How to Say Things with Numbers and Be (Provably) Consistent About It

Alexei Angelides

1 Introduction

In 1931 a young mathematician named Kurt Gödel published results that would transform fledging branches of pure mathematics & redefine the way the public has interacted with the mathematical sciences ever since. Called the Incompleteness Theorems, plural because there are exactly two, many formidable minds have waxed poetic & philosophical about their scope and implications. Some have argued that they can be taken to show that the mind cannot be modeled by a machine. Others have said that Gödel's results show that only God can know all mathematical truths. Even Hofstadter, author of bestselling *Gödel, Escher, Bach* (1979), has said that the Incompleteness Theorems lie at the core of AI. While Goldstein, author of an entire biography entitled *Incompleteness: The Proof and Paradox of Kurt Gödel* (2005), has said that the Gödel theorems address "the central question of the humanities: what is involved in our being human?" In 2020 Wired Magazine published an article on Gödel's theorems that suggests that "Gödelian incompleteness afflicts not just math but – in some ill-understood way – reality." Such ways of thinking about two theorems in a subfield of pure mathematics are inspired, but misconceived. So what do the Incompleteness Theorems say and what exactly is supposed to be incomplete?

1.1 Setting the Stage

Whatever we might think of inspired inferences like the ones above, one thing is certain: the Incompleteness Theorems took direct aim at Hilbert's

Program (HP). In the 30 or so years leading up to Gödel's Incompleteness Theorems, mathematician David Hilbert had proposed to extend the successful results for the completeness of first-order *logic* to mathematics. The Completeness Theorems, as they apply to formal first-order logic, are that the language of first-order logic is sound and complete. For a language to be *sound* means that if a sentence has a proof in that language, then that sentence is true in all possible models of the language. For the language to be *complete* means that if a sentence is true in all models of that language, then it has a proof in that language. More precisely, let's use the language of first-order logic. Suppose the following symbols are defined as follows:

$$\neg := \textit{not} \quad \vee := \textit{or} \quad \wedge := \textit{and}$$

The Completeness Theorem for a first-order logical language $\mathcal{L}$ says that, for example, for any arbitrary sentence p:

$$\text{if } p \vee \neg p \text{ is true in all models of } \mathcal{L}$$
$$\text{then } p \vee \neg p \text{ is a theorem of } \mathcal{L}.$$

That's to say that all logical truths are theorems of $\mathcal{L}$ and hence, $\mathcal{L}$ is complete. Hilbert had proposed to extend those results to the language of arithmetic, and then, if they'd been successful at that, extend them further to analysis and beyond. This proposal came to be known as Hilbert's Program. To carry (HP) out, it would mean that for every sentence in the language of arithmetic and analysis – from easy things like $0 = 0$ and $0 = 1$ to harder things like the Intermediate Value Theorem – there is a proof of it, or its negation, in the language. Especially important to Hilbert's Program was to prove that there is no proof in the language of arithmetic of a sentence *and* its negation, a contradiction of the form $p \wedge \neg p$. That's to say that one of the main aims of (HP) was to use a mathematical proof to show that mathematics is *consistent*. Alas, it is said, it was not to be.

Upon publishing his two theorems, a consensus in the 1930s emerged, that persists today, that they implied that carrying out (HP) was simply not possible. Gödel's First Incompleteness Theorem (G1) states that for any recursively enumerable formal theory T that axiomatizes a sufficient amount of arithmetic there is a formula ϕ in the language of T that is true in its model but neither provable nor refutable using T's axioms and inference rules. Gödel's Second Incompleteness Theorem (G2) states that it is possible to produce for ϕ a formalization of the assertion that T is consistent such that the formula is true in its model, but again neither provable nor refutable in T. In contrast to (G1), (G2) concerns an inherently self-referential statement that is central to the goal of proving the consistency of classical arithmetic. In the tradition following Gödel's first and second incompleteness theorems, with few exceptions philosophers, logicians, and mathematicians have claimed that (G2) entails that for any T satisfying the above conditions T does not prove its own consistency, or that, similarly, no consistency proof of T is formalizable in T. Though the theorem in Gödel (1931) is somewhat different, for the moment, let's refer to (G2) as:

$$\text{(G2)} \qquad T \nvdash Con(T)$$

and distinguish it from the claims:

1. that T does not prove its own consistency; and
2. that no consistency proof of T is formalizable in T.[1]

One task is to understand if (G2) justifies claims (1) or (2). Once we do, we are in a better position to see how (G2) affects (HP).

1.2 Outline & Goals

Our plan is as follows. In section (2) we'll look at the method of arithmetization. Arithmetization is at the heart of (G1) and (G2) and forms the basis for getting a formal mathematical theory T to say of itself that it is consistent, in whatever ways it might. In section (3) we look at claims (1) and (2) above as interpretations of (G2). It turns out that on an extensional interpretation of (G2), (1) and (2) do not follow. But on an intensional interpretation, claims (1) and (2) are true. However, intensional interpretations

[1] Franks (2009) calls (1) and (2) the "Gödelian inferences" because in (1931) Gödel himself was the first to draw them in such a way. But since most logicians & mathematicians post-1931 make these inferences, we'll just refer to them as above.

raise the issue of the correctness of the consistency formula because on such an interpretation, T must recognize the consistency formula as *the* formula that expresses its own consistency. Our question then concerns the correct conditions under which T recognizes its own consistency. In section (4) we show that there is a formalism that is able to recognize that its provable sentences form a subclass of the sentences true of each complete extension of itself, and that in such a formalism Gödel's Second Incompleteness Theorem does not follow. But this leads to a natural research problem: find the mathematically weakest formal theories that fail to recognize properties that are true of complete extensions of itself, that is, the weakest theories for which Gödel's Second Incompleteness Theorem is true. Because, if we can find such theories, then *eo ipso* those theories are upper bounds of theories for which (G2) is *not* true.

2 How to Get Numbers to Say Things

At its heart, arithmetization is a coding of the syntactical symbols of a formal language $\mathcal{L}_T$ of a formal mathematical system T using simple, tractable theorems from number theory to assign numbers to the sentences of T. So, an arithmetization encodes sentences *about* a mathematical theory as sentences *within* that mathematical theory. Take for example, the sentence:

there is a proof in number theory
of the fundamental theorem of arithmetic.

This is a sentence about number theory, about the existence of proofs in number theory. An arithmetization of that sentence would produce a true theorem of number theory, say $2 + 2 = 4$, that encodes the sentence above by assigning each syntactical element in the sentence to numbers in number theory. Hence, an encoding is a mapping that enables expressions about a formal theory to be embedded into the formal theory itself, so that after application of the procedure, sentences of the formal theory assert that certain properties of natural numbers hold of sets of natural numbers, and these sentences make indirect assertions about certain properties of formulae holding of sets and sequences of expressions. It's in this sense that an arithmetization can get numbers to say things.

Let's see how by looking at an example. Consider the table below.

symbol	♯ assignment	meaning
$\neg$	1	not / negation
$\wedge$	3	and
$\rightarrow$	5	if . . . then . . .
$\exists$	7	there exists
(	9	left parenthesis
)	11	right parenthesis
v_i	$i + 13$	variable
$\mathbf{0}$	2	zero
$\mathbf{S}$	4	successor
$=$	6	equals
$+$	8	addition

The table can be seen as a representation of a kind of mapping of logical and arithmetical symbols in the language to numbers. Now let's take a sentence in the language of formal arithmetic:

$$\neg\exists v_0(\mathbf{S}v_0 = \mathbf{0})$$

This sentence says:

it's not the case that there exists a number v_0 such
that 0 is the successor of that number.

Or, in more familiar terms, zero is not the successor of any number. Our table lets us map each one of the syntactical elements in the formal sentence above to a number so that we get the following sequence:

$$< 1, 7, 13, 9, 2, 13, 4, 2, 11 >$$

We could just take the sum of this sequence – 62 – as the sentence's number but then it would not uniquely represent the sentence because there are infinitely many ways to decompose 60, such as $20 + 20 + 13 + 7 + 2$. And that way of decomposing 60 represents the ill-formed, nonsensical string:

$$v_7 v_7 v_0 \exists \mathbf{0}$$

In order to make sure that the number is unique, Gödel's main insight was to use the Fundamental Theorem of Arithmetic. That theorem, first proved in Euclid's *Elements*, says that every integer greater than one can be represented as the unique product of primes. Hence, for example, 264 has a unique prime factorization as:

$$(2 \times 2 \times 2) \times 3 \times 11$$

which is just:

$$2^3 \times 3^1 \times 11^1$$

So let's use this basic fact from arithmetic and map our sequence into exponents of products of primes:

$$2^1 \times 3^7 \times 5^{13} \times 7^9 \times 11^2 \times 13^{13} \times 17^4 \times 19^2 \times 23^{11}$$

This is then the unique number associated with the sentence above, and is called its Gödel Number, denoted as:

$$G\sharp(\neg\exists v_0(\mathbf{S}v_0 = \mathbf{0}))$$

or simply as the product itself.[2] In Gödel (1931) there are of course some differences, and Gödel himself only offered sketches. It was sharpened by successive mathematicians in the years following the publication of his two theorems, but the essence of arithmetization is captured by the above.

One of the main consequences of this method is that not only can it be

[2] More explicitly, f is a function mapping each symbol of a first-order language $\mathcal{L}_T$ of a formal arithmetic T to a specific natural number by partitioning the symbols of T into predicate parameters and logical symbols, partitioning the natural numbers into even and odd numbers, and defining f such that f maps parameters to even numbers and logical symbols to odd numbers. Then a Gödel Number of a sentence in $\mathcal{L}_T$ is the value of a function g that maps sequences of f to exponents of products of primes.

used to encode theorems of arithmetic and analysis in $\mathcal{L}_T$ but it can also encode sentences in $\mathcal{L}_T$ *about* theorems and axioms and proofs in T. Arithmetization gets numbers to say things about other numbers in arithmetic and analysis and it gets numbers to say things about proofs, theorems, and axioms in arithmetic and analysis. For example, taking our example above, we can use arithmetization to say:

> the formula with Gödel Number $2^1 \times 3^7 \times 5^{13} \times 7^9 \times 11^2 \times 13^{13} \times 17^4 \times 19^2 \times 23^{11}$ begins with negation.

Since negation was assigned the number 1 in the table, that sentence can be understood as a sentence about how the prime factorization of the number above begins with 2^1, namely that:

> there is some number v_0 such that $v_0 \times 2 = 2^1 \times 3^7 \times 5^{13} \times 7^9 \times 11^2 \times 13^{13} \times 17^4 \times 19^2 \times 23^{11}$ and there is no integer v_1 such that $v_1 \times 4 = 2^1 \times 3^7 \times 5^{13} \times 7^9 \times 11^2 \times 13^{13} \times 17^4 \times 19^2 \times 23^{11}$.

Hence, the artihmetization allows us to prove, in T, that the large number above has 2 as a factor but not 4 which, when decoded, is a true sentence about the syntax of a formula in T. It's in this specific way that arithmetization allows us to convert sentences about proofs and theorems of T into sentences about prime factorizations.

Most importantly, we can consider predicates such as:

x is the Gödel Number of a proof in T
of the formula with Gödel Number y

which is translated into an arithmetical predicate in T of the form:

$$Prf_T(x, y)$$

Let n be the Gödel Number of that predicate. Now consider the predicate:

z is the Gödel Number of a proof in T
of the formula with Gödel Number n

that is, an arithmetical predicate in T:

$$Prf_T(z, n)$$

But now, the proof in T refers to itself, and it is this basic insight that Gödel uses to prove (G1) and (G2). Consider, for example, the predicate:

$$\neg \exists x Prf_T(x, y)$$

This says that it's not the case that there is a proof in T with $G\natural(x)$ of the formula with $G\natural(y)$. Let $G\natural(n)$ be the Gödel Number of that formula. Now substitute n for y in the first formula:

$$\neg \exists x Prf_T(x, n)$$

and call the Gödel Number of this formula $G\natural(g)$. What does this formula say? It says that there is no proof in T of the formula (itself) with no proof in T. In short it says of itself that it is not provable in T. By the construction of the sentence it is true in T. But, of course, because what it says is true, it follows that $G\natural(g)$ is not provable in T.

This is the crux of how (G1) works. We've produced a true but unprovable sentence in T. Note that the arithmetization procedure for (G1) only relies on the formula being "numerically correct," meaning that the Gödel Numbers assigned at each step must be unique, and that each formula must express some unique property of a Gödel Number's prime factorization. But what about when we turn to (G2)? In many ways (G2) appears to follow directly from (G1). The consistency of T is expressed by the sentence:

$$Con(T) := \neg \exists x Prf_T(x, \perp)$$

which says that it's not the case that there exists a proof x in T of a contradiction, where $\perp$ means a formula of the form $p \wedge \neg p$. Then, let $G\natural(Con(T))$ be the Gödel Number for the Gödel sentence that says that there is no proof of $Con(T)$ in T. By the construction, what it says is true, but because of what it says, it follows that there is no proof of $Con(T)$ in T. QED it seems.

3 Intension & Extension

What I want to turn to now is claims (1) and (2) above. For although the inference from (G1) to (G2) seems direct, (G2) is much more sensitive than (G1) to how an arithmetization is done. The basic problem is as follows. Feferman (1960) claims that, unlike (G1), interpreters of (G2) are confronted with a unique problem. Now (G1) states that for any formal system T, there is a formula true in a model of T but not provable using T's axioms and rules of inference. If formalized this statement only expresses the fact that *some* sentence is true but formally underivable in T, but does not tell us which sentence, nor if T need recognize the sentence that is formally underivable. Insofar as T contains some primitive recursive arithmetic, then it is possible to recursively arithmetize a sentence for T that satisfies (G1). But for the proof of (G2) T must recognize that the sentence that is formally underivable is a sentence that expresses, or "says" that T is consistent. Feferman writes:

> [i]n broad terms, the applications of the method [of arithmetization] can be classified as being *extensional* if essentially only numerically correct definitions are needed, or *intensional* if the definitions must more fully *express* the notions involved, so that various of the general properties of these notions can be formally derived. (Feferman (1960), 35)

For extensional examples Feferman lists (G1), the undefinability of truth, and the undecidability results for various theories. For examples of the intensional type he lists (G2), results of relative consistency strengths between theories, and logics for ordinal analysis. For results of the first type one need only know that arithmetization of the metamathematical concept picks out a unique numerical class. But for results of the second type one must know that the numerical class picked out correctly expresses the metamathematical concept. Since our analysis of this distinction will aid us in understanding how, if at all, claims (1) and (2) are consequences of interpretations of (G2), in what follows we sharpen it.[3]

[3] By "picking out a numerical class" is meant either bi-numerate or numerate. See Feferman (1960), 51.

3.1 Intension and Extension

In order to sharpen up the distinction between extensional and intensional applications of arithmetization, let's first consider a few concrete examples. Gödel (1931) states the theorem for the ω-consistency of a formal system. For the system that Gödel studied, *Principia Mathematica* (PM), it is said to be ω-consistent just in case when (PM) proves $\phi(n)$ for all n, where ϕ is an arithmetical formula, then $\phi(n)$ is true in the standard model of arithmetic for each n. After Gödel, Rosser (1936) strengthened Gödel's result and along the way found a predicate distinct from the one that expresses the proof relation "x is a proof in T of y" for a theory T to obtain extensionally improved theorems concerning provability in metamathematics. Let:

$$Prf_T(x, y)$$

be a numerically correct definition of the proof relation for T. Then the Rosser predicate is the relation $Prf_T^R(x, y)$ defined by:

$$Prf_T(x, y) \wedge \neg\exists z\, [z \leq x \wedge Prf_T(z, neg(y))].$$

Under the assumption that T is consistent, then $Prf_T^R(x, y)$ and $Prf_T(x, y)$ are extensionally equivalent. But the Rosser proof predicate says that x is a proof of y in T such that there are no proofs z in T shorter than x that prove the negation of y, and therefore fails to express the proof relation in T, which is instead expressed by $Prf_T(x, y)$. Hence, while $Prf_T^R(x, y)$ and $Prf_T(x, y)$ are extensionally equivalent, they are not intensionally equivalent. Here is, then, an example in which it is possible to use predicates with "non-standard" intensions in order to obtain extensional improvements (in the sense of a wider scope) upon existing theorems.

In a similar direction, let T be a formal system containing some arithmetic, let $Prf_T(x, y)$ be a numerically correct definition of the proof relation as above, and consider the relation $Prf_T^C(x, y)$ defined by:

$$\Big(Prf_T(x, y)$$
$$\wedge$$
$$\neg\exists u \exists w \exists z \big((u \leq x \wedge w \leq x \wedge z \leq x)$$

$$\wedge$$
$$(Prf_T(u, z) \wedge Prf_T(w, neg(z))))\Big).$$

Again if T is consistent, then $Prf_T^C(x, y)$ is extensionally equivalent to $Prf_T(x, y)$. If we now define a consistency predicate, $Con^C(T)$, as:

$$\neg\exists x\exists y\exists z\big(Prf_T^C(x, z) \wedge Prf_T^C(y, neg(z))\big)$$

then, since the above is an instance of a first-order validity, it is possible to prove $Con^C(T)$ in T. But does that sentence express the consistency of T? Feferman claims that it does not. In his view the predicate $Prf_T^C(x, y)$ is "intensionally incorrect, so we can ascribe no clear intensional meaning to the result." Moreover, in contrast to Rosser's manipulation of the intension of the definition of provability in order to obtain an improved extensional result, with this definition "we cannot formally derive other properties of provability in terms of the definition $[Prf_T^C(x, y)]$," and hence "we see no results of extensional interest which follow from the proof of $[Con^C(T)]$" (Feferman (1960), 37). Note, though, that there are two claims implicit in Feferman's argument. First, he suggests that it is we who ascribe the intensional meaning to a metamathematical result. Second, he suggests that if no fruitful consequences of "extensional interest" result from the use of predicates with "non-standard" intensions, then that counts as evidence against believing that the predicate expresses a meaningful concept. Let's return to these in a moment, once we've further sharpened the problem.

3.2 Sharpening the Problem

Franks (2009) makes the distinction between extensional and intensional interpretations of metamathematics by arguing that on an extensional interpretation, statements about formal systems are "theory-independent" in the sense that such statements are independent of what the formal system counts amongst its theorems and proofs. Hence, in the case of a consistency sentence, it is possible for there to be, from our "theory-independent" point of view, various extensionally equivalent means of expressing the consistency of a formalism regardless of whether the formalism itself proves them to be equivalent. But on an intensional interpretation of metamathematics, "statements about mathematics [formal systems of mathematics] are always part of a mathematical theory," or "theory-dependent," in the sense that statements about the system depend entirely upon what is provable in it (Franks (2009), 7). Hence, on an intensional interpretation, a function is not defined unless the formalism proves that it satisfies the relevant existence and uniqueness conditions, two formulae cannot be counted as equivalent unless the formal system proves it, and a formal system proves a formula just in case *that* formal system proves the "correct" formalization of the statement that it proves the formula. But note that both extensional and intensional interpretations presuppose standards of correctness. From the extensional point of view, that standard is tied to whether a metamathematical predicate picks out extensionally identical numerical classes, independent of the details of the formal system for which it is defined. From the intensional point of view, it is tied to how a formal system proves (or fails to prove) that a formula is a correct expression of the metamathematical concept *for* that particular system, and hence, the standard varies dependent upon the details of the formal system for which the predicate is defined.

For example, Hilbert and Bernays (1939) set out three conditions aimed to furnish standards for the intensional correctness of a proof predicate. In order for a proof predicate to be intensionally correct, T must satisfy:

(HB1) if $\vdash_T \phi$ then $\vdash_T Prov_T(\phi)$;

(HB2) $\vdash_T Prov_T(\phi \to \psi) \to Prov_T(\phi) \to Prov_T(\psi)$;

(HB3) $\vdash_T Prov_T(\phi) \to Prov_T(Prov_T(\phi))$.[4]

While (HB1) ensures that T recognizes that its proofs are recursively enumerable, (HB2) ensures that T recognizes that all its proofs are closed under *modus ponens*, and (HB3) ensures that T recognizes that all its provable theorems are provably provable. Hilbert and Bernays show that these three conditions are sufficient for the proof of (G2). Hence, one need only verify

[4] In Hilbert and Bernays (1939) the derivability conditions stated are slightly different than the three conditions stated above, which are the improved conditions found in Löb (1955). For the original reference see Hilbert and Bernays (1939), 295*ff*.

that a formal system satisfies these conditions in order to prove (G2) without being required to provide an explicit definition of the proof predicate. However, there are at least two problems with their approach. First, their motivation is to find the conditions sufficient for (G2). But if the possibility of a consistency proof is in question, then it cannot be the case that the standards for intensional correctness are just those conditions that permit the proof of (G2), since that begs the question. Second, while it is true that given a formal system that satisfies these conditions one may assume that (G2) is provable without an explicit definition of proof, it is not true that given an explicit definition of proof for a formal system, there is a general method for testing whether it satisfies these conditions. Hence, these conditions are instrumental for having the resources required to prove (G2), but fail to fix the standard for correctness according to how it varies with the details of the particular formal system for which it is defined.

In response to this, Feferman (1960) proposes a general approach to intensionality that follows Hilbert-Bernays, but attempts to overcome its limitations. Feferman identifies a set of conditions that the proof predicate must meet such that it correctly expresses the concept for any formal system for which it is defined. Given a formal system T, one associates with it the class of all formulae $\tau(x)$ that numerically define the set of axioms of T. Then, by formalizing the concept of *logical* proof, one associates with each τ a formula $Prf_{\tau}(x, y)$ and a sentence Con_{τ}. Hence, on his approach, the proof relation is fixed by the set of non-logical axioms of a formal system, and the proof and consistency predicates are built out of the concept of logical proof underlying those axioms. He writes that on his approach "whenever a formula $\tau(x)$ can be recognized to express correctly that x is an axiom of T, the associated sentence Con_{τ} will be recognized to be a correct expression of the proposition that T is consistent." Due to the generality of the approach, "all intensionally correct statements of consistency for familiar theories can be obtained as special cases" (Feferman (1960), 38). Any changes made to the formula must be from the "inside," that is, made by varying the proof definition based on choosing subclasses τ' of the class of formulae τ. Feferman contrasts his approach with approaches like Rosser's that, he claims, are changes made from the "outside," a designation that is intended to suggest that, while useful in some respects, such changes are artificial because they involve changes in the concept of logical proof.

3.3 Problems for Interpreting (G2)

Thus far we have looked at some examples of how the extension and intension of a proof predicate come apart, and looked at two examples of approaches to providing conditions that allow for some generalizations of (G2). Our question is now whether the extensional or intensional approaches entail claim (1) or (2) or both. Hence, the basic questions are whether the extensionalist is able to secure the inference from his interpretation of (G2) to claims (1) and (2), and whether the intensional approach defended by Feferman (1960) is able to secure the inference from his interpretation of (G2) to claims (1) and (2). For the extensionalist, the task is to find general versions of (G2) such that the conditions that it must meet in order to be derived are met by all correct formalizations of consistency. Detlefsen (1986) calls this the "stability problem," and writes that the extensionalist must "show that every set of properties sufficient to make a formula a fit expression of T's consistency is also sufficient to make that formula unprovable in T (if T is consistent)" (Detlefsen (1986), 81). Hence, if some correct formalization of consistency does not meet the conditions, then the inference from (G2) to (1) or (2) is "unstable." On the other hand, the intensionalist must find a set of necessary and sufficient conditions for metamathematical predicates such that a T-proof of those conditions amounts to a demonstration of the intensional correctness of those conditions in T. But here the problem is slightly different, since the intensionalist must, in addition, explain how the proposed conditions are constitutive of the metamathematical concept.

How does the extensionalist face his task? Suppose, for the moment, that his standard of correctness is that the predicates must be numerically correct in the sense that they pick out identical numerical classes. Consider again the proof predicate $Prf_T^C(x, y)$. Above we mentioned that it is numerically correct, and that T need not recognize (in the sense of prove) that the two predicates are equivalent but that it suffices for us, from our theory-independent point of view, to be able to recognize the equivalence. If one builds the "usual" consistency statement out of $Prf_T(x, y)$:

$$Con(T) := \neg\exists x Prf_T(x, \bot)$$

then while T fails to prove $Con(T)$, it does prove $Con^C(T)$. That is, while (G2) goes through for $Con(T)$ it does not go through for $Con^C(T)$. Hence, since there is a T-proof of $Con^C(T)$ and since it is formalizable in T, the inference from (G2) to (1) and (2) fails if the extensionalist's standard is that a proof predicate and consistency statement *merely* be numerically correct. It seems that there are two ways to rectify the situation. First, he might argue for the claim that there is a theory-independent concept of consistency expressed by $Con(T)$ that $Con^C(T)$ fails to express. But this approach requires that he have in hand a theory of meaning or content that supports his claim. Hence, accepting or rejecting $Con(T)$ as expressing uniquely the consistency of T will depend on our inclination to accept or reject the underlying theory of meaning. Or, he might argue that some standard other than numerical correctness governs the arithmetization of metamathematical concepts. But this approach changes the standard because of (G2), and so begs the question in that it assumes that the inference from (G2) to (1) or (2) is valid rather than establish it on independent grounds.

It might be the case that there are other means for the extensionalist to defend the inference from (G2) to (1) or (2), but the prospects do not seem encouraging. On the other hand, on the intensional interpretation of (G2), the inference from it to (1) or (2) is immediate. To understand why, consider the following. Gentzen's *Hauptsatz* asserts that for every classical proof there is a corresponding "cut-free" proof that does not use classical indirect proof methods (that might be longer, but combinatorially simpler). It is possible, then, to build a "Gentzenian" proof predicate $Prf_T^G(x, y)$ that is extensionally equivalent to $Prf_T(x, y)$. Hence, if the cut-rule is admissible for T, then it is possible to build two formulae, $Con(T)$ and $Con^G(T)$, that are extensionally equivalent. However, depending upon our choice of formalism, T might be proof-theoretically rich enough to arithmetize $Con(T)$, but not be rich enough to prove the formalization of the *Hauptsatz*, and hence, T might not prove that $Con(T)$ and $Con^G(T)$ are equivalent. Hence, from the point of view of some formalisms, $Con(T)$ and $Con^G(T)$ do not express the same concept. It follows that at most one of the two predicates is intensionally correct for those formalisms. Hence, if such a formalism does not prove one of the two formulae, and recognizes the one not provable as an expression of its own consistency, then (1) follows immediately from (G2) for this formalism. Moreover, for the intensionalist there is no other option than to prove in the formalism that the predicate that expresses that formalism's consistency is unprovable, and so (2) follows from (1). It seems, then, that on an intensional understanding of metamathematical predicates, (1) and (2) follow immediately from (G2) for specific formalisms. But our questions are now twofold. First, is it possible to obtain results that do not depend upon the details of specific formalisms? Second, what are the standards that constitute a formula's intensional correctness?

In Feferman's view, it is possible to answer these two questions at the same time. He argues that the goal is to find conditions that are constitutive of metamathematical concepts that permit generalizations of (G2) such that one need not restrict the arithmetization of provability to specific formalisms. Predicates that meet said conditions permit the inference from (G2) to (1) and (2). Let $Proof_T(x)$ represent the unary predicate "x is a proof in T;" let $Prf_T(x, y)$ represent the binary predicate "x is a proof in T of y;" and let $Pr_T(y) := \exists x Prf_T(x, y)$ represent the predicate "y is a theorem of T". In order for a provability predicate for T to be intensionally correct, Feferman claims that it must satisfy the following conditions:

(i) $\vdash_T \forall\phi\forall\psi\,(Pr_T(\phi) \wedge Pr_T(\phi \to \psi) \to Pr_T(\psi))$;

(ii) $\vdash_T \forall\phi\,(Pr_T(\phi) \to Pr_T(Pr_T(\phi)))$;

(iii) $\vdash_T \forall x(Proof_T(x) \to Pr_T(Proof_T(x)))$;

(iv) $\vdash_T \forall\phi\forall x(Prf_T(x, \phi) \to Pr_T(Prf_T(x, \phi))))$.[5]

[5] Note that the list above is incomplete, but captures Feferman's most important conditions that a formal system must meet in order for its provability predicate to be intensionally correct. For the original list and further reference, see Feferman (1960), 60*ff.* Also note that, for simplicity, reference to Gödel numbering in the predicates is omitted.

For each condition, if T satisfies it, then T expresses a concept constitutive of provability. Condition (i) expresses the concept that theoremhood is closed under *modus ponens*. Condition (ii) expresses the concept that provability is idempotent, or more suggestively, transparent. Condition (iii) expresses the concept that all T-proofs are formalizable in T. Condition (iv) expresses the concept that all T-proofs of closed formulae are formalizable in T. Feferman claims that a "minimal" standard governs the choice of conditions on intensional correctness. At the least, such conditions must preserve the logical—in contrast to the mathematical or finitistic—steps in a derivation. Hence, for Feferman, provability in T reduces to logical provability.

For example, Feferman derives an important consequence from his approach that shows how conditions (i)-(iv) preserve a derivation's logical steps. Theorem 5.9 (Feferman (1960), 68) demonstrates that it is possible to choose a subformula $\tau *$ of the class of formulae τ that strongly represent the axioms T of a consistent recursive extension $\mathcal{T}$ of (PA) such that:

$$\vdash_{PA} Con^{\tau *}(\mathcal{T}).$$

In other words, there is a formula that is a numerically correct (strong) representation of the consistency of a set of axioms extending (PA) such that (PA) proves the extension consistent. On the face of it, theorem 5.9 (and corollary 5.10) appears to contradict (G2). But Feferman argues that the appearance is just that, and that his approach makes the problem clear. He writes that "one particular conclusion we can draw is that the formula [$\tau *$], although it extensionally corresponds to [T], does not properly express membership in [$\mathcal{T}$]" (Feferman (1960), 69). That is, $\tau *$ is extensionally correct but intensionally incorrect because it fails one of the conditions (i)-(iv), specifically condition (ii). Hence, we have the following:

Theorem 2.0: If $\vdash_{PA} Con^{\tau *}(\mathcal{T})$, then $\nvdash_{PA} Pr_{PA}(\tau *) \rightarrow Pr_{PA}(Pr_{PA}(\tau *))$.

Moreover, since theorem 2.0 is formalizable in (PA) by conditions (iii) and (iv), and (PA) is complete for the proof-predicate, we have the following:

Corollary 2.1: $\vdash_{PA} \Big(Con^{\tau *}(\mathcal{T}) \rightarrow \big(Pr_{PA}(\tau *) \wedge \neg Pr_{PA}(Pr_{PA}(\tau *))\big)\Big)$.

Both say that (PA) itself recognizes that there are intensionally incorrect formulations of provability that are nonetheless extensionally correct. That is, if it proves its own consistency, then it *knows* that it has failed to produce a predicate that expresses the correct concept. For Feferman, it follows that it knows that it has failed to preserve the logical steps in its derivations.[6]

There are at least three points in Feferman's analysis to which one might apply some pressure. Franks (2009) identifies two. He claims that one might demand a defense of Feferman's conditions (i)-(iv) on the "grounds that they seem from one point of view rather strong and from another point of view too weak" (Franks (2009), 122). Feferman's conditions are "too weak," for Franks, because "certain basic properties about provability do not appear in Feferman's list, and in fact cannot" (ibid.). He cites reflection principles:

$$(Ref) \qquad Pr_T(\phi) \rightarrow \phi$$

that express the concept that all provable formulae are true and that, by Löb's theorem, are unprovable in formalisms satisfying conditions (i)-(iv). But the objection misfires because Feferman makes no claim for his conditions governing all possible properties, but rather just those properties that express the logical concept of proof. By contrast, (i)-(iv) are "too strong," for Franks, because if ϕ is provable, by condition (ii) there exist infinitely many (Gödel) numbers $\#\phi, \#Pr_T(\phi), \#Pr_T(Pr_T(\phi)), \ldots$, and that "seems like an ontological assumption very far removed from the notion of ϕ's provability" (Franks (2009), 123). But this objection misfires as well. The formulae obtained by iterating condition (ii), given ϕ's provability, are syntactic, and hence, make no claims about the ontology of numbers. At best, such formulae only make claims in a model-theoretic interpretation of the formalism, and even then, it is unclear how formalisms are committed to ontology. Hence, Franks is not entitled to conclude that "Feferman's proposal is incomplete as a method for the explicit arithmetizations needed for

[6] Feferman's approach clearly permits the inferences to (1) and (2). Feferman pursues arithmetization as it appears within formalisms irrespective of a predicate's numerical correctness. Hence, it is possible to claim that if a consistency predicate for a formalism is intensionally correct, then it is unprovable in that formalism, from which (1) follows; and that there are consistency proofs formalizable in T only if the consistency predicate's formalization is intensionally incorrect, from which (2) follows.

a fully mathematical treatment of metatheory" (ibid.). In what follows, we begin to develop our proposal alongside Feferman's third pressure point.

4 Inside Consistency

Thus far, we introduced the distinction between (G2) and what it "says," that is, claims (1) and (2). Then, we analyzed the distinction between extensional and intensional interpretations of metamathematics as it arises in discussions of inferring claim (1) or (2) from (G2). In section (3.2) it was argued that the prospects for inferring claims (1) and (2) from (G2) on an extensional interpretation of metamathematics are poor. Then it was argued that on an intensional interpretation of metamathematics, the prospects are much better, but that the intensionalist must, in addition, explain how the chosen conditions constitute the metamathematical concept being arithmetized. Then we showed that two objections by Franks failed to undercut two pressure points in Feferman's approach, and suggested that there is at least a third point. In section (3.1) we saw that Feferman suggests that if no fruitful consequences of "extensional interest" result from the use of predicates with "non-standard" intensions, then that counts as evidence against believing that the predicate expresses the correct concept. But note that his suggestion opens up a route to the following alternative approach to arithmetization. Are there changes in the logical concept of proof, i.e., "changes from the outside," that might be warranted by how formalisms generate "fruitful consequences" relative to our choice of provability predicate? In this section our goal is to begin to develop an approach to arithmetization the main feature of which is that one is warranted to believe that a consistency predicate expresses the correct concept, even if it skirts (G2), on the condition that its use generates fruitful consequences. Our claim is that these consequences constitute a unique kind of "internal" mathematical evidence for intensional conditions that express different concepts of proof.

4.1 Evidence and Arithmetization

Let's return, for the moment, to Feferman's claim that the standard for the choice of conditions that govern a predicate's intensional correctness is that they must express the logical concept of proof. Other conditions might express a concept, but in Feferman's view, unless it meets all four conditions listed above, that concept fails to be the logical one. Feferman (1960) does not discuss in much depth the reason behind his claim that the concept expressed must be the logical concept, nor does he discuss why conditions (i)-(iv) jointly express it. However, in the formalisms that Feferman studies every proof predicate is either a "change in the logical concept of proof," or it is provable, in the formalism, that it is equivalent to the "standard concept," and hence, in Feferman's view intensionally correct. For these formal arithmetics, the "standard concept" is the concept of a formal deduction: a sequence of formulae that are instances of axioms or obtained from the axioms by applying the inference rules finitely many times. Moreover, because the arithmetics Feferman studies are strong arithmetics such as (PA), variations in the formalism's proof theory are insignificant.[7] Hence, for example, since (PA) proves the *Hauptsatz*, it is provable in (PA) that every cut-free proof is equivalent to a proof that uses cut; and since (PA) proves that every Hilbert-style proof is equivalent to a Gentzen-style proof, (PA) doesn't recognize the difference between the two proof-systems. Implicitly, then, it seems that Feferman is committed to the claim that the concept of proof that is expressed by a formalism need not capture differences in the formulation of the proof theory nor needs to be responsive to the particular details of the proof-theoretic capabilities of a given formalism.

In part, this implicit commitment is the means by which Feferman ensures that his characterization of the conditions constitutive of the concept of proof (for formalisms) remains fully general. But, as noted above, there

[7] In fact, Feferman (1989) shows that his approach to arithmetization may also be utilized for weaker formal arithmetics such as (PRA) and its conservative extensions. There, the solution to the problem of the inference from (G2) to (1) and (2) is solved through what he calls a "finitary inductively presented logic," where if a formalism satisfies a set of inductive conditions, then it satisfies (1) and (2). However, there are still weaker formalisms for which the problem resurfaces. See the discussion below.

are formalisms that fail to satisfy conditions (i)-(iv). How should we treat such formalisms? Feferman's response might be to argue that since such formalisms fail his conditions, the proof predicates for such formalisms fail to express the logical concept of proof, and hence, fail to reflect or present its own logic. But then Feferman seems to lose traction on the claim that his characterization of the conditions constitutive of the concept of proof hold in general. There are at least two responses to the problem of formalisms that fail to satisfy Feferman's conditions. One might, as Franks suggests, argue on *a priori* grounds that his choice of conditions are not constitutive of the concept of proof. Or one might, as suggested above, argue that the failure of an arithmetization to meet one (or more) of Feferman's conditions yet produce consequences of "extensional interest" constitutes evidence against that condition for the particular formalism. On the first route, one assumes, "monistically," that the conditions a formalism must meet in order to express *the* concept of proof are identical for each formalism irrespective of its proof-theoretic capacities. On the second route, one denies that the monistic approach carries any weight and assumes, "pluralistically," that each formalism implicitly expresses its own concept of proof. If we pursue the first route, then we must explain why our choice of alternative conditions are *a priori* constitutive of the concept of proof but Feferman's are not. But if we pursue the second route, then we must find means to extract conditions for each member of a (possibly) countable set of formalisms and show that such conditions constitute counterexamples to Feferman's conditions.

Above we suggested that, because one must explain why the choice of conditions does not vary with details of the formalism—that is, why the representation of a formalism's metatheory within the formalism must meet the same conditions irrespective of its specific mathematical and proof-theoretic capacities—pursuing the monistic route is unlikely to yield much fruit. Let's pursue the second, pluralistic, route for a moment. Consider the following. Recall that, above, we mentioned that if T is as strong as (PA), then it proves the *Hauptsatz* and hence proves the equivalence between $Pr_T(y)$ and $Pr_T^G(y)$. However, if T is not strong enough to prove the *Hauptsatz*, then T fails to recognize the equivalence, and hence, from T's standpoint only one of the two provability relations is intensionally correct. Feferman seems to imply that in such cases the "standard" construction is the correct one. But that one arithmetization is correct while another is incorrect is a claim that it is possible to make only from outside of T's standpoint, i.e., from the point of view of a proof-theoretically "richer" theory in which *we* know that the two formula are equivalent but that T fails to prove this. From the point of view of T itself, neither formula has more or less claim to correctness. It might be the case that, from within T, we want a formula that expresses T's provability relative to a proof-theory that is syntactically simpler because it does not contain the cut-rule. In such a case, the "natural" choice is the Genztenian predicate. Or, it might be the case that we want a formula that expresses T's provability relative to a standard classical proof-theory containing the cut-rule. In such a case, the "natural" choice is the standard predicate. Hence, if one is unwilling to conclude that such formalisms fail to express a proof concept, then for formalisms that fail to meet conditions (i)-(iv) and that fail to prove the equivalence between non-standard and standard metamathematical concepts, one *must* find the means to formulate such concepts by varying the concept of proof "from the outside."

What variations are permissible? Franks (2009) proposes an approach to the arithmetization of metamathematical concepts that he contrasts with Feferman's "logical" approach by claiming that it is "fully mathematical." First, some background. An equation of the form:

$$f(\vec{x}) = 0$$

where the unknowns $\vec{x} = (x_1, \ldots, x_n)$ are integers and f is a function of the integers is a *Diophantine equation*, and a Diophantine *problem* asks whether or not a given Diophantine equation has a solution in the integers. In Herbrand (1930) it is shown that metamathematical questions such as "is ϕ a theorem of the formalism S?" are equivalent to particular Diophantine problems.[8] Herbrand's Theorem asserts that a formula (possibly with

[8] As is well known, Hilbert (1900) asked, in his famous Tenth Problem, whether in general there is an algorithm that tells us whether or not a given Diophantine equation has a solution. Matiyasevich (1970) answered Hilbert's question negatively, so that there

quantifiers) is provable in the predicate calculus if and only if there exists a tautological disjunction (its Herbrand expansion) in which all variables are replaced by closed numerical terms. Kreisel (1951) and (1952) claimed that Herbrand's theorem provides a means of extracting "constructive content" from metamathematical questions. If we ask if a Diophantine problem has a solution by effectively substituting closed numerical terms for the unknowns, then the metamathematical question that corresponds to the Diophantine problem can be answered in any formalism that proves those terms to be total functions. Franks claims that this approach provides a better solution to the problem of an intensionally correct consistency predicate: a Diophantine equation produces for every formalism T (the metamathematics) an intensionally correct formulation of the consistency of S (the formalism in question) just in case the equation has a solution in T's provably total functions. Hence, on his approach, the consistency statement for a formalism varies with the functions the metamathematics proves total, and the result of arithmetizing the question of whether an equation has a solution in T's terms is a formula that corresponds to "the statement of S's consistency 'as', one might say, 'T thinks about the question'" (Franks (2009), 10).

4.2 Franks on Consistency

Let's dig in to the details of the view for a moment. Suppose that there exists a solution in T's provably total functions for a Diophantine equation. Then, by Herbrand's Theorem, we may construct a *Herbrand disjunction* that states that one (or more) of T's provably total functions is a solution to that equation such that S proves the disjunction. Since S proves the disjunction, it is possible to construct, in T, a consistency predicate for S from the Herbrand disjunction that is accurate for it from T's point of view. However, since it is possible that T might be proof-theoretically stronger than S, the question arises as to whether it is possible for S to construct its consistency predicate on its own. If so, then since S proves the Herbrand disjunction and the consistency predicate is built out of the Herbrand dis-

is no general method for determining if so or not.

junction, then it ought to follow that S proves its own "consistency," where the arithmetization of consistency does not result in the standard consistency predicate, $Con(S)$, but rather $Con^H(S)$, its Herbrand-consistency. Franks answers this in the affirmative. More precisely, let $Pr_T^H(\phi)$ be the Herbrand-provability predicate for T, read as "there is a Herbrand proof of ϕ," and which says that for a finite set T' of the axioms of T there is a Herbrand-disjunction of ϕ proved in T' in which closed numerical terms and functions are substituted for variables such that the resulting quantifier-free disjunction is a propositional tautology. Then, the Herbrand consistency of T, $Con^H(T)$, is the assertion that no finite set T' proves a Herbrand-disjunction with a contradiction (or $\bot$) as the end-formula of the proof. Note that there are two features of this approach that indicate that it is a change "from the outside" in the concept of proof. First, in the arithmetization of provability, and hence in the construction of the consistency statement, one must arithmetize the concept of proof using a predicate that expresses it for propositional logic. Second, in formalisms that do not prove Herbrand's Theorem, $Pr_T(\phi)$ and $Pr_T^H(\phi)$ are not equivalent, and neither are their consistency statements. Hence, as discussed above, for such formalisms one is faced with a choice: either stick with the standard statement $Con(T)$ or change the concept of proof from the outside.

For concreteness, consider a kind of formalism that we shall call, after Buss (1986), Bounded Arithmetic (BA).[9] Let Con_β and Con_β^H be two consistency formulae for (BA) such that β is a recursively enumerable arithmetization of the axioms of (BA), where the former is the standard consistency predicate while the latter is its Herbrand consistency. No member of this class proves Herbrand's Theorem. Hence, it follows that:

$$(BA) \nvdash Con_\beta \leftrightarrow Con_\beta^H. \quad (*)$$

On an intensional interpretation, only one of the two formulae can be accurate from (BA)'s point of view. Since Con_β is not provable in (BA), whereas

[9] Buss (1986) investigates the computational properties of a class of formalisms, the "bounded arithmetics," which he denotes as S_k^i, and which include Q, a quantifier-free schema of induction, plus an axiom asserting that exponentiation is a total function.

Con^H_β is, Feferman might claim that this speaks against the intensional correctness of Con^H_β. But Franks draws an orthogonal conclusion. He argues that ($*$) provides an analysis of (G2) for (BA). Though the quantified formula that is the result of arithmetizing the standard consistency statement for (BA) is true in the sense that there exist numerical terms that can be substituted for the variables, such terms do not belong to the class of functions that (BA) proves to be total, and hence, do not belong to its provably total recursive functions. "Thus," he writes, "the unprovability of the standard consistency statement in bounded arithmetic appears merely to be a consequence of the fact that there are function symbols in these theories' languages that they do not prove to be functions" (Franks (2009), 150). For Franks, the unprovability of Con_β in (BA) follows from the fact that (BA) contains expressions for functions that (BA) does not prove to be total. But the fact that a formalism contains redundant expressions seems to be a poor reason to believe that Con_β expresses the consistency of (BA), and Franks takes it to be evidence against the intensional correctness of Con_β. Hence, he concludes, here's a case in which a "non-standard" concept of consistency expressed by (BA)'s Herbrand-consistency appears to be preferable.

Let's reflect on this approach for a moment. While we have argued that the idea that the evidence for or against the choice of an intensional predicate for a formal system ought to arise from the constraints of that system itself, Franks' choice of Herbrand-consistency as an alternative to the standard formulation raises some questions. First, in general the existence of a solution given a Diophantine equation is undecidable, though cases of it are decidable. Jones (1980) shows that, given a Diophantine equation, if there exists a decidable algorithm for its solution, then the degree of the equation must be strictly less than four. Hence, a formal system in which its Herbrand-consistency is provable must have a set of provably total recursive functions whose solutions are in a Diophantine equation of less than four degrees. Hence, and in plain(er) English, formal systems in which Herbrand-consistency is preferable to the standard formulation of consistency and provable where the standard is not have extremely limited mathematical application.[10] Second, Franks might counter that it might be possible to prove the Herbrand-consistency of a formal system stronger than the bounded arithmetics, just in case there exists a solution to a Diophantine equation of degree greater than or equal to four in that formal system's provably total recursive functions. But then the existence of a consistency proof for a formalism depends upon open mathematical problems involving the solvability of Diophantine equations, many of which are unknown and whose general problem, as we remarked above, is undecidable. Hence, the question of a system S's consistency depends upon the existence of a solution to a Diophantine equation in S's provably total recursive functions. It seems, then, that we've foisted the problem of a consistency proof for a formal system onto the solvability of open mathematical problems, and that direction of dependence is in conflict with Franks' claim that an intensional proof-theoretic analysis ought to contribute to mathematics.

But the problems in Franks' approach go deeper and appear to affect many approaches to the problem of consistency when it's formulated as a problem of arithmetization. That is, either we tinker with the arithmetization of consistency for particular formalisms in order to skirt (G2) or we preserve and generalize (G2) and stick with its canonical arithmetization.[11] In other words, when our approach begins with the question of which arithmetization is the proper one we are faced with a dilemma. Either an arithmetization of consistency is suited, a la Franks, to the particular formalism for which it's formulated but its applicability is not fully general, or an arithmetization is the canonical one, a la Feferman, and its applicability fully general but unsuited to the particular means with which a given formalism is formulated. On the first approach a formalism can be said to more correctly express its own consistency but then its application is extremely limited, while on the second approach arithmetization for specific formalisms is absorbed into the canonical expression of consistency but its application is fully general. Hence, we seem to have a dilemma. Either the

[10] Indeed the limit, and hence the extent to which a system may prove its own Herbrand-consistency, is fixed to just the so-called set of bounded arithmetics studied in Buss (1986).

[11] Cf. Solomon Feferman (2012).

inference from (G2) to claims (1) and (2) fails for extremely weak theories with limited applicability, or the inference holds but we lose the explanation for why it holds. In part, this is because Franks primarily puts pressure on the question of the arithmetization of the consistency predicate, while in fact the issue goes deeper than that. Indeed, as we shall argue in our conclusion, the issue here lies with how the provability predicate expresses the concept of proof. Hence, we shall begin to spell out our own, "third pressure point," to Feferman's 1960 argument and develop some of its consequences while indicating further directions for proof-theoretic research. Our main question is when is a change in the concept of proof for a given formal system warranted and when is it not?

5 Conclusion

Recall that in section (3) we looked at Feferman's suggestion that if no consequences of extensional interest result from the use of predicates with intensions that fail to meet one of conditions (i)-(iv), then that should count as evidence against believing that such a predicate expresses the correct concept. Our claim there was that the suggestion opened up a third pressure point in our discussion of arithmetization—namely, the idea that if an arithmetization with a non-standard intension of provability does produce a consequence with extensional interest, then that ought to count as evidence for that non-standard intension. Let's consider this more precisely for a moment. In proof theory, reflection principles express a form of soundness—if ϕ is provable, then ϕ is true—and as such their use within a base theory T applied to a proof-theoretically stronger theory T' appears to be desired. For a reflection principle, when localized to a weak theory T, expresses the concept that all theorems derived via T are true. On this suggested approach for a theory we then have:

Theorem 3.0: If $\vdash_T Pr(\phi) \rightarrow (\phi)$, then $\nvdash_T Pr_T(\phi) \rightarrow Pr_T(Pr_T(\phi))$,

and hence:

Corollary 3.1: If $\vdash_T Pr(\phi) \rightarrow (\phi)$, then $\nvdash_T \bot$

Since Corollary 3.1 is formalizable, it follows that if T encodes a reflection principle for itself, then T fails to prove that it is inconsistent, and hence, does not know that it is not. Our claim is that this isolates a special, combinatorial concept of mathematical proof. Under this concept the inferences from (G2) to claims (1) and (2) fail. But immediately at least two natural questions arise. First, since this is far too general—ϕ here is unrestricted for the purposes of illustration—are there natural formalisms in which condition (ii) fails that one can show a similar bootstrapping procedure for the arithmetization of the consistency statements for such theories? Second, and more importantly, when a reflection principle holds and condition (ii) fails, it follows that the set of axioms for which said principle holds *not* recursively enumerable. Is this too large a concession for such an apparently small payoff? On the first question our hunch is that there is, and on the second our hunch is that it is not, but we leave both to future research.

Our claim here is not that this constitutes the only approach. Rather, where Feferman's analysis of the concept of proof reduces to what he calls the "'logical" concept, and where Franks believes to have isolated the purely "mathematical" concept of proof in an arithmetization, our claim is that we have isolated a combinatorial concept of proof. Whatever the proposed conditions, they must at least be compatible with the sense of the formula's intensional correctness and capture how this justifies the consistency formula as a correct expression of consistency for the particular formalism under consideration. But for the pursuit of Hilbert's Program using limited means, this seems to be as it should.