is explicated two pages hence by the theorem (J20). The semantic content, involving expressibility in J, is best broken into two points. Firstly, "pointwise" correct (i.e., numeralwise correct) but deviant encodings do not express the PCS which is their extension. This point has nothing to do with the weakness of the logic J. Secondly, and this is Jeroslow's real point, even the natural descriptions are not adequate expressions of some arbitrary PCS. Jeroslow puts the point somewhat misleadingly; hyphenating '"express" in the logic J' would help. What J doesn't have is the power to show that its natural description is correct - is the minimal solution to the inductive clauses. This is important, for in order for some J to establish the Second Theorem, it for-

malizes a proof of the First Theorem; and thereby has to talk of provability. We shall soon see the technical cashing in of this intuitive motivation, but an analogy might serve a purpose here.

Suppose I wish to establish that P cannot prove a certain sentence and that that sentence says that it is not provable. I would have to argue, inter alia, that a certain predicate expresses provability. Moreover, if I were to theorieze about predicates expressing provability, I would, as we have seen, not give an extensional account. That is, the conditions on a predicate, for it to express provability, are stricter than co-extensiveness. So, I would in essence be giving an intensional semantics for a small piece of discourse. Now what I have just had to (hypothetically) do, to establish that P cannot prove a certain sentence and that that sentence says that it is not provable, is just what J has to do to prove the First Theorem and thereby certify the Second. It has to contain an intensional semantics for proof predicates. What Jeroslow is saying is that sufficiently weak Js are "semantically incomplete" - they cannot establish that their descriptions are correct, though they are correct (see also p. 113).

Jeroslow further shows that for K-complete theories there is an invariance over the particular concatenation function chosen.

In Jeroslow's approach the natural notion of K-description corresponds to formalizing the inductive definition of a derivation in an arbitrary PCS. Since independent PCSs can be used for many syntactic categories, Jeroslow is thus able to get natural descriptions of number, term, variable, formula, etc., derivatively. (pp. 34-37)

So far, however, none of this has been applied to the Second Theorem. This comes about through the arithmetization of Proposition 4. As I suggested above and show in more detail in section 10, this move is the semantically significant one; for it introduces another "layer of language" that Jeroslow's machinery (helpfully) keeps separate. Proposition 4 mentions words μ and their images $\ulcorner\mu\urcorner$; in discussing it (or formalizing it) we (or the formalizing theory) must have names for them. $\overline{\ulcorner\mu\urcorner}$ is the special name for $\overline{\mu}$. The notion is described in $\underline{P}$ by a formula, SpTm(x,y), involving the beta-function. The definition yields, where $A_n(x)$ is the K-description of PCS An

$$\vdash A_n(x) \supset (\exists!y)\,\mathrm{SpTm}(x,y)$$

$$\vdash \mathrm{SpTm}(\ulcorner\mu\urcorner, y) \longleftrightarrow y = \overline{\ulcorner\mu\urcorner}$$

The proof of proposition 4 needed only P1-P3 and the definition of a K-description. P1-P3 formalize easily (P2 needs a complex substitution operation but it can be done.). (P1), for example: $\mathrm{Thm}(x)\ \&\ \mathrm{Thm}(y) \longleftrightarrow \mathrm{Thm}(x * \ulcorner\wedge\urcorner * y)$. A

theory J proves its conjunctivity if it proves the preceding formula, and similarly, it proves its substitutivity and deductivity if it proves the arithmetizations of P2 and P3.

In the subsequent Theorem, (J20), it is well to keep in mind that we have now advanced from the level of (our) remarking on the powers of J to describe an arbitrary PCS, to the level of a theory, J, remarking on its powers to describe an arbitrary PCS. I quote (J20) in full from [Jer2], p. 40:

> (J20) Let * be a K-complete concatenation in a quantifier logic J, and suppose that $\underline{F}(x)$ is a K-description of a PCS F in J. Suppose also that J proves its conjunctivity, substitutivity, and deductivity. Let $\overline{F(v_0)}$ be $A*\overline{v1}*B$. Then we have
>
> $$\vdash \underline{F}(x) \supset (\exists y)(\underline{A}_n(x) \wedge SpTm(x,y) \wedge Thm(A*y*B)).$$

Jeroslow sums up the relation of his results to standard encodings as follows:

> We consider two realms of idealized objects, the first consisting of all the strings in an alphabet A, the second of all the non-negative integers.
>
> We have languages for discussing both realms, the language of "letter" and "concatenation" for the first realm, and "number" and "plus" and "times" for the second realm. If instead of using some name 'a' for a letter a in A of the alphabet A, we use $\overline{a}$ as name, and instead of using the name "concatenation" we use the name *, then true fini-

> tist statements in the first language (the language for strings) correspond to true finitist statements in the second language (the language for arithmetic) under the usual interpretation of $\bar{a}$ as a number and * as a function of numbers. Furthermore, this fact of a correspondence can be seen finitistically, at least in the case of the concatenation * discussed in detail above.
>
> This means that, while a and the number designated by $\bar{a}$ are distinct entities, as far as knowledge is concerned we shall know as much about one as about the other, so that there is no harm in identifying the two entities in all linguistic situations.
>
> Let us therefore identify the two objects a in A and the number z designated by the closed term $\bar{a}$ Then, given the (finitist) fact of the Chinese Remainder Theorem, the standard description F(x) given by (7) of a formal system F corresponds precisely, word-for-word (modulo the cited Theorem), to the description of F as consisting of all[26] those strings one can deduce from the axioms by repeated use of the production rules. ([Jer1] p. 15)

When one is considering formal systems whose theorems form an r.e. set, Jeroslow's methods are at least as powerful as Feferman's. Adapted to quantifier-free systems, they actually extend the Gödel Second Theorem to systems that are rather weak. Moreover, by avoiding passage through a theory of recursive functions, Jeroslow's account serves its explanatory function better than Feferman's: After all, there is no need to explain the connection between formal systems and recursiveness if the identification of formal systems with PCSs is accepted.

---

[26] Jeroslow should say "all and only". I discuss the isomorphism of the second paragraph on p. 108f. Jeroslow's passage from talk of truth in paragraph two to "knowledge" in paragraph three is also discussed later.

Chapter 7

PRELIMINARY MORALS

In the light of the preceding accounts what can we now say about the cavils of earlier sections? In additon, what about the cases that Feferman covers and Jercslow doesn't - the non-RE-formulas that extend familar theories? As far as this later question goes the Feferman and Jeroslow approaches actually have ccmplementary insights to contribute.

Feferman's (5.9) tells us that there are extensionally correct descriptions of the axioms of not unusual sets of theorems (e.g., the theorems of P) whose corresponding consistency sentences (corresponding, of course, to the description) are provable. If one wishes to phrase the Second Theorem in terms of formal systems then formal systems are going to have to be individuated more narrowly. After all the same object can hardly both prove and not prove its own consistency. If the rigorous technical accounts of Feferman and Jeroslow show that the property Z, mentioned at the beginning of section 3, is indeed a respectable mentionable, then this application of Leibniz' Law is surely one to which no objection can be taken. More-

over, both the F-account and the J-account supply us with the appropriate intensional objects - the formulas t on the one hand and PCSs on the other.

This does not yet settle the question of whether (8) is true. Can some formal systems prove their own consistency? This could be answered by answering 'Do non R.E. t's represent formal systems?' in the negative. One is then saddled with supporting this answer. On the J-approach, the truth of (8) is built in; and built in via the identification of PCSs with formal systems. We thus have two, intimately related Church-type Theses. They have in common the claim that some descriptions do not describe formal systems, irrespective of the recursion-theoretic character of the set of theorems picked out. As I have argued before (pp. 16-17) stipulating such a claim, in the interest of being able to state the Second Theorem clearly and cleanly (e.g., as (8)), is not an option.

Of course, the distinction between RE- and non-RE-formulas is not ad hoc; what is missing is the connection between RE-formulas and our conception of a formal system.

What does a formula like α^* say? Let A* be the formal system thatα^* describes, and A the formal system that α describes. x is an axiom of A* just in case x is an axiom of A and the set of all axioms of A (with godel number) less than or equal to x forms a consistent theory. There is a sense in which α^* describes a subsystem of A, for A* is con-

sistent regardless of A's consistency. But why is A* not a formal system? One might suspect the notion of formal rules of proof conflict with non-REness. As we shall shortly see, the culprit is the unbounded universal quantifier in the consistency clause of of the definition of α^*. (The presence of this quantifier also blocks a formal derivation of an undecidable formula on hypothsis of Con_{α^*}.)

Consider the task of checking whether an arbitrary object, x, is an axiom of A*. To show that x is an axiom, infinitely many statements are needed: 1 is not the godel number of a proof of a contradiction from axioms $\leq x$, 2 is not the godel number of a proof of a contradiction from axioms $\leq x$, ... Such a procedure violates our conception of a formal system, involving as it does non-effectively given conditions axiomhood and hence proofhood. This spells out the informal consideration that the extension of α^* is fixed as the extension of α only on hypothesis of consistency - which hypothesis is not effectively verifiable.

We can construct an interesting puzzle about deviant formal systems such as the ones introduced by Feferman. Consider $\underline{P}$. Take a non-deviant consistency sentence for $\underline{P}$, ConP and let $P_1 = P \cup ConP$, and in general, $P_{n+1} = P_n \cup ConP_n$. Let $P_\omega = \cup P_n$. Clearly $\underline{P}_\omega$ is consistent - if not the proof of the contradiction would be in some $\underline{P}_n$. But then some $\underline{P}_n$ would be inconsistent and each $\underline{P}_n$ is consistent. Moreover, P_{n+1}"knows that". The above reasoning is elementary and $\underline{P}_\omega$ can prove the

consistency of each P_n. Why doesn't this show that P_ω can prove its own consistency?

The answer is that $\underline{P}_\omega$ is not a formal system (on the conception of a formal system that we have been examining) and so that it is self-"provably" consistent is, as above, not surprising. This is to take $\underline{P}_\omega$ as given by its mode of presentation above - and of course that mode guarantees its consistency[27].

There is a slight sleight-of-hand in the presentation of this brain teaser. Each theory along the way results from the addition of an axiom, $\mathrm{Con}P_n$ to the previous theory and closure under rules of proof. The P_n are axiom sets and $\underline{P}_n$ the associated theory. P_ω is the union of all the axioms sets P_n and $\underline{P}_\omega$ is its associated theory. $\underline{P}_\omega$ is not given as the union of the $\underline{P}_n$ - that's a different puzzle. (Technically, of course, the problem is that of another universal quantifier, occuring in the description of P_ω. Let α_0 numerate P in $\underline{P}$, α_0 RE. $\alpha_{n+1}(x) \equiv \alpha_n(x) \vee x = \dot{\mathrm{C}}\mathrm{on}_{\alpha_n}$. $\alpha_\omega(x) = \exists n\, \alpha_n(x)$. What our above argument shows is $\forall n \vdash_{P_{n+1}} \dot{\mathrm{C}}\mathrm{on}_{\alpha_n}$ and hence $\forall n \vdash_{P_\omega} \dot{\mathrm{C}}\mathrm{on}_{\alpha_n}$. It then takes a dubious step to get $\vdash_{P_\omega} \dot{\mathrm{C}}\mathrm{on}_{\alpha_\omega}$; namely to $\vdash_{P_\omega} \forall n \dot{\mathrm{C}}\mathrm{on}_{\alpha_n}$ and $\vdash_{P_\omega} \forall n \dot{\mathrm{C}}\mathrm{on}_{\alpha_n} \leftrightarrow \dot{\mathrm{C}}\mathrm{on}_{\alpha_\omega}$.)

In fact what the provable statement amounts to is that every stage is consistent. This is analagous to the situa-

---

[27] Feferman has shown that there is an RE-formula whose extension is P_ω. Of course the consistency sentence formed from it is not provable in $\underline{P}_\omega$.

tion with reflexive theories, the theories for which Feferman proves (5.9) (5.9)[28].

The above are some of the ways in which the work of Feferman and Jeroslow illuminate the nature of the deviant predicates and supply the individuation conditions needed to quarantine them. The theses associated with each account have received support as correct explications of the notion of a formal system. In a subsequent section I will have much to say about what this shows about the nature of these accounts, when applied to the problems first raised in sections 2. and 3. As inevitable as I think the journey to my views in section 10 is, there are some who disagree. Feferman seems, implicitly, to demur (see p. 56); but the only explicit disagreement is taken up in the next section.

---

[28]Feferman's paper on transfinite recursive progressions is the serious working out of the idea behind this anecdotal system. The idea goes back to [Turing]. The intensional correctness of the proof predicate is foundationally important here; the idea is to achieve epistemologically secure extensions of standard formal systems. This is done by having the reflection principles express the soundness of each system on the way up. This is another instance where "express" is to be taken seriously in order for technical results to have philosophical (in this case, foundational) interest.

Chapter 8

OTHER VIEWS

Very little has appeared in the philosophical literature concerning the recalcitrance of the Second Theorem. Kreisel's hybrid pieces are extremely important, but the only article specifically on this topic has been Michael D. Resnik's "On the Philosophical Significance of Consistency Proofs" [JPL 3 (1974) pp. 133-147].

Resnik first relates the importance of consistency proofs to Hilbert's program - in particular the demand for finitistic consistency proofs. He then seeks to explain the relevance of the Gödel results to this demand. He claims that the First Theorem does not bear much on the program. This is because the Gödel sentence is an <u>ideal</u> sentence[29]. Resnik takes the content of the First Theorem to be that certain formal systems are incomplete and hence cannot prove all truths. But, says Resnik, it was no part of the Hilbert program that they should - just the truths expressed by <u>real</u> sentences. A version of Gödel's First Theorem can be

---

[29]See Hilbert's "On the Infinite" in [P-B], especially pp. 143-149.

[30]'Can be' since, in order to obtain a simple A, only particular constructions of Pr will do. That Pr(A) $\longrightarrow$ A is

proved[30] according to which Pr(A) → A is unprovable for a particular universal A. The purpose of a formal system for Hilbert was to allow manipulation of ideal sentences to facilitate the proof of real sentences. Since Peano arithmetic contains at least finitist reasoning, Pr(A) → A is not established finitistically.

Resnik would reply, on Hilbert's behalf, that A is not real; its schematic version is. (The schematic versicn of A is just A with its quantifiers dropped.) A schematic sentence counts as provable if each instance of it is provable in the ordinary way. And that each instance of this schematic formula is so provable is itself finitistically provable. (Since we are establishing provability we may assume consistency.) This reply on Hilbert's behalf only shifts him from one horn of a dilemma to another. Because of ω-incompleteness the First Theorem <u>also</u> dooms the schematic interpretation as inadequate, as we shall now see.

Horn one is the observation that for some sentences A, Pr(A) → A cannot be established finitistically. Replying that those A's don't matter because their schematic versions are adequate replacements for them is futile; they aren't. This is revealed by the phenomenon of ω-incompleteness (the existence of unprovable universals of which every numeral

---

provable iff A is, a strengthening of this version of G1, is really a version of G2; requiring that Pr be non-deviant.

instance is provable). This is a First Theorem result. And this does not just show that a universal and its schematic version don't mean the same thing, but also that one cannot replace the other in a provability context, salvae veritate.

So Resnik's attempt to minimize the relevance of the First Theorem to Hilbert's Program founders. Resnik is right to point out that the non-formalizability of arithmetic truth doesn't doom the Hilbert program; but the limits on the formalization of (finitist) proof do.

However, once the First Theorem is taken into account, then the Second Theorem does become relevant to the Hilbert program in a number of ways. For, without the fact of the First Theorem in view, a Hilbert-type program could concentrate on a few formalisms which held hope of codifying mathematical practice. Since, for each such we can construct a patently correct proof predicate, the above remarks apply; that is, we get an A of the right form. The initial involvement of the Second Theorem is not in its statement, but in the techniques developed in order to state it. For, given essential undecidability, a Hilbert-type program would adopt and consider a large range of formalisms. And the argument above, concerning $Pr(A) \rightarrow A$, although a version of the First Theorem, then requires intensionally correct provability predicates for a wide range of formal systems. Then the Gödel results entail that if Pr is correct then it is not self-provably correct. So each modification of proof

leads to a proper extension. (This is the idea behind ordinal logics, via reflection principles. To paraphrase Kreisel: If a formal system formalizes finitist proof then it doesn't know it. For then $Pr(A) \rightarrow A$ is true, but, for some A, not provable. In fact it is provable in arithmetic that $Pr(A) \rightarrow A$ is provable iff A is provable. So $Pr(A) \rightarrow A$ represents the truth that the formalism doesn't "know". This result is taken up in more detail in the section on the modal interpretation.)

The techniques of the Second Theorem thus make it relevant to a Hilbert-type program, though not in the way Resnik wished to argue. He goes on and wants to conclude that the Second Theorem tells us that there is no answer to a certain kind of sceptic (one who demands that the consistency proof aspect of Hilbert's program be carried out). Feferman's work is introduced as a possible counter to this. First, however, Resnik introduces the provability-of-every-instance deviant consistency sentence (see p. 38), shows it to be provable, and correctly points out that it is not intuitively a consistency sentence. He concludes: "[T]he very weakness of this sense of consistency casts doubt upon Hilbert's suggestion that schemata can be used as approximations to unbounded universal quantification." But this is, of course, an insight to be garnered from the First Theorem; once again defeating Resnik's denigration of it as irrelevant to the Hilbert Program.

Resnik then quotes Feferman's (5.9) as the possible reply to skepticism. He rightly shoots this down, as the deviant consistency sentence doesn't express consistency - actually doesn't express consistency independently of the consistency of the formalism involved[31].

What I find puzzling about Resnik's piece, apart from his remarks on the First theorem, are the morals he draws. He says: "Perhaps the conditions on the axiom predicates must be given in terms of their syntactic form or their intensions. The latter alternative is repellent not only because it is vague but also because it introduces intensionality into mathematics." I find this a particularly confusing (or confused) passage, for the following reasons:

As stated by Resnik, the latter alternative is vague; but surely such vagueness is not an intrinsic feature of the project of giving conditions on axiom predicates. I will later argue that the Feferman and Jeroslow accounts are in fact semantic accounts - and vagueness is certainly not a fault of either of these accounts. Resnik's second reason for finding the latter alternative repellent reflects a mere unargued prejudice. More vitally, I am confused by the

---

[31] If the system is inconsistent the deviant consistency sentence says that a certain finite subsystem is consistent (and by reflexivity this is guaranteed provable). So to show that the consistency" sentence is a consistency sentence of the intended theory, one would have to prove consistency. Of course, this is sufficient to disqualify it as an answer to the skeptic.

dichotomy he presents. There are legitimate questions as to what the correct syntactic theory will be so as to allow a correct and adequate semantics - but these hardly creates Resnik's mysterious dichotcmy.

The predicates will hardly be proof predicates (or any such kind of predicates) without interpretation; and interpretations run off syntactic representations[32]. It also is instructive that, for the case of proof predicates, we shall have an intensional semantics whose entities are themselves syntactic objects. So I can find no sense in Resnik's objection. I will return to the methodological issues to be confronted here after outlining yet another principled treatment of provability.

---

[32]Perhaps this point is not familar phrased this way, but it is exemplified even at the level of propositional calculus. Let a conjunction be the string formed by placing an & between sentences, a disjunction the string formed by placing a V between sentences. Such strings are themselves sentences. Now a disunction (conjunction) is true if either (both) part(s) is (are) true. Let A, B, and C stand for sentences. Then A&BVC is a sentence. Let C be true and A and B false. Then by the semantic rule for disjunction A&BVC is true if A&B is true or C is true. Since C is true, A&BVC is true. By the rule for conjunction A&BVC is true if A is true and BVC is true; they aren't so A&BVC is false. Thus do semantic theories run off syntax - bad syntax, bad semantics. (I owe this kind of example to Harold Levin, who puts it to more sophisticated uses.)

Chapter 9

THE MODAL TREATMENT

There is another technical solution to the phenomena surrounding the Second Theorem, which it will be useful to add to our repertoire. Historically, it originates from a paper of Gödel's ["An Interpretation of Intuitionistic Sentential Calculus", Ergebnisse eines Mathematischen Kolloquiums, IV, p 39 (1931)], the derivibility conditions of Hilbert-Bernays, the work of Lob and recent work in modal logics. The Gödel paper supplies an interpretation of Heyting's sentential calculus "by means of the concepts of ordinary sentential calculus and of the concept "p is demonstrable"." An ordinary sentential calculus is supplemented by an operator B and the axioms[33]:

(35) $Bp \rightarrow p$

(36) $Bp \rightarrow (B(p \rightarrow q) \rightarrow Bq)$

(37) $Bp \rightarrow BBp$

and the rule of inference: BA can be inferred from A. Call this system S4 (because it is). Gödel warns us than one

---

[33]The B stands for Box, $\Box$, which I will also use.

cannot interpret B as "demonstrable in a definite formal system." For if the formal system contains arithmetic, then, since $B(Bp \rightarrow p)$, $B(0 \neq 0) \rightarrow 0 \neq 0$ would be a theorem and, so would $\neg B(0 \neq 0)$, contradicting the Second Theorem. Is there a notion of demonstrable answering to S4? The axioms seem true to conception of demonstrability. If p is demonstrable then there is a demonstration of p and exhibiting that is a demonstration of the demonstrability of p. This validation of (37) trades on our feeling that demonstrations, or proofs, are ultimately recognizable as such. We certainly respect this intuition in formalization.

The rule of necessitation seems plausible on similar grounds - but note that it hauls the strength of S4 into whatever system B is being used to describe. And this is a culpable move. For while $Bp \rightarrow p$ is true and a theorem of S4, we should not expect the system S4 is describing, the B-system, to be able to prove each instance of $Bp \rightarrow p$ just because S4 can prove $Bp \rightarrow p$[34]. Is the problem just the application of necessitation to (35)? If so, the system attained by taking the theorems of the system whose modal axioms are (36) and (37) with the rule A/BA, adding every instance of (35), and closing under modus ponens, should be a candidate for truths about formal provability.

---

[34]Kreisel, in Ord. Log..., constructs a system obeying axioms (35)-(37).

Consider the Lob axiom.

(38) $B(BA \rightarrow A) \rightarrow BA$

Note that if (38) is added to the original system, a rather undesirable system results. However, (38) is true for the interpretation of B as formally provable. If we take (36), (37) and A/BA, then in that system $BA \rightarrow A/A$ is a derived rule. Moreover, when we consider an actual theory, e.g. P, and a provability predicate, such as Bew, for P, then $\vdash_P Bew(\ulcorner A \urcorner) \rightarrow A$ only if A is a theorem. These results can be summarized and made precise as follows. (see [Boolos] for details)

Let G be a modal propositional calculus whose axioms are all the tautologies, all sentences of the form $B(A \rightarrow A') \rightarrow (BA \rightarrow BA')$, and all sentences of the form (38). The rules are modus ponens and necessitation (A/BA). (The primitives are $\rightarrow$, B, and $\perp$, see [Boolos], chap. 1.) $BA \rightarrow A$ is not a theorem of G. Let G* be a modal propositional calculus whose axioms are the theorems of G, each instance of $Bp \rightarrow p$, and which is closed under modus ponens.

A realization is a function that assigns to each sentence letter a sentence of the language of P. The translation A^{φ} of a sentence A under a realization φ is defined by:

i) $p^{\varphi} = \varphi(p)$

ii) $\bot^{\varphi} = \bot$

iii) $(A \rightarrow B)^{\varphi} = (A^{\varphi} \rightarrow B^{\varphi})$

iv) $(BA)^{\varphi} = Bew(\ulcorner A^{\varphi} \urcorner)$

where '$\ulcorner \urcorner$' denotes the Gödel numeral function.

It turns out that G not only has nice properties as a modal logic with a Kripke semantics (complete and decidable) but also the property that $\vdash_G A$ iff $\vdash_P A^{\varphi}$, for every realization φ. It is in this sense that G contains all the provable (in P) facts about provability in P. As we saw, it does leave out some facts - viz, all the theorems of G* that are not theorems of G. (It also leaves out facts that are intrinsically quantificational in nature. I shall return to this point.)

Our deviant provability predicates are ones which fail to satisfy (36), (37) and necessitation. If B(y) is a provability predicate then $D(y) = B(y) \ \& \ y \neq \ulcorner 0 = 1 \urcorner$ is not (violating (36) though satisfying (37)).

Our previous accounts of the deficiencies of the deviant predicates relied on our seeing their deviance from a standard notion of a formal system. One deviant consistency sentence, e.g., asserted the consistency of a trivially consistent subsystem - which given consistency, is coextensive with the intended system. Alternatively we pointed to the non-effective nature of the implicit rules of proof - that is, we took seriously the description of the formal system as intrinsic to its formality. Now we have another method.

Our deviant consistency sentences had the presumptively undesirable property of violating the Second Theorem. But the Second Theorem is arguably not an intuitively apparent property of provability (and expressibility). The conditions on a provability predicate are. We have indicated good reasons for their truth[35]. The system G*, constructed above, has the property of reporting all truths of provability in P (though only modulo the notion of translation defined above). G* is not closed under necessitation.

Since G (and G*) are decidable, these completeness theorems allow us to easily obtain many metamathematical results about P. One of particular interest for our subse-

---

[35]This may be a little cavalier with respect to the truth of $Bp \rightarrow BBp$. Roughly speaking, troubles arise for it if we consider an informal notion of proof and the problem of surveyability. That is, proofs so long that it is impossible to comprehend them. And there are after all arbitrarily long proofs. Now I think such examples, if set out in detail, trade on conflating "seeing" a proof, in some holistic way, and verifying it. And since we are talking of formal provability, we are conceiving of proofs each of whose steps are verifiable. If we have a formal proof, however long, then it is elementary to certify it as a proof.
What often comes to mind in this context is the problem of verifying a proof procedure, or a description of a method for generating a proof. Here, certain results in computation theory are relevant. It is known that there is no program that will verify the correctness of a program, since that would be to solve the halting problem. More fine-grained results concern the complexity of (non-universal) program-checkers. A recent case in point is the solution of the four-color theorem (nee conjecture). Someone committed to some version of the a priori of mathematical knowledge would locate the proof in the verification that the program for searching the cases is correct.

quent discussion is the DeJongh-Sambin Theorem. Let A(p) be a sentence of G modalized in p. Then there is a sentence H, not containing p (but possibly containing the other letters of A) such that $\vdash H \leftrightarrow A(H)$ and $\vdash_G B(p \leftrightarrow A(p)) \rightarrow B(p \leftrightarrow H)$.

Consider $\neg Bp$, a sentence of G modalized in p. We know, by the fixed point theorem, that there is a sentence S, such that $\vdash_T S \leftrightarrow \neg Bew(\ulcorner S \urcorner)$. This is equivalent to $B(p \leftrightarrow \neg Bp)^{\varphi}$ being true for $\varphi(p) = S$. If A is <u>any</u> sentence modalized in p and there exists an S such that $B(p \leftrightarrow A)^{\varphi}$ is true when $\varphi(p) = S$, then S is said to be a fixed point for A. The DeJongh-Sambin Theorem tells us that every sentence modalized in p has a fixed point and that all fixed points of A are equivalent. Moreover there is a deictic fixed point. <u>Deictic</u> sentences of arithmetic are the translations of the letterless ones of G. The truth-values of such deictic sentences are effectively calculable, and whether or not a deictic sentence is provable is decidable, so that there is a decision procedure for the provability of these fixed points. For instance, a Gödel sentence is provable just in case $\neg Bew(T)$ is provable, i.e., just in case arithmetic is inconsistent.

These are the essential facts that we shall need concerning the modal interpretation. However, certain possible confusions should be forestalled. There is a purely formal sense in which these results about G represent a reduction from grade 2 to grade 1 modal involvement, via the transla-

tion mapping. [Montague] is thought to have shown this impossible. But G^{φ} does not have all the "modal" properties that Montague requires; in particular $\neg \vdash_{\mathrm{p}} (B(p) \rightarrow p)^{\varphi}$, for some S, $\varphi(p) = S$. Montague understresses an important condition if his theorem is to bear an interesting interpretation. The elementary theory, T, has to be adequate for elementary arithmetic (which he says) <u>and</u> '⌜⌝' is to be interpreted as an elementary <u>arithmetic</u> function. Alternatively: T is to be adequate for elementary syntax and '⌜⌝' is a <u>function</u> symbol (i.e. has the semantics of a function) with the appropriate metatheorems. That is, if we godel number, then '⌜⌝' is our name for a certain definable function of arithmetic; if we directly formalize into some rich enough concatenation theory then '⌜⌝' has be a functional term. Why? Well, loosely put, the characterization of '⌜⌝' as a primitive might be given a semantics with no corresponding deductive power. More precisely, 'the denotation of ⌜A⌝ is A' is a schema of the metalanguage laid down as the semantics for '⌜⌝', not as a meta-theorem arising from a syntactic characterization. If this is done then the theory will not be able to prove certain concatenation theoretic facts involving sentences containing '⌜⌝'. But the semantics will supply correct truth-values.

A working out of this idea has been presented by

---

[36]Skyrms' account was given as a paper at UNC-Chapel Hill.

Skyrms[36], though not, I think, in service of my point. I synopsize: Let L_o be a language wih a model theory. Two theories are derived from L_o: L_M which is L_o closed under truth functions and modal operators; and L_ω, which is defined metalinguistically, in a uniform way, from L_o (including a model theory). Each sentence of L_M is assigned a sentence of L_ω. The set of sentences of L_M whose L -correlates are true in all models of L_ω is the <u>model theory induced by L_o</u>. Mimimal assumptions about L_o and its models: Each model assigns denotations to pieces of L_o and all sentences are assigned 1 or 0. L_o contains PC and sentences are finite. L_M gets the obvious definition. L_ω is defined as the union of the L_n, where the sentences of L_{n+1} are the <u>smallest set</u> such that i) $*Q(S)$ is a sentence of L_{n+1}, if S is a sentence of L_n. ii) truth functional closure. C: $L_M \rightarrow L_\omega$ is such that i) S has no modalities $\rightarrow C(S) = S$, ii) S is $\Box R \rightarrow C(S) = *Q(C(R))$, and iii) commutes with truth functions.

The interest comes in the construction of models for L_ω. Note that the construction of L_ω is overly cumbersome. It could have been defined <u>just</u> like L_M, with $*Q$ behaving like $\Box$. So it must be the model theory that makes this interesting.

My synopsis is as accurate as memory permits.

Models are given by denotation functions f_n. The model for L_{n+1}, f_{n+1}, is the smallest extension, f, of f_n, such that i) $f(Q(S)) = S$, ii) $f(*X) = 1$ iff $X = Q(S)$ and $f_n(S) = 1$ for all f_n, and iii) truth functions. The model of L is the union of the f. It is vital to note that the interpretation of Q is <u>not</u> as a function. Q's interpretation is given metalinguistically piecemeal, so that the Q(S) are really constant terms.

To make this clear consider an alternative ccnstruction to Skyrms'. Add to the category of terms of L_0, $q_0, q_1, \ldots$ and a "predicate" *. If S is a sentence of L_0, S is a sentence of L. For all i, $*q_i$ is a sentence of L_ω. Close under truth functions. Let $S_1, S_2, \ldots$ be an enumeration of the sentences of L_ω. Let Q(n) = the first q_i not occurring in $\{S_1, \ldots, S_n\}$. If $S \neq *q$, $f(S) = f_0(S)$. $f(q_i) = S_{Q^{-1}(q_i)}$. $f(*q_i) = 1$ iff $ff(q_i) = 1$ for all f.

This is only a notational variant of Skyrms' construction, but it makes clear that Skyrms' Q is not a function term; my metalinguistic Q is. Skyrms uses this to clarify Montague's theorem. For, *Q satisfies all of Montague's modal laws, yet L_M is clearly consistent. It is not a counterexample to Montague, because, as I have tried to emphasize, 'Q' is not a predicate, and so cannot be diagonalized.

There are a number of ways to enrich the Skyrms scheme to yield problems. Whatever the original domain, Skyrms

introduces new entities, namely the sentences, which obey no laws. We may either relate these sentences to the regular domain, or introduce a theory of sentences directly. If the old domain contains numbers, and we relate the sentences and the numbers by godelization, and the base theory is adequate for elementary syntax, then we've got Montague. What about the other way?

If we think of applying Skyrms' construction, and if the quotation device is to be an adequate reflection of quotation in English, then surely we can ask a little more out of it. But if we add enough, we will be able to "norm", to diagonalize, and we'll have Montague's theorem again.

The moral is that G (and G*) have nothing to do with the logic of necessity. This is why the existence of a Kripke semantics for G is less interesting philosophically, for our immediate concerns (except insofar as it yields a decision procedure), than the Solovay completeness theorem. For G the Kripke modeling is (at present) a purely formal characterization theorem, whereas its completeness in G is a completeness result[37]. This is parallel to the situation in the logic of necessity. Characterization theorems, using topological methods, existed for various modal systems. The

---

[37]I am marking a distinction here that isn't always made, and am appropriating the word 'completeness'. The reader may regard the completeness/characterization distinction as my own coinage, for the duration of this section.

Kripke modelling can count as a response to Quinean doubts insofar as it formalizes our informal Leibnizian notion. Because of its relation to the intuitive semantics I have dubbed the Kripke results 'completeness results'; and the topological models 'characterization theorems'. The advent of Kripke semantics didn't undercut later Quinean scruples, but, because possible world models are the natural models for modal logic, they helped clarify what the issues were. Quine and Kripke have been clear on this (at least lately, see, e.g., "Worlds Away"). But Leibnizian intuitions have nothing to do with G. G is interesting because of the Solovay completeness result, not the Kripke style.

Lastly, it should be pointed out that although G is "completely true" to Bew it doesn't follow that G uniquely characterizes Bew. Are there other predicates of arithmetic that make G true? Yes. Let Sent be a predicate of P that defines sentencehood. Let $(\bot)^{\psi} = \bot$, $(A \rightarrow B)^{\psi} = A^{\psi} \rightarrow B^{\psi}$, $(BA) = \mathrm{Sent}(\ulcorner A^{\psi} \urcorner)$, and $p^{\psi} = \varphi(p)$, where φ is an assignment of sentences of P to letters of G. Thus the image of A under ψ is just like the image of A under φ, with Bew replaced by Sent. It is trivial that ψ takes theorems (of G) to theorems (of P).

What about the converse? Is G complete for Sent? Well, $\vdash_P \mathrm{Sent}(\ulcorner \bot \urcorner)$. But $\neg \vdash_G B\bot$, since $\neg \vdash_P \mathrm{Bew}(\ulcorner \bot \urcorner)$, and mere 1-consistency insures that.

Consider two translations, the Bew one, and another. Suppose $\vdash_G A \leftrightarrow \vdash_P A^{\varphi}$ and $\vdash_G S \rightarrow \vdash_P S^{\psi}$ and suppose that the image of A is a theorem ($\vdash_P A^{\varphi}$) and the image of S is not ($\neg \vdash_P S^{\psi}$). Can A = S? Clearly not.

But let φ and ψ be as above. The biconditonal is Solovay's completenss theorem. If A were the formalization of the inference from 1-consistency to the unprovability of Bew($\ulcorner\bot\urcorner$) (i.e. 1-CON $\rightarrow$ $\neg$BewBew($\ulcorner\bot\urcorner$)); and if S^{ψ} = 1-CON $\rightarrow$ $\neg$SentSent($\ulcorner\bot\urcorner$); then we would seem to have a problem. For, A^{φ} is a theorem, S^{ψ} isn't but it looks as though their preimages are identical. Of course the answer is that there is no preimage; there is no letterless sentence of G whose translates are as required. The arithmetization of 1-consistency is not a deictic sentence. It, and w-consistency, require quantificational structure.

Let A^{φ} be CON $\rightarrow$ $\neg$Bew$\neg$Bew($\ulcorner\bot\urcorner$). This is the formalization of the Second Incompletenss Result. So A^{φ} is provable. Since A^{φ} is really $\neg$Bew($\ulcorner\bot\urcorner$) $\rightarrow$ $\neg$Bew$\neg$Bew($\ulcorner\bot\urcorner$), S^{ψ} should be $\neg$Sent($\ulcorner\bot\urcorner$) $\rightarrow$ $\neg$Sent$\neg$Sent($\ulcorner\bot\urcorner$), whichis clearly provable. And, patently, A = S. In general, given a letterless sentence of G and a predicate of arithmetic, call it Pred, that is "true to G" (i.e. G is sound for it), then if the Bew-translate is provable so is the Pred-translate (conversely if G is complete for Pred).

The above leaves open the question as to whether there is a predicate of arithmetic for which G is sound and com-

plete and not Bew. More precisely, does there exist a predicate, Pred, such that Pred-translation from G to P preserves theoremhood and (45) (or (46) or (47))

(45) $\neg$(Bew$\leftrightarrow$ Pred)

(46) $\neg \vdash$ Bew$\leftrightarrow$ Pred

(47) $\vdash\neg$(Bew Pred)

is true? Even for (47) the answer is yes. Let Pred(x) = Bew(x) V $\neg$Sent(x). (47) holds, since P can verify that non-sentences are not provable. Soundness and completeness hold for <G, Pred> for the same reasons they hold for <G, Bew>. Any term occurring in a Pred-translate as argument to Pred is a term denoting a sentence - hence Bew(t) and Bew(t) V $\neg$Sent(t) will have the same truth-value and provably so. So every Pred-translate has a corresponding, co-provable, Bew-translate.

This only shows that we asked the wrong question. Let us restrict Pred's extension to sentences. The proof predicate for some consistent[38] extension of P, say P with the godel sentence as an additional axiom, will be just like Bew. This predicate is not coextensive with Bew, though not provably not coextensive. To show some extension not coex-

---

[38]If not consistent, then Pred is Sent and <G, Sent> is not complete.

tensive with Bew requires showing something not a theorem of P, and P cannot do that. Obviously this generalizes to Bew(⌜A→ ⌝).

A moment's reflection will reveal that this non-categoricity is not only not surprising but desirable. The intent of G, and G*, is a codification of certain general facts about provability - but not enough to fix the reference. This generality is reflected in the notion of translation employed above. Since 1931 we know better than to try to fix such a reference - even on P. As pleasant as P is it would have cast suspicion on G if it were limited to P. In this regard it is helpful to see, not just that <G, P> is sound and complete, but that G be true to a reasonable notion of provability. This I essayed some pages back. The project is, of course, more apt for G*, G being reserved for what is provable about provability.

This kind of intuitive, or first-principled, justification of G, or G*, can be formalized if we can independently formalize the notion of a formal system. This is just the vaue of the Jeroslow treatment. Jeroslow's constructions are patently an explication of our (Hilbert's) notion of a formal system, from which Löb's Theorem is forthcoming. Since the provability of the translate of the characteristic axiom of G, in P, is the provability of Lob's Theorem, we have a justification of G. This is analagous to justifying a somewhat baroque characterization of recursive function by

showing its equivalence to a more Bauhaus version. The strength of this justification depends on just how plausible and patent the Jeroslow formalization of formal systems and their syntax is. I think an examination of it revealed that it is patently plausible and patent.

Chapter 10

PROOF THEORY AS SEMANTICS

I take it as a piece of received wisdom that mathematics is the paradigm realm of the extensional. At least I think many philosophers have believed something that could be put that way. Furthermore, I take it that asked to defend such a claim, one would typically mention two (not unrelated) facts: The existence, due to Frege/Tarski, of a certain sort of semantics, and the seeming absence of intensional contexts in mathematical discourse. It is, of course, necessary to clarify what is meant by the claim that mathematics is the realm of the extensional, and how the two vaguely indicated facts support the claim. (It is not hard to find supporters of the claim. It is implicit in Frege, and explicit in Russell and Whitehead. See below p. 103)

Before briefly amplifying and clarifying the above, I would like to make a few observations concerning the process of "extensionalizing". These observations are not intended as a precise explication, but rather a sketch of some philosophical folk wisdom. In the course of time many mathematical theories have been applied to various areas of knowledge. Certainly physics, but also population genetics, information theory, traffic control, economics, metaphysics, etc., have in some sense been mathematized.

In some cases of mathematization, extensionalization is claimed. Set theory is the great extensionalizer - with discourse, pure or applied, about properties, relations, functions, etc. being extensionalized. Whatever this is supposed to mean in any precise way, those who claim particular successes in the extensionalizing project are wont to dismiss objections that advert to features present in the unreduced discourse by claiming those features to be unimportant, dispensible, or part and parcel of the incoherence of intensionality.

Alternatively, the intensionality may be replaced by a feature that is equally repugnant to the extensionalist. A piece of (third-grade) modal discourse is not salvaged for Quine by the existence of an applied Kripke semantics for it; such a semantics being committed to dubious entities and relations. Although the admirable precision and clarity of mathematization was often accompanied by what I have been calling "extensionalizing", the case of Kripke semantics makes it clear that the clarity of the pure mathematics is not always inherited by its applications.

What I will show is that the language of metamathematics is intensional, that the available theories that are applied to it are adequate and provide an intensional semantics for it. Construed as semantic theories, these rigorous accounts are not subject to objections based on commitment to dubious entities and relations. This is simply because the seman-

tics, albeit intensional, is committed only to syntactic objects and relations among them. Moreover, there are no viable extensionalizing alternatives; in the sense in which the semantical theory of Russell/Smullyan might be thought to be an alternative to that of Frege/Church. Finally, I will sketch the connections that these results have with philosophy of language. First, the promised clarification.

What are said to be extensional are contexts of a language. 'Language' here must mean interpreted language. Intensionality is signalled by a failure of extensionality and extensionality is to be defined in terms of an extension function from expressions to a domain. Contexts will be extensional or non-extensional with respect to this extension function[39]. These notions cease to be purely formal if the extension function is an adequate extension function - i.e. actually assigns the denotation of names to names, extensions of predicates to predicates, etc. What this means, of course, is that we are talking of languages that we know or are making up.

Failure of extensionality is not a gurarantee of intensionality. 'Intensionality' connotes an involvement of semantic notions such as meaning, analyticity, synonymy and

---

[39]A purely syntactically specified language may have a notion of extensionality associated with it only insofar as it has, e.g., rules that are interpreted as truth preserving and a symbol interpreted as equality.

so on. And, after all, predicates formed from a predicate and the quotation operator, like ' ' ' has five letters' or ' ' ' rhymes with vodka', are non-extensional but permit of substitution of co-related expressions; where the relation in question is neither same extension as nor same meaning as, nor is it characterized in an intensional vocabulary[40].

Intensional contexts were noted by [Brentano] and exploited by Frege to serve as data for arguments establishing the two-dimensional nature (sense/reference) of a semantics for a natural language. A semantic theory for a particular language has to be empirically adequate. A semantic theory that did not allege synonymy when synonymy was present would be inadequate - if synonymy is thought to be a pre-theoretic notion of sufficient coherence to be theorized about. The importance of Frege's use of intensional contexts is that it links the intensional and referential portions of semantic theory. For a semantic theory inadequate to facts about meaning will also be inadequate to facts about more respectable notions such as reference and truth[41].

---

[40] I owe this point to Richard I. Nagel.

[41] I am omitting an important point. The syntax of the language plays a large role in intensionality arguments. It supplies the notions of context, predicate, name, etc., which are needed to characterize extensionality. As Russell/Smullyan showed, this can be critical.

This last point deserves some comment and comparison to my earlier remarks on what I labelled Mates' point (p. 18). In considering a language, we might note that there are certain meaning based relaticns or properties, say synonymy and analyticity, that we might incorporate into the semantics. That is, if we thought such properties and relations could or should be theorized about. Suppose we thought otherwise. That is, that facts of reference and truth were the coherent core of our messy pre-theoretic intuitions about meaning. Then we would seek a semantics based only cf such notions. Meaning would only enter in such translation contexts as those condoned by Mates; where the relevant facts of the matter are stipulatory. However, the Fregean point is that this picture comes a cropper - there is no adequate theory of reference and truth (let alone of meaning) that ignores meaning. For, in intensional contexts, the referent of a phrase is its sense. Metalinguistic contexts are brought down into the object language by intensional contexts. If Frege is right we do not have to first independently establish that synonymy, for example, is a pre-theoretic notion of sufficient coherence to be theorized about.

Frege, in developing this kind of argument, neglected to supply a detailed theory of meaning. His purposes centered on mathematics, for which, he believed, only the referential portion is needed. Russell and Whitehead ([R&W1910], p. 8) wrote: "[M]athematics is always concerned with extensions

rather than intensions." This has remained the accepted and entrenched view.

This view has been supportable on a number of grounds. The language of mathematics has *prima facie* fewer kinds of constructions than the rest of natural language. Adverbs and tense are absent; and, seemingly, intensional contexts. Moreover, many problems of logical form were solved early for mathematics - by Frege and Russell, *inter alia*.

Furthermore, the Hilbert school, along with progress in recursion theory, produced a well-understood syntactic theory, seemingly adequate to the intended semantics. First-order languages emerged as the languages into which to regiment. Tarski formalized the semantics for first-order languages, giving what we would now call purely referential semantics, and the picture was complete.

The picture is that of mathematical discourse regimented into a first-order language (the syntactic theory) interpreted by a referential semantics (although Tarski himself might not have put it this way). This is, of course, the picture as it appears to us. Hilbert, the finitist, had little interest in the referential semantics. Conspicuous in their absence are semantic notions like sense, synonymy, analyticity, etc.

The seeds for the breakdown of this picture were planted by Hilbert himself and germinated by Gödel. Hilbert, by inviting the formalization of mathematics and by insisting

that metamathematics is mathematics, established proof theory as a candidate for the mathematical vocabulary. Gödel showed how it could all be done without expanding the ontology; he showed how to represent syntax in arithmetic.

For the First Incompleteness Theorem in the form of an incompleteness result very little information about provability is encoded. Even in the form of a remark that there are true but unprovable sentences of formal arithmetic very little is needed concerning truth -- only that one of a sentence and its denial is true. The proof of the Second Incompleteness Theorem requires that significantly more than correct extension be true of the provability predicate.

I have argued (in sections 2 and 3) that in order to state the Second Theorem (and related results) in an interesting way, an account of them must contain a real (non-punning) notion of expressibility for certain syntactic notions. Two questions naturally arise connecting this need and the accounts of Feferman and Jeroslow and the modal account. Can these accounts be construed as (pieces of) non-referential semantic theories? Are they correct and adequate? As I have urged previously, the latter question is a question distinct from whether the accounts produce acceptable generalizations - purely mathematical results about a large class of objects. The latter is consistent with the accounts given being <u>ad hoc</u> devices of proof theorists designed to "make the proofs go through."

That is, we might, for example, merely see the Feferman account as follows: Formulas extending the axioms are divided into two categories: category I and category II. This clever, but intrinsically meaningless, division is such that all sentences formed in a certain way from the formulas of category I are underivable. This is not true of category II. Moreover, amongst the sentences so formed from category I formulas are sentences we have traditionally called consistency sentences. Finally, stretching, the bounds of this purely non-semantic viewpoint, we recognize that the mode of formation of these sentences from the open formulas syntactically mimics definitions of consistency; where the formulas play the role of the predicate is an axiom. Note that not even this last remark can pretend to establish the whole class of consistency sentences as consistency sentences, since it doesn't distinguish between category I and category II.

I have been at pains to argue that the viewpoint I have just sketched, of the Feferman account as a clever technical device to permit true generalizations over formalisms, cannot yield our usual gloss of the Second Theorem, in which the Second Theorem is taken to be about consistency. In order to justify such a gloss, the first question (Can these accounts, those of Jeroslow and Feferman, be construed as non-referential semantic theories?) must be answered yes. And this makes the second question (Are the accounts correct

and adequate?) of interest. Let us begin with the first question, which will lead us to the second. I approach the first question by indicating what the required semantics would be like, and showing that the F- and J- accounts can be described as being like that.

The language whose semantics is at issue is a stipulated language of elementary proof theory - LEPT. LEPT is an interpreted, informal language, a fragment of natural language. The intended interpretation is the one that we learn when we learn the meaning of such terms as 'consistency', 'formal derivation', 'universal formula', 'variable', etc. Moreover, LEPT's domain consists solely of syntactic objects. (48)

(48) No consistent formal system, T, that is sufficiently strong can prove T's consistency.

is not in LEPT. What I have in mind is that (48) is not in LEPT because what formal systems prove are formulas, and that's as fancy as LEPT gets. However, the semantics for an LEPT sentence like (4) ought to supply us with a useful premise in an argument to (48).

So, LEPT's domain contains strings and sequences of strings, not propositions. A formalization in LEPT would presumably contain enough equipment to do the concatenation theory of those strings and to handle certain inductive definitions. If we were carefully formalizing LEPT we would set up a uniform mode for describing formal systems[42]. Not

surprisingly, Jeroslow and Feferman do just this. It should be noted that many formalizations in LEPT attempt to be ideologically parsimonous; in particular, the notion of set membership is avoided. This observation anticipates an ultimate "reduction" to arithmetic. But this parsimony also tells us that LEPT may be formalized in fairly impoverished set theory, e.g. ZF-Infinity, were we to want to do that. Such issues do not directly concern us here. It should also be noted that, although LEPT is a language, a formalization in LEPT is one that is true to LEPT's interpretation. Without our earlier argument as to the desirability of construing (48) (i.e., giving enough of a theory so that one can establish its truth), concern with LEPT would not on its face involve us with intensional issues. LEPT is an extensionalist's dream - a language purged (by stipulation) of that-nominalizations, like 'that T is consistent', and whose quantifiers range over symbols. The conflation of use and mention is not our road of good intensions. What occurs after 'proves' in LEPT is a name of a sentence, or a variable whose values are sentences.

---

[42] This, and such moves as taking the symbols of formalisms to <u>be</u> numbers, is just to give formalisms an abstract syntax. It is justified by noting that such an abstraction preserves all relevant features - orthography not being relevant. It doesn't matter which mode we pick provided we are able to mirror syntactic relations.

As is well-known, formal systems are often taken to remark on themselves. LEPT is a language for remarking on formal systems. A formalization in LEPT, then, might be taken to remark on formal systems. How do formal systems remark on anything? A sentence of some formal system is provable and that sentence is interpreted as some remark. The sentences of many formal systems have standard interpretations, given in standard manners (see. p 18). Let us consider those formal systems which are of arithmetic. Here we have formal systems remarking on numbers. Note the not surprising fact that such remarking is relative to an interpretation. How does LEPT get into the act? A famous isomorphism is set up: LEPT is reinterpreted. Each member of its domain is effectively assigned a unique number, and the predicates of LEPT are assigned extensions appropriate to the requirements of isomorphism. This isomorphism is godel numbering. Dub this reinterpreted LEPT, numerical LEPT, or NLEPT. The truths of LEPT are truths of NLEPT. NLEPT is about numbers. Hence a formal system with an arithmetic intended interpretation remarks in NLEPT and derivatively (via the godel numbering) in LEPT. If the theorems of the formal system are true then its remarks in LEPT are true. LEPT is a language that talks of formal systems. Thus do formal systems remark about themselves. Put another way, formalisms can be reinterpreted to be "of syntax" rather than "of arithmetic". Our innocuous isomorphism yields lit-

tle of interest. The only relationships preserved will be purely logical ones, just because we have an isomorphism between interpretations which respects the semantics of first-order logic. That is, the requirements of isomorphism are simply that the mapping commute with the logical operators in the Tarski truth definition clauses. More vividly, the isomorphism will ignore features of NLEPT and LEPT peculiar to either one. Sometimes, i.e. for some purposes, these features are important and interesting. In terms of Mates' point, the isomorphism ignores the manner of the interpretation, so that what remark about formalisms a sentence whose interpretation was given in arithmetic vocabulary makes, is mysterious.

Among the important and interesting relations are the entailment relations, which are dependent on the respective (and different) formalizations of P and LEPT[43]. They , of course, are typically not preserved. It can be seen from this that while the translation of LEPT to NLEPT is innocuous if the only purpose to be served is preservation of truth, for purposes of theorizing it is unhelpful. These remarks should be regarded as glossing my invocation of Kreisel on p. 56 ff, concerning fine-grained distinctions of theories.

---

[43]'Different' connotes here that the formalization of the one will not be the isomorphic translate of the other. See below.

What are the correct entailment relations to impose on LEPT? Presumably those plausible on LEPT's interpretation and not NLEPT's. The odds are simply against the translate into LEPT of, e.g. Euclid's Theorem being a vital and interesting fact about formalisms. Note also that this translation cannot be given by the isomorphism alone. Conversely, anyone who has pondered what any of Godel's original forty odd definitions say when expanded into unabbreviated number theoretic language, would hardly take them as foundational truths. The sheer size of the constants is discouraging.

The predicates and objects of LEPT are familar to readers of logic books where definitions of formal derivation, universal formula, etc., abound. Looking at those with an eye to formalization (i.e., inter alia, giving all formalisms a common alphabet) we are led to Jeroslow (if we formalize directly) or Feferman (if we go in NLEPT first and then formalize with an eye on the LEPT interpretation). Neither is a proof in T of nor T is consistent are primitives. In particular, the definition of proof of is inductive and Jeroslow's treatment makes this explicit and critical to the characterization of a canonical provability predicate. As we saw in our discussion of Jeroslow, a method for excluding the extraneous (and deviating) matter from a proof predicate can be derived from principled considerations concerning the standard definitions of formal derivation.

The peculiar virtue of the Jeroslow treatment derives from its direct formalization of LEPT. By avoiding the intervention of an alien subject matter (numbers) in axiomatizing, Jeroslow produces a very weak theory. In other approaches conservative, with respect to number theoretic sentences, extensions of (what turn out to be) stronger arithmetic theories are used. This produces technical virtues for Jeroslow, concerning extensions of the Godel Second Theorem to weak theories, but I have in mind a philosophical virtue. Jeroslow's treatment allows us to regard all of the theorems of his minimal formalization as consequences of formalizing the basic definitions of proof theory. This means that the exclusion of extraneous or deviating matter is based on principled considerations. This feature of the Jeroslow approach enables one to argue that no fact not constitutive of the meaning of basic proof-theoretic terms is included in his formalization of LEPT. The formalization in LEPT is not a theory yielding merely truths in LEPT, but truths arising from the basic definitions. This will play a role in some of my later remarks concerning the construal of the Jeroslow approach as a semantic theory. Explicitly in the Jeroslow approach, and (I shall show) implicit in the others, is the introduction of a third "layer" of language. The proper semantics for LEPT will be couched in some language that mentions the syntactic structures of LEPT; while LEPT itself is a language for talking of formal languages.

Matters get confusing because all these layers are eventually collapsed into the language of arithmetic. It is important to separate the layers so as to see the nature and justification of the embeddings -- which cannot be seen from post-collapse perusal.

Why have I been talking of the proper semantics for LEPT? I insinuated such talk by talking of proper formalizations of LEPT - ones which do more than produce only truths in LEPT. LEPT is a lanugage for talking of formalisms. But we want to talk of predicates of LEPT, such as *is a proof of*, in formalizing LEPT's semantics. Of course, LEPT itself is a candidate for the language of such a semantics. One question is - what are the truths of this semantics? A partial answer is - not all the truths in LEPT. Thus the preceding emphasis on parsimony.

We have, in essence, already seen a major desideratum for a semantics for LEPT. If the Second Theorem is to be stated in LEPT then LEPT's semantics must be intensional; i.e. distinguish between coextensive predicates of LEPT. Our clear separation of LEPT and its semantics is obfuscated, as I have mentioned, by the identity of the languages of each and by the following mode of semantic descent: Itemize the analyticities of LEPT by a weak formalization in LEPT.

From these observations we may distill two senses in which formalisms may be said to talk about themselves;

senses which have typically been amalgamated, as one of them plays no manifest role in the First Theorem. One sense is the mode adumbrated above, which exploits the Gödel numbering isomorphism. The other sense is based on the existence of a semantics for LEPT. I shall detail each of these two senses so as to clarify their distinction. Let us dub the first, the isomorphism sense; and the second, the semantic sense.

Although the isomorphism sense is described above, a redescription will be helpful. The previcus description indicated how formalisms can be said to remark on themselves; namely <u>via</u> a purely extensional translation (preserving first-order logical relations) into LEPT. Now we may view this isomorphism as supplying a semantics for LEPT. Since the isomorphism respects only extensicns, what this semantics says LEPT says is sensitive only to standard arithmetic truth conditions on which it is parasitic; not even on their "standard manner" of presentation. So the implicit isomorphism semantics counts deviant and non-deviant consistency sentences as making the same remark. This is merely to cast into different terminology some of the earliest observations of this thesis.

The semantic sense does not give LEPT a semantics parasitic on an arithmetic interpretation. Given a formalization in LEPT and a semantics for it true to the intended interpretation, then we have this semantics giving sense to

the claim that formalisms remark on themselves. What is an example of such a semantics? Well, I take it to be obvious at this juncture that I have three examples in mind -- call them Jeroslow-semantics, Feferman-semantics and Modal-semantics. Furthermore, these differ from the isomorphism semantics in being non-extensional.

A semantics for LEPT is to be given in a very restricted theory, though one adequate for the purpose. The semantic task is easier here than for English not because the semantic complexity is less (though it is), but because we have a clear grasp of what the primitives and correct definitions are. It turns out, moreover, that entailment relaticns are the primary emphasis of this semantic theory. The data would include such home truths as that A's being a theorem entails A's being a sentence; as well as others that follow from the basic notions of proof theory. Given this synoptic view, I am going to indicate how each of the three treatments sketched above fits it. I shall also use it to answer some questions that have been raised, explicitly or implicitly, in some of the literature surrounding the Second Theorem; and to account, on an other than analogical or intuitive basis, for certain semantic-like remarks in the literature.

The distinction just made between the isomorphism semantics and a semantic semantics can be seen from another point of view. The conflation of the two senses in which formal-

ism remark about themselves is abetted by the transparency of the First Theorem. At the level of sentences, extensionality promotes equality of truths; whereas, the Second Theorem requires the separation of the analytic from the merely true. If we do not ascend from LEPT the definitional truths will be indistinguishable from other truths; definitonal coextensiveness will be indistinguishable from mere coextensiveness. If we need to make such distinctions (and I have been arguing that consideration of the Second Theorem and related results indicate such a need) we end theorizing about LEPT, not speaking LEPT. Attributions of analyticity to sentences of LEPT should not be confused with (assertions of) the analyticities themselves. This third "layer of language", the language of a semantics for LEPT, mentions constituents of LEPT as well as use them[**]. This overview can be anchored by some pertinent particular considerations: The "third layer" is the common thread I shall unravel from each of the three treatments.

[**]It struck me, after sorting this out, that Kreisel deserves some credit for seeing the point first. I think. At any rate the remark of Kreisel's had always puzzled me: "Thus Gödel's Second Theorem would be stated: If a system S is consistent and a formula can be <u>proved</u> <u>in</u> S to express the consistency of S, then A cannot be proved in S." The italicized occurence of 'proved' was the puzzle. It is, we now see, something of a red herring. It really should read 'is an assertion of LEPT's semantics.' When all three layers are arithmetized, such distinctions are lost. Only a few extra turnstiles remain to remind us of their passing.

LEPT is an interpreted language and theories may be couched in LEPT. Some theories in LEPT are particular or narrow, and certain truths in them will not be truths of all theories couched in LEPT. Some propositions of LEPT do follow from the meanings of the lexicon of LEPT and have the status of analytic truths of LEPT. Some examples have been given, and an interesting case is discussed on p. 121. A semantics for LEPT will be a theory that assigns entities to structural descriptions of LEPT sentences and their parts, in such a way as to distinguish the class of analytic sentences, to mark off certain entailments, etc. I shall now describe the three treatments as semantic theories. (see also p. &&, earlier.)

Feferman's treatment regiments LEPT as the "dotted" language. In this language are terms that are to be interpreted as referring to syntactic entities. The crucial feature of Feferman's semantics is the presence of the term α, used to refer to open formulas of arithmetic. The predicates, PR-formula and RE-formula are predicates of such open formulas. The semantic theory for this regimented LEPT includes a standard referential semantics (induced by the "identification" of syntactic objects with numbers, and the notions of numeration and binumeration) augmented by the PR and RE predicates. Crucially, two α's of identical extension need not both be RE, or PR. In this dotted language coextensive substitution is not truth preserving; this is Feferman's Thm. 5.9.

Feferman's semantic theory, containing such terms as 'binumeration', 'RE-formula', etc., is a fairly abstract semantics. He indicates, however, the obvious natural application of it. The α's are to be thought of as rules for selecting out axioms, and RE rules are ones that characetize formal systems. More precisely, the dotted proof predicates, which are associated with their inductive definitions involving an occurence of (defn. (4.1), p. 88), are identified with formal systems.

The various modal approaches, whether Löb's, or the highly developed one that utilizes G, can be thought of as producing the relevant analyticities of LEPT directly[45]. LEPT is once again regimented -- into a language including terms for syntactic objects and the predicate Bew. The modal approach mentions sentences containing Bew, and '$\vdash_G$' is readable as 'is an analytic truth of LEPT'. Since this semantics is merged with the standard one that assigns extensions to predicates like Bew, we get the familar lack of extensionality. Of course, <u>via</u> the Solovay result, we get the semantics reduced into arithmetic, and post-reduction, all the turnstiles look alike ('$\vdash_T$'). This reduction, and corresponding reductions to a common arithmetic theory in the other approaches, are hindrances to seeing the layers separately.

---

[45]This remark needs qualification, which is supplied a few pages hence.

As remarked previously, Jeroslow's treatment contains what we are now construing as a semantics for LEPT. The definitons of proof theory are given in a language that contains such predicates as 'SpTm', 'Thm', etc. Canonical proof predicates are defined by again recurring to the inductive characterization of formal proof. And formal systems are Post Canonical Systems. This identification plays the role here that Feferman's use of open formulas, interpreted as giving the rules for picking out the axioms, did in the F-semantics. Many PCSs have identical sets of productions.

A first glance at the necessary conditions on provability predicates often puzzles people. Why should, for example, it have to be provable in arithmetic that provability predicate, B, be such that $B(A \rightarrow C) \rightarrow (B(A) \rightarrow B(C))$? Why not just true? What is revealed by the above discussion is that the answer is that such conditions are not required to be provable in arithmetic. They are just assertions of a semantics for LEPT. Translated into the language of arithmetic they become theorems; this accounts for the "extra" turnstiles. That is, in all these cases, there is a third theory whose formalization is represented by those "extra" turnstiles - and whose appearance as turnstiles of some arithmetic theory is simply a result of the post hoc translation.

This last point has been mentioned previously, in connection with a remark of Kreisel's. My explanation of the "extra-turnstiles" has a coincidentalist ring. That is, I have argued that a semantics, and an intensional semantics at that, is needed to explicate Gödel's Second Theorem. Since both the objects being interpreted and their interpretations (the objects of the semantics) are syntactic objects, the semantic theory can be couched in the language of LEPT and formalized; furthermore, it is a weak theory easily arithmetized.

That LEPT's semantics <u>can</u> be formalized, and with equipment already to hand in any arithmetization endeavour, may explain why it is always presented <u>as</u> formalized. Kreisel only seems to be asserting that it must be formalized and that it must be formalized (<u>via</u> arithmetization) in the theory whose consistency sentence is at issue. Feferman shows that it is a sufficient condition that this be shown; i.e., that, as Kreisel sums it up, "a formula can be proved in S to express the consistency of S". I have been attempting to show that this is true, but a "coincidence".

If taken seriously as a necessary condition, it would mean that a theory too weak to formalize EPT's semantics, that is, the language of its own proof theory, doesn't have any consistency sentences. The temptation to say this arises because such a theory is too weak to formalize the First Theorem, and hence, too weak to have the Second Theorem

proved for it. Why this is so is the proper reason for asserting that weak theories can't assert their consistency. In proving the Second Theorem the formalism for which the Second Theorem is being proved is required to "know" (prove) that a formula expresses provability (see (J20)). It can do this if it is K-complete. K-completeness is, roughly, a requirement that the formalism contain its proof theory's semantics.

This doesn't take us all the way. There is seemingly the option of saying that weak theories have consistency sentences, possibly provable ones. Such theories would have to be so weak that they couldn't discern their consistency sentences from deviant "consistency" sentences. I think that there are persuasive reasons for discarding this option, and this last fact is one such reason. Although I have avoided talking about the epistemological content of the Gödel Theorems, I will say that this situation, a formalism that proves consistency but doesn't know it, seems not to have the epistemological content that a Hilbertian might have had in mind.

An other problematic aspect arises because none of the truths of mathematics, meta- or otherwise, are contingent. Many of our semantic intuitions concerning intensional contexts and analyticity connect to ones concerning necessity and contingency. These intuitions have recently been schooled by the development of possible world semantics.

Unfortunately their bearing on the semantics of mathematical discourse, particularly the cases I have been describing, is dubious; particularly if the possible world semantics is taken "literally", as Stalnaker, for example, takes it.

In the cases under consideration this aspect presents itself as the problem of distinguishing between the analytic truths of LEPT and the non-analytic ones- which we cannot perforce call 'the merely contingent ones'. Many truths in the language of LEPT follow from any precise identification of the formalisms involved. Many of these I have implicitly ruled out as analyticities; that P is consistent, for instance, or that Q cannot prove ccmmutativity of addition and P can. The latter fact follows inexorably from the standard specifications of Q and P. The former fact, that P is consistent, also fcllows from the specification of P. For the moment let us assume that this fact, that P is consistent, can be justifiably ruled a non-analytic truth of LEPT. Then we can see how to describe the distinctiсn between canonical and "contingent" (that is, non-analytic) reference. Recalling that the deviant proof predicate could be described as extending the theorems of the intended formalism on hypothesis of consistency, we can see how the extension of such a predicate depends on a non-analytic fact.

This only shows how we might draw such a distinction, for some cases, and how it would fit in with our previous

observations. It remains to show how it is implemented and why where the line gets drawn is appropriate. Our immediately previous observations are relevant. For, the distinction is really between truths in LEPT and truths about LEPT. And, if this is right, it is a clear case of the separation of metaphysical modal notions from semantic ones.

What is the proper semantic theory for LEPT? Well, it depends whose account one takes as gospel. It is most explicit in the Jeroslow account, but on all accounts it is never stronger than P. So, amongst other things, Gödel proved that that P is consistent is not an analytic truth of LEPT; i.e., is not a semantic fact.

As I suggested earlier, the modal account would seem to plump for G as the theory of analyticity and entailment for LEPT. It would be tempting to read '$\vdash_G$' as 'is an analyticity of LEPT'. Unfortunately G seems too strong. Many of its theorems seem to be what I have been calling non-analytic truths of LEPT. $\Box p \rightarrow \Box\Box p$ is a difficult case. Counterexamples to this would involve very weak systems- and we might want to say that by 'formal provability' we (LEPT) mean(s) sufficiently strong formal provability. The temptation is reinforced by the weakness of G's deductive apparatus. G is not a first-order quantifier theory with induction; such a theory might cause us to worry that more than the semantic facts were being represented. I was lauding the Jeroslow approach for its avoidance of this problem.

With G's weak deductive apparatus (it's a sentential logic), if we can justify the axioms semantically, we are a long way toward justifying G as a semantics. Since G is weak, the rule of necessitation is harmless if we assume, as above, that formal provability means sufficiently strong formal provability.

Jeroslow's theory, J, is very weak; but all the accounts are adequate to the semantic data examined. That is, they respect the cavils of the early sections. What justifies restriction to such weak theories, other than methodological parsimony? There are two sorts of justification - one epistemological, the other semantical. Jeroslow essentially alludes to the former in the passage quoted on p. 71f. It seeks to salvage the epistemological core of the Hilbert program. The semantic theory is a finitist theory about (the concept of) provability. As we have seen just above, Godel proved that the consistency of P is not provable in such a finitist semantics. The semantic justification is simply the observation that the analytic truths are just those that follow from the meanings of the predicates involved - the meanings are given by the inductive definitions; and anything in addition to inductive ability on the relevant predicates is to go beyond the minimal core of meaning-based truths. Note that we have been more than once driven to put lower bounds on the strength of the semantic theory.

Curiously enough these sketches of two sorts of justification for a semantic theory yield yet another deep parallel to issues in natural language semantics. Some writers have claimed, implicitly or explicitly, that the adequacy of a semantic theory is to be judged solely on whether the correct semantic properties and relations are predicted. Others place restrictions on how a semantic theory produces its semantic descriptions - the restrictions based on demanding that the theory be "psychologically real". The parallel to the Hilbert demand is close. In the next and final section I will have more to say on this and on other connections to issues in natural language semantics.

Chapter 11

POSTLIMINARY CAVILS

LEPT is a technical fragment of English. LEPT requires, and has, an intensional semantics. So English requires an intensional semantics. This argument is perhaps reminiscent of arguments for the extensional insufficiency of context-free grammars for English from their extensional insufficiency for some fragment. These are in turn perhaps reminiscent of an argument that some set is non-recursive because because some subset of it is. Since the last argument is invalid, one might suspect that the others will, at best, need additional premises to make them respectable.

This has been done for the case of English grammars in ways that preserve the spirit of the original argument. Roughly speaking, the strategy involved looks like this: If one can argue that the fragment is "independent", in the sense that a semantics for the whole yields a semantics for the fragment by mere restriction, then the argument is bolstered. This is done by defining an equivalence relation between grammars stricter than <u>same productions</u>, thereby strengthening accompanying notions of adequacy. If a grammar for the whole assigns certain structures to sentences, then it assigns the same structures to sentences that reoccur in the fragment. Then it can be argued that those

structures are incorrect for the fragment, and so, incorrect for the whole.

Without going into more detail, my point is simply that the argument represented by the first two sentences of this section, it might be hoped, could be rectified in a similar way. It certainly does seem clear that LEPT is a well-defined isolatable fragment; and that a semantics for all of English should contain a separable semantics for it. Analogies with certain intensionalizing moves in natural language semantics are striking. The parallel between strong equivalence of grammars and individuation by PCSs being only one case. While the project of filling in the argument seems plausible, I have not worked out the precise lacunae to be supplied.

There is a less dubitable, but weaker, conclusion, supported by the preceding sections. There is nothing as convincing as a counterexample. Demonstrating a logical lapse in someone's argument is not as strong as giving a counterexample to the conclusion. The counterexample should not, of course, merely shift the subject of dispute to the coherence of the counterexample.

My claim is that I have found two, maybe three, counterexamples to the claim that intensional semantic theories are incoherent. Now maybe nobody ever really claimed that; but if anyone did, it was Quine. I have in mind a strong version of this attribution of incoherence. Namely, that inten-

sional semantics are intrinsically incoherent. This shouldn't be confused with the claim that a pure intensional semantics is somehow an incoherent piece of mathematics. Kripke cleared that much up on the modal front. Here, though, we have an applied intensional semantics (innocently constructed, by the way, to solve purely technical prcblems) that cannot be faulted on ground of rigor nor on the incoherence or murkiness of the entities and relations appealed to; they are finitary syntactic objects and relations among them - a Quinean ideal.

I would like to briefly take up some remarks I made toward the end of the previous section concerning psychological reality. It has begun to strike many people working natural language semantics that many available semantic theories are not candidates for competence models of human linguistic behavior. Indeed, some people who only believe in performance or competence theories have argued to the incorrectness of semantic theories committed to infinitary objects. I shall consider the issue as one concerning the method of implementation, in a finitary mode of representation, of an infinitary semantics.

Fortunately, to make this point vivid, one need not mount a careful destruction of arguments against psychologically unreal semantics - counterexamples are at hand. One such I will only allude to. Workers in computer science have described semantics, for computer languages, that are

committed to infinitary objects. They have, moreover, no trouble in talking of successful implementations ("performance models") of such languages, nor any trouble in dealing with algorithms to realize the semantics ("competence models"). Indeed, the need for a notion of sucessful implementation prompted the creation of computer language semantics. These implemenations are, of course, on those paradigms of finitude, computers.

The usual semantics for first-order arithmetic is non-computational. Yet it is not usually argued that it is an incorrect semantics on the grounds that we dcn't compute the references and truth-values in the manner of the semantics. This is so because no one takes the semantics as a performance or competence model. (Except, I guess, intuitionists and some finitists.)

Finitists, in producing primitive recursive arithmetic and similar systems, have tried to give competence models of human mathematical behavior. Of course they were motivated by epistemological concerns: only such knowledge as could be directly represented in our finitary heads would be knowledge. Hilbert, we might say, regarded finitist systems as representations of knowledge. We could have putatively infinitary knowledge to the extent to which such a system adequately represented such knowledge; and if we could show the system consistent. The problem of adequate, or correct, representation I shall skip, although it is a key issue.

I should like, in the light of distinctions made in this thesis, to suggest a certain delineation of a Hilbertian program as sketched above. Proof theory is the theory in which Hilbertian justifications are carried out. So proof theory must be finitary. The semantics for the language of proof theory, the semantics of LEPT, is a weak theory. Yet it does not follow from the semantics of LEPT that any, sufficiently strong, system is consistent. So even if our semantic knowledge constituted a secure epistemological base, assertions of consistency exceed it. Thus, while mathematical knowledge may be necessary, it isn't analytic. I take this to be a form of realism.

I hope I have established what I said I would. I feel sure that I have a least shown that the connective theses I have uncovered do not have, as I remarked on p. 14, "mathematics' clarion certainty and precision."

Appendix A

BIBLIOGRAPHY

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

[Boo2os]
Boolos, G. "The Iterative Conception of Sets", Journal of Philosophy

[Brentano]
Brentano, . "The Distinction Between Mental and Physical Phenomena", in R. Chisholm, ed. Realism and the Background of Phenomenology, Glencoe (1960)

[B&J]
Boolos, G. and Jeffrey, R. Computability and Logic, Cambridge (1974)

[Fef]
Feferman, S. "Arithmetization of Metamathematics in a General Setting", Fundamenta Mathematicae, XLIX (1960)

[Fef2]
Feferman, S. "Transfinite Recursive Progressions of Axiomatic Theories", Journal of Symbolic Logic, XXVII (1962)

[H-B]
Hilbert, D. and Bernays P. Die Grundlagen der Mathematik Berlin (1939)

[Jer1]
Jeroslow, R. "On Godel's Consistency Theorem", unpublished manuscript

[Jer2]
Jeroslow, R. "On the Encodings Used in the Arithmetization of Metamathematics", unpublished manuscript

[Jer3]
Jeroslow, R. "Consistency Statements in Formal Theories", Fundamenta Mathematicae, LXXII

[Kreisel]
Kreisel, G. "Mathematical Logic" in T.L. Saaty, ed. Lectures on Modern Mathematics, vol III, New York, London and Sydney (1965)

[Krsl]
Kreisel, G. Review of [Fef], 4913 in Mathematical Reviews

[Lob]
Löb, M.H. "Solution of a Problem of Leon Henkin", Journal of Symbolic Logic, XX (1955)

[Montague]
Montague, R. "Syntactical Treatments of Modality ..." in R. Montague Formal Philosophy, New Haven (1974)

[TaMoRo]
Tarski, A., Mostowski, A. and Robinson, R. Undecidable Theories, Amsterdam (1960)

[Mostowski]
Mostowski, A. Sentences Undecidable in Formalized Arithmetic, Amsterdam (1964)

[Mates]
Mates, B. Elementary Logic, second edition, New York (1972)

[Turing]
Turing, A. "Systems of Logics Based on Ordinals" in M. Davis, ed. The Undecidable, Raven Press (1965)

[Webb]
Webb, J. "Gödel's Theorems and Church's Thesis: A Prologue to Mechanism" unpublished manuscript