Proving Unprovability

Bruno Whittle

July 19, 2016

1 Introduction

Suppose that we have accepted some mathematical theory T (i.e. some axioms and rules of inference). Is there some natural, generally applicable way of extending T to a theory S that can prove a wide range of things about what it itself (i.e. S) can prove, including a wide range of things about what it cannot prove, such as claims to the effect that it cannot prove certain particular sentences (e.g. $0 = 1$), or the claim that it can never prove both a sentence and its negation (i.e. it is consistent)? Prima facie, one would have thought that the answer would be 'yes'. For if we accept a given theory, then one would have thought that we also accept that it is consistent (as well as other such claims about what it can prove). For example, if we accept a given theory of sets, then one would have thought that we also accept that it is consistent. But then a theory representing everything that we believe about a certain subject (e.g. sets) must be one that can prove such claims about itself; that is, must be along the lines of S as described.

However, typical characterizations of Gödel's second incompleteness theorem, and its significance, would lead us to believe that the answer must in fact be 'no'. For characterizations of this theorem tend to be along the lines of: no consistent formal system meeting certain relatively undemanding conditions can prove its own consistency. Thus, the following, from the first lines of Panu Raatikainen's *Stanford Encyclopedia of Philosophy* entry on the incompleteness theorems, is representative:

> The first incompleteness theorem states that in any consistent formal system F within which a certain amount of arithmetic can be carried out, there are statements of the language of F which can neither be proved nor disproved in F. According to the second incompleteness theorem, such a formal system cannot prove that the system itself is consistent (assuming it is indeed consistent). [2015]

According to this statement of the second theorem, no consistent formal system 'within which a certain amount of arithmetic can be carried out' can prove its own consistency. If that is correct, then it would seem that the answer to our question must indeed be 'no'. That question, recall, is: given some theory T that we accept, is there some natural, generally applicable way of extending it to a theory S that can prove a range of things about what it itself can prove, including a range of things about what it cannot prove (such as its consistency)? For it is plausible that any theory that we are capable of accepting will at least correspond to a formal system (i.e. when formalized). Thus, if the answer is to be 'yes', then both the theories that we start with and the extensions are going to be (or correspond to) formal systems. But then as long as the initial theory T contains a reasonable amount of arithmetic, the extension S will have to be (or correspond to) a consistent formal system that itself contains a reasonable amount of arithmetic *and* that proves its own consistency—which is impossible, according to this statement of the theorem.

Further, typical characterizations of the significance of this result are apt to reinforce the impression that the answer must be 'no'. For example, in his paper 'What Gödel's Incompleteness Result Does and Does Not Show', Haim Gaifman characterizes what the result *does* show as follows.

> Any deductive system, T, that formalizes mathematical reasoning must leave something outside: its own consistency, expressed as Con(T), cannot be derived in it. As we remarked above, a computer that proves theorems generates proofs in some formal system… If the computer can "know" only what it can prove, then it cannot know that it is consistent (that is, never produces a contradiction)… A mathematician realizes by self-reflecting on his own reasoning that his inferences can be formalized by such-and-such deductive system. From which the mathematician can go on to infer that the system in question is consistent… Gödel's result shows however that self-reflection cannot encompass the whole of our reasoning; that is, it cannot comprehend itself within its horizon. [2000b: 466–69]

What Gödel showed was that if T is a formal system containing a certain amount of arithmetic, then T can prove 'coded' versions of claims about what it itself can prove (i.e. using gödel numbering). He showed further that if T is consistent, then it will not *in this way* be able to prove its own consistency. That is, it will not be able to prove its coded consistency statement, $\mathrm{Con}(T)$. However, although Gaifman *starts* here with claims about coded consistency statements, the claims that he goes on to make seem clearly to be about consistency statements quite generally. And these seem to require that the answer to our question must be 'no': they are presumably implicitly restricted to theories meeting some minimal constraints (e.g.

concerning how much arithmetic is contained); but if there *is* some natural, generally applicable way of extending any given theory to one that can prove its own consistency, then it would be hard to see these claims about what computers *cannot* prove, and about the limits of mathematical self-reflection, as anything other than false. What is supposed to justify these more general claims? The thought is of course that the argument of the second theorem will apply not only to the proof of coded consistency claims, but also to any reasonable alternative way of proving things about what is provable in a given theory.

The aim of the present paper, however, is to explore a positive answer to our question. It is to develop natural ways of extending any given theory that we might accept to one that can prove a range of things about what it itself can prove, including its consistency. If this can be done, then it would show that the thought behind typical characterizations of the significance of Gödel's result—that the argument of the result will apply to any reasonable way of proving things about provability—is mistaken.

1.1 Proof and Truth

The general approach that I will pursue is one that would seem to be very natural, but which, surprisingly, does not seem to have been explored. This is to follow the lead of recent (and not so recent) approaches to truth and the Liar paradox. In particular, approaches that develop accounts of languages that contain their own truth predicates. This approach to the question we are concerned with would seem to be natural because in each case—i.e. the problem of how languages can contain their own truth predicates, on the one hand, and the problem of how theories can prove things about their own provability, on the other—the obstacles that stand in the way of a straightforward solution would seem to be very similar. Thus, in the case of truth, the principle obstacle is the Liar paradox, i.e. the paradox that results from sentences that say of themselves that they are not true, together with other paradoxes of the same family. These show that languages containing their own truth predicates cannot have all of the properties one would have expected. Similarly, in the case of provability, the main obstacle arises from sentences that say of themselves that they are unprovable in the theory in question, as well as other sentences belonging to the same family (see below). Just as in the truth case, the existence of such sentences shows that theories about their own provability cannot have all of the properties one would have expected.

To illustrate, suppose that S is a theory about, among other things, provability in S, and suppose that $\mathrm{Pr}(x)$ is a predicate of S intended to mean *provable in* S.

Suppose also that c is an individual constant of the language of S such that S proves c = '$\neg\mathrm{Pr}(c)$'.[1] The existence of such a constant shows that S cannot have all of the properties that one would have expected of a theory concerned with its own provability. There are a number of ways of seeing this point. But the simplest is perhaps as follows. Prima facie, one would have expected a theory about its own provability to prove every instance of the following schema (for sentences A: by an instance of this schema I mean the result of replacing *both* occurrences of the first letter of the alphabet with a given sentence).

(1) $\mathrm{Pr}(\text{`}A\text{'}) \rightarrow A$

After all, if S is an acceptable theory, then anything it proves will be true. And this fact about provability in S—i.e. that it is factive—is a basic and important one. So one would expect schema (1)—which constitutes the most straightforward expression of this fact—to be provable in S. However, an instance of (1) is $\mathrm{Pr}(\text{`}\neg\mathrm{Pr}(c)\text{'}) \rightarrow \neg\mathrm{Pr}(c)$, and so given c = '$\neg\mathrm{Pr}(c)$' and classical logic we will have $\neg\mathrm{Pr}(c)$. That is, if S proves (1) and is closed under classical consequence, then it is unsound (i.e. proves falsehoods): for it will prove ¬Pr(c), which says that it itself is unprovable in S. Even worse, one would expect S to be closed under the following rule.

(2) A / $\mathrm{Pr}(\text{`}A\text{'})$

But then, if S proves $\neg\mathrm{Pr}(c)$ and is closed under classical consequence, we will have an outright contradiction, i.e. $\mathrm{Pr}(c) \wedge \neg\mathrm{Pr}(c)$.

It is not simply that $\neg\mathrm{Pr}(c)$ is similar to a Liar sentence in that it says of itself that it does not have a certain property. It is also that the schema and rule that lead to the problem are similar to the schema that leads to the problem in the truth case: i.e. the truth schema: 'A' is true iff A. For (1) and (2) are of course weakenings of the truth schema, but with provability in place of truth.[2]

[1] One could make versions of the points below without assuming that the language of S contains quote-names: e.g. one could instead just suppose that the intended interpretation of c is $\neg\mathrm{Pr}(c)$. However, I will assume that the language does contain such names, since this simplifies things. Similarly, in place of $\neg\mathrm{Pr}(c)$ one could consider a sentence that says of itself that it is unprovable using general syntactic resources such as a diagonal function. One would expect a theory about its own provability to have the means to express such a function, even if its language does not contain an individual constant along the lines of c. However, I will assume that the language of S contains such an individual constant, since that allows a more straightforward presentation.

[2] An alternative argument showing that the existence of sentences such as $\neg\mathrm{Pr}(c)$ means that S cannot have all of the properties one would expect is, essentially, the argument of Gödel's sec-

I am not claiming that sentences such as $\neg\mathrm{Pr}(c)$, which say of themselves that they are unprovable in a given theory, give rise to *paradoxes* in anything like the way in which Liar sentences do—or indeed that they give rise to paradoxes at all. They do however give rise to limits on what theories can prove about themselves that are in many ways similar to the limits on what languages can truthfully express about themselves that Liar sentences give rise to.

And the parallel goes much wider. For there are of course many variants of the Liar sentence that give rise to paradoxes in a similar way: e.g. 'Liar-cycles', Curry sentences, Yablo sentences etc. And so it is with provability: that is, there are many variants on our sentence $\neg\mathrm{Pr}(c)$ that give rise to limits on what theories can prove about themselves in a similar way to that in which $\neg\mathrm{Pr}(c)$ does, and these are often similar to the sentences that give rise to variants of the Liar paradox. Here is just one example (but there are many more). A simple variant of the Liar paradox results from the following Liar-cycle (using T for truth): a denotes $\neg T(b)$, and b denotes $T(a)$. For a parallel example with provability, suppose S proves d = '$\neg\mathrm{Pr}(e)$' and e = '$\mathrm{Pr}(d)$'. Using (1), (2) and classical logic one can get a contradiction from these sentences similarly to as in the $\neg\mathrm{Pr}(c)$ case.[3]

It is thus natural to look to accounts of languages containing their own truth predicates as a guide in our attempt to find theories that can prove a range of things about their own provability. Specifically, in §2 I will consider classical approaches (i.e. theories that are closed under classical consequence), while in §3 I will explore non-classical ones. Concerning the latter I should forestall a worry: for it might seem extremely radical—not to say downright wrong-headed—to consider non-classical approaches to provability. After all, we are concerned here with which sentences can be derived in certain formal systems (or with theories that correspond to such), and surely it is as clear that classical logic is correct in this domain as it is that it is correct in arithmetic (for example). However, the suggestion is certainly *not* that classical logic fails to preserve truth in this context. It is simply that relaxing the requirement that our theories be closed under classical consequence

ond theorem (which I will discuss in §2.4 below). This shows that if S is closed under classical consequence and consistent, then it cannot prove $\neg\mathrm{Pr}(\text{`}c \neq c\text{'})$, for example, and also satisfy the Hilbert-Bernays-Löb derivability conditions. Again, however, the principles involved are similar to ones that one would naively expect to hold for truth: the derivability conditions correspond to weakenings of the truth schema together with a schema to the effect that modus ponens is truth preserving (which follows from the truth schema in classical logic).

[3]An instance of (1) is $\mathrm{Pr}(\text{`}\mathrm{Pr}(d)\text{'}) \rightarrow \mathrm{Pr}(d)$, giving $\mathrm{Pr}(e) \rightarrow \mathrm{Pr}(d)$. But another instance is $\mathrm{Pr}(\text{`}\neg\mathrm{Pr}(e)\text{'}) \rightarrow \neg\mathrm{Pr}(e)$, giving $\mathrm{Pr}(d) \rightarrow \neg\mathrm{Pr}(e)$, and thus $\neg\mathrm{Pr}(e)$. (2) then gives $\mathrm{Pr}(d)$ and so $\mathrm{Pr}(e)$.

allows for the satisfaction of desiderata that would otherwise be unsatisfiable. In particular, we will be able to give theories that prove that 'provability-liars' such as $\neg \mathrm{Pr}(c)$ (where c denotes this sentence) are unprovable in the theory in question—without thereby committing ourselves to these sentences *themselves* being provable (which would violate soundness and presumably also consistency). The approaches to provability I will explore are inspired by approaches to truth of Kripke [1975] (§§2.1–2.4), Gupta [1982] (§2.5) and Gaifman [2000a] (§3).

By way of preview, I should say something about the way in which I will think about provability predicates in this paper; specifically, when I will count something as being a provability predicate for a given theory, and thus when I will count a theory as being (at least in part) about its own provability. The basic stance is straightforward: I will take a predicate letter P to be a provability predicate for a theory S just in case it is introduced with that intention, i.e. the intention of being used to mean *provable in* S.[4] There is of course nothing special about provability here: similarly, a predicate letter Q is a natural number predicate, for example, just in case it is introduced with the intention of being used to mean *natural number*. This is certainly *not* to say that the project of the paper is easy, however. It follows from this stance that it is easy to produce some theories that are about their own provability. But the aim of the paper is to produce theories that can prove a whole range of things about their own provability, including a whole range of things about what they cannot prove—and to simply produce *some* theories concerned with their own provability is of course very far from doing this.

This basic view about provability predicates seems very natural, although I will consider objections to it in §§2.2 and 2.4. One of these concerns an alternative view that is sometimes implicitly assumed, but which doesn't seem to have been argued for in any serious way. This is that a theory contains its own provability predicate only if it satisfies the Hilbert-Bernays-Löb derivability conditions (see §2.4). As we will see in §2.4, although there is a lot to be said for the claim that, other things being equal, it is desirable for a theory concerned with its own provability to satisfy these conditions, there is *very little* to be said for that to the effect that any theory concerned with its provability must satisfy these, or indeed the weaker claim that any 'reasonable' theory so concerned must. The comparison with truth will once again be instructive in relation to these claims.

One upshot of the paper will be that characterizations of the significance of Gödel's second theorem must be nuanced in a way that we now realize that char-

[4]For simplicity I focus on the case of predicate letters, but similar remarks apply to compound predicates.

acterizations of the significance of Tarski's theorem on the undefinability of truth must be. Thus, one might initially be tempted to claim that Tarski's theorem (or the argument of the theorem) shows that no reasonable language contains its own truth predicate. We now know, however, that such claims are false. Characterizations of what the theorem shows must be more restricted: it shows that no language meeting a range of—by no means non-negotiable—conditions contains its own truth predicate. Similarly, one might initially be tempted to claim that Gödel's result (or the argument of that result) shows that no reasonable theory concerned with its own provability can prove its own consistency (cf. the quote from Gaifman above). We will see, however, that just as in the case of Tarski's theorem this claim is false. What the result shows is rather that no theory meeting a range of—again, far from non-negotiable—conditions can prove its own consistency. This is still, to be sure, an important fact, but it is significantly weaker than is commonly claimed.

I should, however, mention a concern that one might have about the analogy that I am drawing between proof and truth. Specifically, one might worry that in at least one respect this is misleading. For one might think than an important difference between the two notions is that truth is 'up for grabs', while provability is not; or, at least, formal provability is not, which is what is relevant here. After all, the literature on truth contains a wide variety of non-standard languages that are each proposed as the best way of extending a language to one that contains its own truth predicate. In apparent contrast, if we are given some theory T, it is surely completely fixed what it means for something to be provable in that theory. These two facts are compatible, however. For despite the range of proposals in the literature on truth, once we focus on one of the specific languages from this literature, e.g. the strong Kleene, least fixed point proposal of Kripke [1975], it is then completely fixed what it means for a sentence to be true in this language—just as in the provability case. Conversely, although once we are given a specific theory, it is determined what it means to be provable in this, there is nothing to stop us from producing a wide range of different theories—e.g. with very different axioms and rules—as ways of handling different subject matters. This concern about the analogy would thus seem misplaced.

1.2 Different Approaches

There are three sorts of approaches to modal notions, such as necessity, knowability and provability, that one finds in the literature. Firstly, one can use a predicate that belongs to the language that one starts with to express the notion in the question. For example, one can use a predicate of the language of arithmetic to express

provability in a given theory. This is of course the method that Gödel used in proving the incompleteness theorems. Secondly, one can add to one's initial language a predicate or operator intended to express the relevant notion. This is the sort of approach that one finds in standard textbooks such as Hughes and Cresswell [1996].[5] Thirdly, one can base a modal logic on a predicate in one's initial language, e.g. on a provability predicate of the language of arithmetic. That is, one can consider translations of modal formulas in terms of this predicate (i.e. where $\Box$ is translated as the predicate in question), and ask which modal formulas are such that their translations are always provable in a given theory, or such that these are always true.[6]

The approaches pursued below are of the second sort: we will add a new predicate to the language of the theory that we start with. What is distinctive about the approaches below is simply that the intended interpretation of this new predicate is provability in the theory that is being constructed (i.e. that we are extending our initial theory to).

I should, however, discuss some examples of the first sort of approach that—like those below—are aimed at giving theories that can prove a range of things about their own provability, including their consistency.[7]

One family of approaches use theories with consistency 'built-in'. Thus, rather than a standard theory S, one would use a variant $S^\star$ that has consistency built-in. For example, $S^\star$ might be such that something counts as a proof in $S^\star$ if (i) it is a proof in S of some sentence A, and (ii) there is no shorter (or equally short) proof in S of a sentence B with $A = \neg B$ or $B = \neg A$. Alternatively, $S^\star$ might be such that something counts as a proof in $S^\star$ if it is a proof in S such that the set of axioms of S that are shorter than (or equal to) the longest axiom used in this proof is consistent. As long as S is consistent, $S^\star$ (understood in either way) will prove the same theorems as S. Further, if PA* is such a variant of Peano arithmetic (PA), for example, then it will be able to prove its own consistency (using gödel numbering). However, such approaches yield theories that can prove their own consistency only by apparently trivializing the question. For in the context of PA*, for example, the question of consistency no longer has the significance that it does in that of PA. After all, if U is the theory of Frege's *Grundgesetze* (including basic law V), then $U^\star$ will be consistent, despite being clearly inadequate. Thus, in the context of theories with consistency built-in, the analogue of the question of consistency

[5] See also, e.g., Kaplan and Montague [1960], des Rivières and Levesque [1988] and Halbach, Leitgeb and Welch [2003].

[6] See, e.g., Japaridze and de Jongh [1998] and Artemov and Beklemishev [2005].

[7] For discussion of these approaches, see Detlefsen [1986], Visser [1989] and Feferman [1990].

would seem to be that of whether the theory would be consistent even if it did not have consistency built-in—which is of course just the question of whether the original theory is consistent. And PA* can of course no more answer that question than PA can. For this reason, these approaches do not seem to give a very satisfying solution to the problem of how to give theories that can prove a range of things about their own provability. In contrast, however, the approaches of this paper will not trivialize the question of consistency in any such way.

A distinct but closely related family of approaches would instead simply use non-standard provability predicates. Thus, rather than changing one's theory—i.e. changing the condition that something must satisfy to count as a proof—one just uses a different predicate to talk about this theory. For example, rather than using a predicate of the language of arithmetic that (under the intended interpretation) corresponds to the property of being provable in PA (i.e. expresses an arithmetic property that encodes this property), one would use a predicate of this language that corresponds to a distinct but coextensive property, such as that which a sentence A has if there is some proof of it in PA such that there is no shorter (or equally short) proof of a sentence B with $A = \neg B$ or $B = \neg A$. (This latter property is of course that of being a theorem of PA*, in the first sense mentioned above.) Using such a predicate PA can *in a sense* prove claims about its own provability, including the claim that it is consistent: i.e. PA can prove claims such as $\neg \mathrm{Pr}^\star(\ulcorner 0 = 1 \urcorner)$ and $\forall x \forall y(\mathrm{Neg}(x, y) \rightarrow \neg \mathrm{Pr}^\star(x) \vee \neg \mathrm{Pr}^\star(y))$, where $\mathrm{Pr}^\star(x)$ is a non-standard provability predicate. ($\ulcorner 0 = 1 \urcorner$ is an abbreviation of the gödel numeral of $0 = 1$, and $\mathrm{Neg}(x, y)$ corresponds to the is-the-negation-of-relation.) These approaches seem to change the subject, however: what is being proved is *not* that $0 = 1$ has the property of being unprovable in PA (for example), it is rather that $0 = 1$ has some other property that happens to be coextensive with this one. After all, there is no natural interpretation of the language of arithmetic under which predicates such as Pr* express properties that correspond to the property of being provable in PA, and the axioms and rules of PA would be unnatural given any such non-standard interpretation of this language. In contrast, the approaches below will not change the subject in any such way: the relevant predicates will express the property of being provable in the theory under natural interpretations of the language, and the axioms and rules will be natural given this interpretation.

Finally, there is the possibility of using such a non-standard provability predicate as a means of introducing a new predicate, say Pr, into the language. The basic idea would be to define a translation between sentences of the extended language and those of the original one, where occurrences of Pr are translated by a non-standard provability predicate. One would then extend one's theory by adding an

axiom to the effect that Pr applies to a sentence iff the non-standard predicate applies to its translation. Given the intended interpretation of the base language, Pr would apply to exactly those sentences that are provable in the extended theory. Further, as with the previous family of approaches (in terms of such non-standard predicates), the extended theory could use Pr to prove a range of things about its own provability, including its own consistency. Again, however, there seem to be reasons for preferring the approaches pursued below. Firstly, insofar as the intended interpretation of Pr is provability in the extended theory—rather than provability in some quite different theory that happens to have the same theorems—to define Pr in terms of such a non-standard predicate would seem highly unnatural. Secondly, at least some of these theories will lack an apparently desirable property that many of those considered below (namely, those of §§2.1 and 2.3) possess. Specifically, the property that for any sentence A, A is provable iff $\mathrm{Pr}(\text{`}A\text{'})$ is; and $A$ is refutable (i.e. $\neg A$ is provable) iff $\mathrm{Pr}(\text{`}A\text{'})$ is refutable. In at least some cases, if Pr is introduced by means of a non-standard predicate, then one will not have the second half of this property (i.e. (III) of §2.1 below).[8] It is possible that some theories where Pr is introduced in this way will possess this property, but without some argument to that effect, there does not seem to be any reason to expect this.

2 Classical Approaches

The first approaches I will consider, then, are inspired by the approaches to truth of Kripke [1975]. The basic idea behind these is that one can (at least to a great extent) learn the meaning of 'true' by being told that one may assert (or deny) of a sentence that it is true precisely when one may assert (deny) the sentence itself. Thus, one will see that one may assert "'snow is white' is true', 'some sentence containing 'snow' is true', "'snow is white' is true' is true', and so on. Kripke gives a variety of constructions of languages containing their own truth predicates based on this

[8] In particular, this is true if we use the second sort of non-standard predicate mentioned above (i.e. in terms of consistent sets of axioms). (I am grateful to a referee for this journal for suggesting the following argument.) Suppose that our initial theory is PA, Pr_0 is the standard provability predicate of the language of arithmetic, $\mathrm{Pr}^\star$ is a non-standard predicate of the sort in question, and S is the extension of PA that results from introducing Pr via $\mathrm{Pr}^\star$. S will prove $\neg\mathrm{Pr}(\ulcorner\neg\mathrm{Pr}_0(\ulcorner 0 = 1\urcorner)\urcorner)$: reasoning in S, suppose $\mathrm{Pr}(\ulcorner\neg\mathrm{Pr}_0(\ulcorner 0 = 1\urcorner)\urcorner)$; then $\mathrm{Pr}_0(\ulcorner\neg\mathrm{Pr}_0(\ulcorner 0 = 1\urcorner)\urcorner)$ (since PA proves $\forall x(\mathrm{Pr}^\star(x) \rightarrow \mathrm{Pr}_0(x))$); but then $\mathrm{Pr}_0(\ulcorner 0 = 1\urcorner)$ (by the second incompleteness theorem); and thus $\mathrm{Pr}(\ulcorner\mathrm{Pr}_0(\ulcorner 0 = 1\urcorner)\urcorner)$ (since for any Σ_1-sentence A of the language of arithmetic, PA proves $A \rightarrow \mathrm{Pr}^\star(\ulcorner A\urcorner)$); giving $\mathrm{Pr}(\ulcorner 0 = 1\urcorner)$ and then $0 = 1$. On the other hand, S will *not* prove $\mathrm{Pr}_0(\ulcorner 0 = 1\urcorner)$. That is, we do not have: A is refutable in S iff $\mathrm{Pr}(\ulcorner A\urcorner)$ is (for any sentence A).

idea. I will suggest that broadly similar approaches to provability would also seem to be natural.

The basic idea is simply as follows. Given some initial theory T, one adds a new 1-place predicate letter Pr to the language, intended to mean *provable in* S, where S is the theory that one is extending T to. In S, Pr will be governed by rules that allow one to go from a proof of A to one of $\mathrm{Pr}(\text{`}A\text{'})$, and from a refutation of $A$ (i.e. proof of $\neg A$) to one of $\mathrm{Pr}(\text{`}A\text{'})$ (i.e. a proof of $\neg\mathrm{Pr}(\text{`}A\text{'})$).

I should comment on some features of this way of proceeding. Firstly, the decision to add a new predicate to the language of our base theory T. This follows standard methodology in the theory of truth, but in this context it may come as a surprise. After all, as long as the language of T includes that of arithmetic (or similar), and the theory S we are going to end up with is a formal system, the language of T will already contain a predicate that corresponds to the property of being provable in S. Why then add this new predicate? Simply because doing so allows theories that can prove things about themselves that would be impossible if we stuck to doing things in the language of arithmetic. Secondly, it would in some ways be more natural not to directly add a predicate for provability in S, but rather to add a 2-place predicate $\mathrm{Proof}(x, y)$, intended to mean *x is a proof in S of y*. I will add a 1-place predicate just because this is simpler, but versions of many of the approaches below that instead add a 2-place predicate would also seem possible. Thirdly, I will focus on theories in which (under the intended interpretation) Pr applies to sentences, rather than gödel numbers. This is primarily because my aim is to give natural theories concerned with their own provability, and theories that do this directly, rather than via coding, seem clearly to be more natural. But a secondary reason is that doing things in terms of gödel numbers precludes certain approaches that doing things in terms of sentences allows (e.g. that of §2.5).

2.1 First Classical Approach

I now give the first simple approach along these lines (a more sophisticated one will be given in §2.3). Thus, let L be a first-order language with equality, and let T be a set of sentences of L that is closed under classical consequence. That is, in the main presentation I identify our base theory T with the set of sentences that it proves.[9] However, this is just for simplicity: it is straightforward to give versions of

[9] As in the introduction, I will use 'theory' for a collection of axioms and rules, rather than simply for a set of sentences. But I will also count a set of sentences as a theory, i.e. that whose axioms are the members of the set and which has no rules.

the approaches below that think of T rather as a collection of axioms and rules. Let $\mathcal{L}$ be the extension of L that results from adding the new 1-place predicate letter Pr.

I make the following further assumption about these things.

Assumption 1. (i) For any sentence A of $\mathcal{L}$, 'A' is an individual constant of L.

(ii) There is a classical interpretation of L, $M = \langle D, I\rangle$ (D is the domain and I is the interpretation function), which is the intended interpretation of L.

(iii) For any sentence A of $\mathcal{L}$, $I(\text{'}A\text{'}) = A$.

(iv) There is a formula $B(x)$ of $\mathcal{L}$ that is satisfied in M precisely by the sentences of $\mathcal{L}$. I use $\text{Sent}(x)$ as an abbreviation of this formula.

(v) M is a model of T.

This might seem to be a rather demanding assumption, for two reasons. Firstly, the languages of some mathematical theories will not have intended interpretations in this sense (i.e. with set domains). For example, the language of set theory, assuming that its quantifiers are intended to range over absolutely all sets. As we will see, however, everything that I will say can be extended to such theories. But for simplicity I assume (except when otherwise stated) that T is not such a theory. Secondly, the languages of most mathematical theories do not contain quote-names or formulas that apply to sentences. However, it is straightforward to extend any language, theory and intended interpretation so that they do satisfy assumption 1.[10] Thus, I am in effect simply assuming that we have already done that.

The first way of extending T to a theory that can prove a range of things about its own provability is then as follows.

Definition 1. $\Phi(T)$ is the theory with language $\mathcal{L}$, axioms the members of T and (A1), and rules (RC), (R1) and (R2).

(A1) $\forall x(\text{Pr}(x) \rightarrow \text{Sent}(x))$

(RC) $A_1, \ldots, A_n$ / B if $n \geq 0$ and B is a classical consequence of $A_1, \ldots, A_n$.

(R1) A / $\text{Pr}(\text{'}A\text{'})$

[10] That is, by expanding the language and domain in the obvious way, and restricting the quantifiers in the theory by a new predicate letter intended to apply exactly to the objects in the original domain.

(R2) $\neg A \,/\, \neg \mathrm{Pr}(\text{`}A\text{'})$

That is, a sequence $A_1, \ldots, A_n$ of sentences of $\mathcal{L}$ is a proof in $\Phi(T)$ if for each i ($1 \leq i \leq n$) either $A_i \in T \cup \{(\mathrm{A1})\}$, or A_i results from applying one of (RC), (R1) or (R2) to earlier members of the sequence. For the rest of the paper, I will use S for $\Phi(T)$. If R is a theory with language $\mathcal{L}$, then I use $\vdash_R A$ to mean that A is provable in R; Th_R for the set of theorems of R; I_R for the interpretation function that is just like I except that $I_R(\mathrm{Pr}) = \mathrm{Th}_R$; and M_R for $\langle D, I_R \rangle$.

Is S a natural theory to adopt, given the intended meaning of Pr (i.e. provability in S itself)? In particular, is S sound? That is, do we have $M_S \vDash \mathrm{Th}_S$ (i.e. $M_S \vDash A$ for every $A \in \mathrm{Th}_S$)? If R is a theory in $\mathcal{L}$ with this property, that is, $M_R \vDash \mathrm{Th}_R$, then I say that it is *p-sound* (with respect to M). In fact, it is straightforward to show that S is indeed p-sound. One way of doing this is via a construction from Kripke [1975].[11] However, it is more direct to proceed without going via such a truth-theoretic construction. Further, this argument establishes that S has an important property—(III) below—that does not follow from the Kripkean one (cf. §1.2).

I start by defining sets of sentences as follows.

Definition 2. $X_0 = \varnothing$ and $Y_0 = D - \{\text{sentences of } \mathcal{L}\}$. For any $n \in \mathbb{N}$, X_{n+1} is the set of sentences of $\mathcal{L}$ that are provable in S using (R1) and (R2) at most n times (i.e. in total); and Y_{n+1} is the set of sentences of $\mathcal{L}$ whose negations are so provable, together with the members of Y_0.

If Z and W are disjoint subsets of D, then by $M + \langle Z, W \rangle$ I mean the partial interpretation of $\mathcal{L}$ that extends M by interpreting Pr with $\langle Z, W \rangle$. A sentence A of $\mathcal{L}$ is true (false) in $M + \langle Z, W \rangle$ under the *supervaluationist scheme* if A is true (false) in every classical extension of $M + \langle Z, W \rangle$.

We then have the following.

Lemma 1. *For any $n \in \mathbb{N}$,*

(a) $X_n \cap Y_n = \varnothing$;

(b) *every member of X_{n+1} is true in $M + \langle X_n, Y_n \rangle$ under the supervaluationist scheme.*

[11] Taking Pr as our truth predicate, the members of Th_S are true in Kripke's basic supervaluationist construction. It follows that Th_S is classically consistent, from which it is easy to show that S is p-sound.

Proof. Induction on n. If $n = 0$, then (a) is obvious and (b) is clear given that the members of X_1 are provable without using either (R1) or (R2). So let $n = r + 1$. (a) is clear by the inductive hypothesis. For (b), suppose $A \in X_{n+1}$. There is thus some proof $B_1, \ldots, B_m$ in S that uses (R1) and (R2) at most n times. If there are no uses of (R1) or (R2) in this proof then we are done (as in the $n = 0$ case). So suppose that there is at least one such use, and let B_k result from the last one. Thus for every $j < k$, $B_j \in X_n$. By (b) of the inductive hypothesis, each such B_j is true (under the supervaluationist scheme) in $M + \langle X_r, Y_r \rangle$. Thus since $\langle X_r, Y_r \rangle \leq \langle X_n, Y_n \rangle$,[12] each such B_j is true in $M + \langle X_n, Y_n \rangle$. If B_k results from a use of (R1), then $B_k = \mathrm{Pr}(\text{`}B_j\text{'})$ for some $j < k$. Then since $B_j \in X_n$, B_k is true in $M + \langle X_n, Y_n \rangle$. Further, since B_m is a classical consequence of $T \cup \{(\mathrm{A1}), B_1, \ldots, B_k\}$, B_m must also be true in $M + \langle X_n, Y_n \rangle$. The case where B_k results from a use of (R2) is similar. □

Theorem 1. *S is p-sound.*

Proof. We must show $M_S \vDash \mathrm{Th}_S$. If $A \in \mathrm{Th}_S$, then for some n, $A \in X_{n+1}$. By (b) of lemma 1, A is true in $M + \langle X_n, Y_n \rangle$. But M_S is an extension of $M + \langle X_n, Y_n \rangle$, because $X_n \subseteq \mathrm{Th}_S$, and $Y_n \cap \mathrm{Th}_S = \varnothing$ (by (a) of lemma 1). So A is true in M_S. □

As I noted, however, some mathematical theories will not have intended interpretations in the sense that I have assumed that T has (i.e. (ii) of assumption 1). Can we be sure that S is sound, even if we drop this assumption? Yes, given only the following assumptions: (i), (iii) and (iv) of assumption 1 hold for the intended meaning of L; T is sound for this intended meaning; truth under this intended meaning is preserved by classical consequence; and there is a classical interpretation (i.e. with a set domain) that satisfies (i), (iii–v) of assumption 1. It is then easy to show that every theorem of S is true under the intended meaning of $\mathcal{L}$ (we show that Th_S is consistent as before, from which the claim easily follows). For the rest of the paper I will assume that T does have an intended interpretation in the sense of (ii) of assumption 1. However, any uses that I will make of this assumption can be shown to be inessential along the lines just sketched.

Thus, the soundness of S follows from that of T, and so if we accept T, it would seem reasonable to accept S, too.

S satisfies the minimal conditions that one would expect any extension of T concerned with its own provability to satisfy, i.e. the following.[13,14]

[12] I.e. $X_r \subseteq X_n$ and $Y_r \subseteq Y_n$.

[13] (II) is sometimes called the Kreisel condition: see Kreisel [1953].

[14] (II) on its own *could* be satisfied simply by treating $\mathrm{Pr}(\text{`}A\text{'})$ as a notational variant of A. One might thus worry that it is a toothless requirement. In the context of natural assumptions about S,

(I) S is p-sound.

(II) For any sentence A of $\mathcal{L}$, $\vdash_S \mathrm{Pr}(\text{`}A\text{'})$ iff $\vdash_S A$.

(The left-to-right direction of (II) follows from $M_S \vDash \mathrm{Th}_S$; the right-to-left from the fact that S contains (R1).)

We do not in general have: $\vdash_S \neg\mathrm{Pr}(\text{`}A\text{'})$ iff $\nvdash_S A$. (This is impossible if S is recursively enumerable, contains a reasonable amount of arithmetic, and is consistent.) But we do have the following: as I noted, this is a result that alternative approaches and arguments seem unable to deliver.

(III) For any sentence A of $\mathcal{L}$, $\vdash_S \neg\mathrm{Pr}(\text{`}A\text{'})$ iff $\vdash_S \neg A$.

For the left-to-right direction, suppose $\vdash_S \neg\mathrm{Pr}(\text{`}A\text{'})$. It follows that for some $n$, $\neg\mathrm{Pr}(\text{`}A\text{'}) \in X_{n+1}$. Then by (b) of lemma 1, $\neg\mathrm{Pr}(\text{`}A\text{'})$ is true in $M + \langle X_n, Y_n\rangle$ under the supervaluationist scheme. From which it follows that $A \in Y_n$ and thus $\vdash_S \neg A$. The right-to-left direction is of course immediate from the fact that S contains (R2).

What about 'provability-liars', i.e. sentences A such that $A \leftrightarrow \neg\mathrm{Pr}(\text{`}A\text{'})$ is a classical consequence of $T$? It follows from the classical consistency of $S$ that it proves none of: $A$ (by (R1)), $\neg A$ (by (R2)), $\mathrm{Pr}(\text{`}A\text{'})$ (since it would then prove $\neg A$), and $\neg\mathrm{Pr}(\text{`}A\text{'})$.

Further, unlike approaches that employ non-standard provability predicates (§1.2), the rules of S are entirely natural given the intended interpretation of Pr as provability in S. For, under this interpretation, (A1) holds given that $\mathcal{L}$ is the language of S. (R1) is obviously sound (i.e. it only allows us to prove truths). And the fact that (R2) is sound follows immediately from the fact that S is consistent.

For the rest of the paper, I use $\bot$ for some arbitrary classical logical falsehood of L. We of course have $\vdash_S \neg\mathrm{Pr}(\text{`}\bot\text{'})$. However, S cannot prove its own consistency, i.e. the quantified claim that for any sentence A, either A or $\neg A$ is unprovable (the fact that S cannot prove this follows from (b) of lemma 1). There are slightly more sophisticated approaches that can prove this (§2.3). Before giving such an approach, however, I will do two things: I will make some remarks about ways in which the results of this subsection can be strengthened; and then, in the next subsection, I will consider some objections to the approach given above (versions of which might also be made to the other approaches I will give).

---

however, (II) will place significant further requirements on S. For example, if we assume that S is closed under classical consequence, and proves 'A' $\neq$ 'B' whenever $A \neq B$, then (II) requires that S proves $\exists_n x\mathrm{Pr}(x)$ (for any $n \in \mathbb{N}$), which doesn't follow from the assumptions alone. Similarly, if we assume that S proves $c =$ '$\neg\mathrm{Pr}(c)$', then (II) entails: if S proves $\neg\mathrm{Pr}(c)$ then it is classically inconsistent; which, again, doesn't follow from the assumption alone.

The first way in which the above results can be strengthened is as follows. It is easy to show, given weak assumptions about T, and using essentially the same arguments as before, that S is not only p-sound but also a conservative extension of T (i.e. for any sentence A of L, $\vdash_S A$ only if $\vdash_T A$). Specifically, one can show this as long as for any sentences A and B of $\mathcal{L}$: (a) if $A \neq B$ then $\vdash_T$ 'A' $\neq$ 'B'; and $\vdash_T$ Sent('A'). One can then use versions of the arguments above to show that any model of T can be extended to one of S, which establishes that the latter is a conservative extension of the former.

The second way in which the results can be strengthened is this. If T is given in part by axiom schemes, then—again, given certain assumptions—it can be shown that the above results hold of the extension of S that extends these schemes to $\mathcal{L}$. Specifically, if M contains the natural numbers, then lemma 1 and theorem 1 can (by essentially the same arguments) be seen to hold of the extension of S that contains every instance of induction for $\mathcal{L}$. That is,

$$(A(\mathrm{o}) \wedge \forall n(N(n) \wedge A(n) \rightarrow A(s(n)))) \rightarrow \forall n(N(n) \rightarrow A(n)),$$

where N applies exactly to the natural numbers. Similar remarks apply to all of the other approaches presented below.

2.2 Objections

I will now consider some objections that stem from the observation that although S is sound given the intended interpretation of Pr as provability in S, it is also sound given the interpretation of Pr as truth; specifically, certain varieties of Kripkean truth (i.e. the supervaluationist ones).

The first objection that this might give rise to is as follows. If the theory that I have proposed holds when Pr is interpreted as truth, then by what right do I claim to have added a provability predicate to L, rather than a truth one? But the simple answer is that we have added a provability (rather than truth) predicate because that is the intention with which Pr was introduced. That is, the intended interpretation of Pr is *provable in S* (and not any species of truth), and the intended extension is thus Th_S (rather than the set of sentences that would be true if Pr was interpreted as truth[15]). And, further, the theory that we have proposed for Pr is indeed sound given this interpretation. It is true that it does not *force* this interpretation (in the

[15]These sets will of course in general be distinct. For example, if L includes the language of arithmetic, then the set of truths of $\mathcal{L}$ when Pr is interpreted as truth will contain either A or $\neg A$ for any sentence A of L. But as long as T is recursively enumerable this will not be the case with Th_S.

sense that it is also sound given certain unintended ones). But it is hard to see why this should be regarded as an objection. After all, theories *never* force their intended interpretations. And we know from the Löwenheim-Skolem theorems that most standard mathematical theories do not even force theirs up to isomorphism.[16] We thus seem to be well within our rights in claiming to have added a provability predicate to L.

Nevertheless, the fact that we used theories of truth as our starting point, and have then ended up with a theory that is sound when Pr is interpreted as truth, might make one worry that we have not adopted the axioms and rules that are most natural for *provability*. Rather, one might worry that we have simply adopted those that happen to hold for both provability and truth. It is just a fact, however, that the most natural, general axioms and rules that one might adopt for provability also hold for truth. As I noted in the introduction, this is true not only of (R1) and (R2), and the factivity schema $\mathrm{Pr}(\text{`}A\text{'}) \rightarrow A$, but also of the derivability conditions.[17] The fact that our axioms and rules hold for truth is thus no indication that we have failed to adopt those that are most natural for provability.

But a different objection is as follows. Given that our axioms and rules hold for truth, one might wonder: why bother adding a predicate for provability at all? Can't one get by with just a truth predicate? The most straightforward response to this objection is that one wants a provability predicate (and not just a truth predicate) because one wants to prove claims *about provability* (and not just about truth), and for this one needs a predicate that expresses provability. But to this the objector might reply: if everything that we want to prove about provability also holds of truth, why can't we use a single predicate, i.e. a predicate originally introduced for truth, to express *both* truth and provability? But the answer is of course that many claims that we will want to make about truth do not hold of provability in any reasonable theory. For example, suppose that L includes the language of arithmetic. Then, if $T(x)$ is our truth predicate, and A is a sentence of the language of arithmetic, we will want to assert: $T(\text{`}A\text{'}) \vee T(\text{`}\neg A\text{'})$. However, for no reasonable theory will we want to assert this schema for provability (since no such theory will be able to decide its own Gödel sentence, for example). The fact that S is sound for truth

[16] It is true that some *extensions* of theories fix the interpretation of the new vocabulary, in the sense that this is forced if one takes for granted the interpretation of the old vocabulary. However, given that the interpretation of our language must in general be fixed by something other than laying down a theory, it is hard to see what would motivate any requirement to the effect that the interpretation of 'new' vocabulary must always be so fixed.

[17] Similar points apply to related notions such as that of necessity. For it is also true that most popular modal theories, e.g. S5, are sound when the box is interpreted as *it is true that*.

should not therefore lead us to think that we can get by with a single predicate for both truth and provability.

I should note further that it is straightforward to give a version of the approach above that is *not* sound for truth, if for some reason one wanted to do this. For example, suppose that T is recursively enumerable, and includes Peano arithmetic. Suppose also that L contains a formula $\mathrm{Sent}_L(x)$ that is satisfied in M precisely by the sentences of L, and a formula $\mathrm{Neg}(x, y)$ that is satisfied in M precisely by pairs of the form $\langle \neg A, A \rangle$ for A a sentence of $\mathcal{L}$. One could then extend T to a theory S' that adds to S an axiom to the effect that some sentence of L is neither provable nor refutable (i.e. $\exists x \exists y(\mathrm{Sent}_L(x) \wedge \mathrm{Neg}(y, x) \wedge \neg \mathrm{Pr}(x) \wedge \neg \mathrm{Pr}(y))$). This theory is not of course sound for truth, but it straightforward to show that it is sound for provability in S'.

2.3 Second Classical Approach

I make the following additional assumption about L and M.

Assumption 2. (i) D contains every finite set of sentences of $\mathcal{L}$.

(ii) There is a formula $\in(x, y)$ of L that is satisfied in M precisely when x and y are assigned objects d and e, respectively, such that $d \in e$.

(iii) There is a formula $\mathrm{Conseq}(x, y)$ of L that is satisfied in M precisely when x is assigned a set of sentences of $\mathcal{L}$ and y is assigned a sentence of $\mathcal{L}$ that is a classical consequence of this set.

As before, any mathematical theory R in language K with intended interpretation N can easily be extended so as to meet these assumptions.

(In the following, I abbreviate formulas containing $\in$ in the usual way.)

Definition 3. $\Psi(T)$ is the theory with language $\mathcal{L}$ and the axioms and rules of S, i.e. $\Phi(T)$, together with the following axiom.

(A2) $\forall x \forall y(\mathrm{Conseq}(x, y) \wedge \forall z \in x \mathrm{Pr}(z) \rightarrow \mathrm{Pr}(y))$

I will use U for $\Psi(T)$. As before, Pr is intended to mean *provable in* U. What exactly U can prove will depend on what T can, however. Thus, I will make the following, not completely precise, assumption about T.

Assumption 3. T can prove any claim about sentences of $\mathcal{L}$, finite sets of sentences of $\mathcal{L}$, and classical consequence in $\mathcal{L}$ that is provable in Peano arithmetic via gödel numbering and standard techniques.

For example, I will assume that L contains a formula $\mathrm{Neg}(x, y)$ applying precisely to pairs $\langle \neg A, A \rangle$ for A a sentence of $\mathcal{L}$; that T can prove that every sentence of $\mathcal{L}$ has a negation; that it can prove that for any sentence there is a set containing precisely that sentence and its negation; and so on. Since assumption 3 is imprecise, one could if one preferred simply assume that T contains Peano arithmetic, and do everything in terms of gödel numbers. But, for the reasons stated above, I will persist with doing things in terms of sentences.

Given assumption 3, U will prove the following.

(3) $\forall x \forall y(\mathrm{Neg}(x, y) \rightarrow \neg \mathrm{Pr}(x) \vee \neg \mathrm{Pr}(y))$

For T will prove that for any sentence A of $\mathcal{L}$ there is a set X containing precisely A and its negation; and it will prove $\mathrm{Conseq}(X, \text{`}\bot\text{'})$; but $\vdash_U \neg \mathrm{Pr}(\text{`}\bot\text{'})$ (by (R2)); and so from (A2) we have $\exists z \in X \neg \mathrm{Pr}(z)$; i.e. either A or its negation is unprovable in U. (3) is of course the claim that U is consistent. Similarly, using $\mathrm{Triv}(x)$ for $\forall y(\mathrm{Sent}(y) \rightarrow \mathrm{Conseq}(x, y))$, U proves the following.

(4) $\forall x(\mathrm{Triv}(x) \rightarrow \exists y \in x \neg \mathrm{Pr}(y))$

It will also prove that all logical truths are provable, i.e. $\forall x(\forall y \mathrm{Conseq}(y, x) \rightarrow \mathrm{Pr}(x))$, that conjunctions are provable iff their conjuncts are, i.e.

$$\forall x \forall y \forall z(\mathrm{Conj}(x, y, z) \rightarrow (\mathrm{Pr}(x) \leftrightarrow \mathrm{Pr}(y) \wedge \mathrm{Pr}(z))),$$

that instances of provable universally quantified claims are provable, i.e.

$$\forall x(\mathrm{Pr}(x) \rightarrow \forall y(\mathrm{Inst}(y, x) \rightarrow \mathrm{Pr}(y))),$$

that a sentence that entails $\bot$ when combined with something provable is refutable, i.e.

$$\forall x(\exists y(\mathrm{Pr}(y) \wedge \mathrm{Conseq}(\{x, y\}, \text{`}\bot\text{'})) \rightarrow \forall y(\mathrm{Neg}(y, x) \rightarrow \mathrm{Pr}(y))),$$

and so on.

Further, it is straightforward to show that U is p-sound, essentially as we did with S. To this end, we define sets as follows.

Definition 4. X_0 is the set of sentences of $\mathcal{L}$ that are classical logical truths. For any $n \in \mathbb{N}$, X_{n+1} is the set of sentences that are provable in U using (R1) and (R2) at most n times (in total).

By the *CC-supervaluationist scheme* I mean that which restricts attention to classical interpretations of $\mathcal{L}$ in which Pr is assigned a set of sentences of $\mathcal{L}$ that is classically consistent, and closed under classical consequence. The following are then proved similarly to lemma 1 and theorem 1. (The fact that we are using the CC-supervaluationist scheme, and thus restricting attention to classically consistent interpretations of Pr, eliminates the need for a non-empty anti-extension in (b) of lemma 2.)

Lemma 2. *For any $n \in \mathbb{N}$,*

(a) *X_n is classically consistent and closed under classical consequence;*

(b) *every member of X_{n+1} is true in $M + \langle X_n, \varnothing \rangle$ under the CC-supervaluationist scheme.*

Theorem 2. *U is p-sound.*

(II) and (III) above will also be satisfied. (II) follows from (R1) and $M_U \vDash \mathrm{Th}_U$. The right-to-left direction of (III) follows from (R2). For the left-to-right direction suppose $\vdash_U \neg\mathrm{Pr}(\text{`}A\text{'})$. Then for some $n$, $\neg\mathrm{Pr}(\text{`}A\text{'})$ is true in $M + \langle X_n, \varnothing \rangle$ under the CC-supervaluationist scheme (by (b) of lemma 2). It follows that $X_n \cup \{A\}$ is classically inconsistent, and thus $\vdash_U \neg A$.

If A is a provability-liar, then as with S U will prove none of: A, $\neg A$, $\mathrm{Pr}(\text{`}A\text{'})$ and $\neg\mathrm{Pr}(\text{`}A\text{'})$.

U thus seems to be a natural theory that can prove a wide range of claims about what it can and cannot prove, including the claim that it is consistent. We began with the question: given some theory that we accept, is there some natural, generally applicable way of extending this to a theory that can prove a range of things about its own provability, including its consistency? And I will consider further ways of extending theories below. But we seem already to have done enough to establish that the answer is 'yes'.

2.4 Claims from Gödel's Result

As I noted in the introduction, however, discussions of Gödel's second incompleteness theorem typically contain claims that require that the answer must be 'no'. (For example, that of Gaifman quoted there.) These claims must thus be false.

But what is the source of these mistakes? What Gödel showed was that if T is a formal system that contains a certain amount of arithmetic, then it can prove coded versions of claims about what it can prove. He showed further that if T is consistent,

then it will not *in this way* be able to prove its own consistency. Why think that this justifies the claim that the answer to our question is 'no', or claims along the lines of 'no reasonable theory can prove its own consistency'? The thought (as I said in the introduction) is that the argument of Gödel's result will apply not only to the proof of coded consistency claims, but to any alternative way of proving things about what is provable in a given theory—or at least to any reasonable such. The approach of §2.3, however, would seem to show that such thoughts are mistaken.

This point should though be discussed in greater depth. The argument of Gödel's result certainly does show that there are limits on what a theory can prove about itself. In particular, it shows that if R is a set of sentences of $\mathcal{L}$, which is closed under classical consequence and classically consistent, then it cannot satisfy all of the following.[18]

(5) There is a sentence G of $\mathcal{L}$ such that $\vdash_R G \leftrightarrow \neg\mathrm{Pr}(\text{`}G\text{'})$.

(D1) For any sentence A of $\mathcal{L}$, if $\vdash_R A$ then $\vdash_R \mathrm{Pr}(\text{`}A\text{'})$.

(D2) For any sentences A and B of $\mathcal{L}$, $\vdash_R \mathrm{Pr}(\text{`}A\text{'}) \wedge \mathrm{Pr}(\text{`}A \rightarrow B\text{'}) \rightarrow \mathrm{Pr}(\text{`}B\text{'})$.

(D3) For any sentence A of $\mathcal{L}$, $\vdash_R \mathrm{Pr}(\text{`}A\text{'}) \rightarrow \mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}A\text{'})\text{'})$.

(6) $\vdash_R \neg\mathrm{Pr}(\text{`}\bot\text{'})$

(D1–D3) are (versions of) the (Hilbert-Bernays-Löb) derivability conditions. U satisfies (D1) and (D2), and so by this argument it cannot satisfy (D3) (assuming

[18] The argument is as follows. Suppose R satisfies (5), (D1–D3) and (6). We then have the following.

1.	$\vdash_R G \rightarrow (\mathrm{Pr}(\text{`}G\text{'}) \rightarrow \bot)$	(5)
2.	$\vdash_R \mathrm{Pr}(\text{`}G \rightarrow (\mathrm{Pr}(\text{`}G\text{'}) \rightarrow \bot)\text{'})$	1 and (D1)
3.	$\vdash_R \mathrm{Pr}(\text{`}G\text{'}) \rightarrow \mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}G\text{'}) \rightarrow \bot\text{'})$	2 and (D2)
4.	$\vdash_R \mathrm{Pr}(\text{`}G\text{'}) \rightarrow (\mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}G\text{'})\text{'}) \rightarrow \mathrm{Pr}(\text{`}\bot\text{'}))$	3 and (D2)
5.	$\vdash_R \mathrm{Pr}(\text{`}G\text{'}) \rightarrow \mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}G\text{'})\text{'})$	(D3)
6.	$\vdash_R \mathrm{Pr}(\text{`}G\text{'}) \rightarrow \mathrm{Pr}(\text{`}\bot\text{'})$	4 and 5
7.	$\vdash_R \neg\mathrm{Pr}(\text{`}G\text{'})$	(6) and 6
8.	$\vdash_R G$	7 and (5)
9.	$\vdash_R \mathrm{Pr}(\text{`}G\text{'})$	8 and (D1)

That is, R is classically inconsistent. See e.g. Smith [2013].

that T is sufficiently strong to ensure that U satisfies (5)). When it is assumed that Gödel's argument will apply to any reasonable theory concerned with its own provability, what is being assumed is essentially that any such theory will satisfy (5) and (D1–D3). But why make this assumption? In particular, why assume that any such theory will satisfy (D3)? Well, one might argue as follows. Any reasonable theory concerned with its own provability should be such that if it proves something, then it proves that it proves it (i.e. it satisfies (D1)). But then all of the conditionals in (D3) will be true, and will express important facts about the theory. Thus, if the theory is adequate, it will prove them!

This is a good argument for the claim that, other things being equal, it is desirable for a theory concerned with its own provability to satisfy (D3). But it certainly does not follow that any reasonable such theory will satisfy (D3). For we are simply in an area where we cannot get everything that we want: prima facie, we would also like a theory that can prove its own factivity, i.e. every instance of $\mathrm{Pr}(\text{`}A\text{'}) \rightarrow A$; but we know that this is not going to be possible if certain other perhaps even more desirable conditions are satisfied. We would also like a theory that satisfies (6), but in the presence of the other conditions above that means that (D3) cannot be satisfied. Given the approach of §2.3, a more accurate description of the situation would seem to be as follows. There are reasonable theories that can prove of a range of things about their own provability, including their consistency. Indeed, any given theory that we might accept can be extended to one. There are, to be sure, certain prima facie desirable conditions that these theories will *not* satisfy. But that is *always* going to be true of theories concerned with their own provability, as a range of arguments, including not only that of Gödel's second theorem but also those of the introduction, show.

The analogy with truth is once again instructive here. The conditionals in (D3) are analogues of special cases of one direction of the truth schema (i.e. $T(\text{`}A\text{'}) \rightarrow T(\text{`}T(\text{`}A\text{'})\text{'})$). And one can give an argument just like that above for the claim that any reasonable language containing its own truth predicate must be such that all of these conditionals are true in it, as follows. Surely (the argument would begin) any such language must be such that if a sentence A is true in it, then so is the sentence that says this. But that is just what the conditionals in question say! And so (as in the provability case) these seem to express important facts about the language in question. Thus (one would conclude) if the language is adequate, then surely these conditionals will be true in it. Again, this is a good argument for the claim that—other things being equal—an approach on which these conditionals are true is desirable. It is just that other things are *not* equal, and many of the most important approaches to truth do not deliver the truth of these conditionals: for

example, the approaches of Kripke [1975], Gupta [1982], Herzberger [1982], Gupta and Belnap [1993] and Maudlin [2004].

To illustrate,the parallel in the case of Kripke's approaches is striking. For whenever a sentence A is true in one of Kripke's languages, so is $T(\text{`}A\text{'})$. It would seem, therefore, that what the conditionals say is the case, and that they express important facts about these languages. But despite this the conditionals are not true in these languages. The standard reaction to such limitations of Kripke's approaches, however, is not at all to stick to the claim that no reasonable language can contain its own truth predicate. Rather, it is to recognize that these approaches are natural ways of extending any given language to one that does contain such a predicate, even if these language do not give us *everything* that we might want. I would suggest that our response to limitations in the provability case that are so clearly analogous should be similar.

2.5 Third Classical Approach: Restricted Self-Reference

I will briefly consider one final classical approach. For it should be pointed out that there are in fact theories that satisfy all of the derivability conditions, are closed under classical consequence and classically consistent, *and* prove their own consistency. These are theories with restricted self-reference: e.g. that do not contain resources sufficient to generate provability-liars. Such restrictions in no way limit how much arithmetic is contained (but they do mean that if the theory's language contains that of arithmetic, then one cannot use of gödel numbers in place of sentences). What these restrictions limit are of course syntactic resources (e.g. one cannot have resources sufficient to express a diagonal function for $\mathcal{L}$). So this is not an approach that will show how to extend *any* base theory. But it will show that given any theory not concerned with the sentences of $\mathcal{L}$ (e.g. any standard mathematical theory), one can extend this to a theory that satisfies all of the derivability conditions and proves its own consistency. This would seem to make even clearer how wide of the mark typical characterizations of the significance of Gödel's second theorem are.

In this subsection I maintain assumption 1 but drop 2 and 3. Instead, I make the following. ($\text{SENT}(\mathcal{L})$ is the set of sentences of $\mathcal{L}$.)

Assumption 4. (i) The connectives of L are $\neg$ and $\rightarrow$.

(ii) L contains a 2-place predicate letter Neg such that $I(\text{Neg}) = \{\langle \neg A, A\rangle : A \in \text{SENT}(\mathcal{L})\}$, and a 3-place predicate letter Cond such that $I(\text{Cond}) = \{\langle A \rightarrow B, A, B\rangle : A, B \in \text{SENT}(\mathcal{L})\}$.

Assumption 5. (i) For any individual constant c of L not of the form 'A' for some $A \in \mathrm{SENT}(\mathcal{L})$, $I(c) \notin \mathrm{SENT}(\mathcal{L})$.

(ii) For any n-place predicate letter P of L distinct from =, Neg and Cond, and any $d_1, \ldots, d_n, d_i' \in D$ $(1 \leq i \leq n)$, if $d_i, d_i' \in \mathrm{SENT}(\mathcal{L})$, then: $\langle d_1, \ldots, d_i, \ldots, d_n \rangle \in I(P)$ iff $\langle d_1, \ldots, d_i', \ldots, d_n \rangle \in I(P)$.

(iii) For any n-place function symbol f of L, the range of $I(f)$ is disjoint from $\mathrm{SENT}(\mathcal{L})$; and for any $d_1, \ldots, d_n, d_i' \in D$ $(1 \leq i \leq n)$, if $d_i, d_i' \in \mathrm{SENT}(\mathcal{L})$, then $I(f)(d_1, \ldots, d_i, \ldots, d_n) = I(f)(d_1, \ldots, d_i', \ldots, d_n)$.

We then have the following (Gupta [1982: 9–19]). ($M + X$ is the extension of M that assigns Pr the interpretation X.)

Theorem 3. *There is a unique* $X \subseteq \mathrm{SENT}(\mathcal{L})$ *such that for any* $A \in \mathrm{SENT}(\mathcal{L})$, $M + X \vDash \mathrm{Pr}(\text{`}A\text{'}) \leftrightarrow A$.

What follows for our purposes? To answer this, consider the following extension of T.

Definition 5. $\Theta(T)$ is the theory with language $\mathcal{L}$ and the axioms and rules of S, i.e. $\Phi(T)$, together with the following axioms.

(A3) $\mathrm{Pr}(\text{`}A\text{'}) \wedge \mathrm{Pr}(\text{`}A \to B\text{'}) \to \mathrm{Pr}(\text{`}B\text{'})$

(A4) $\mathrm{Pr}(\text{`}A\text{'}) \to \mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}A\text{'})\text{'})$

(A5) $\forall x \forall y(\mathrm{Neg}(x, y) \to \neg \mathrm{Pr}(x) \vee \neg \mathrm{Pr}(y))$

I will use V for $\Theta(T)$. Of course, Pr is intended to mean *provable in* V. We have the following.

Theorem 4. *V is p-sound.*

Proof. Let X be as in theorem 3. It is easy to see $M + X \vDash \mathrm{Th}_V$, and thus that Th_V is classically consistent. It is then straightforward to show $M_V \vDash \mathrm{Th}_V$ (using an argument similar to those given above). □

V also satisfies (II) (the left-to-right direction follows from theorem 4, and the right-to-left is clear given (R1)).[19] And it clearly satisfies the derivability conditions

[19] V satisfies the right-to-left direction of (III) by (R2), but I do not know whether it satisfies the left-to-right direction.

(D1–D3). The drawback is of course that V can only talk about the syntax of its language in a rather limited way. But this approach does seem to show that even if we insist on (D1–D3), we can still have natural theories that can prove a range of things about their own provability, including their consistency.

3 Non-Classical Approaches

In §2, we considered various theories that can prove a range of things about their own provability, including in some cases their consistency. However, while these theories could prove of many things that they are unprovable, they could not prove this of provability-liars, e.g. a sentence $\neg\mathrm{Pr}(c)$ with $\vdash_T c = \text{`}\neg\mathrm{Pr}(c)\text{'}$. One might thus wonder: even if a variety of natural theories can prove their own consistency, is the unprovability of such sentences an important fact that no such theory can prove about itself? That is, even if 'self-reflection' can encompass consistency, is the unprovability of such sentences always beyond its reach? In this section, I will argue that the answer is 'no', by considering an extension of T that can prove that such sentences are unprovable. For reasons of space, I will consider only one example of a general sort of approach, but I hope that this will be sufficient to illustrate the more general idea.

The fact that the theories of §2 cannot prove that provability-liars are unprovable is analogous to the fact that in the languages of Kripke [1975] one cannot truthfully assert that Liar sentences are untrue. There are, however, approaches to truth that augment Kripke's so as to remedy this. For example, an approach of Gaifman [2000a], which adds to Kripke's a more nuanced treatment of paradoxical sentences, including Liar sentences.[20]

This approach starts with the following idea about Liar tokens. Suppose that at some time t Gottlob utters 'what Gottlob says at t is not true'. If one tries to work out whether this utterance is true, one is sent to consider whether the utterance it is *about* is true, but that is of course just the utterance that we started with. Gottlob's utterance is thus a 'loop' in this sense. Consequently (the idea continues) it is neither true nor false. On the other hand, suppose that after reflecting on all this Bertrand utters 'what Gottlob says at t is not true'. This second utterance is *not* a

[20] I should note that the approach of Gaifman [2000a] that I will describe is mentioned only in passing in that work. The approaches that are the main focus there are concerned exclusively with truth for tokens, whereas that which I will describe is also concerned with truth for types (i.e. sentences). For another approach that augments Kripke's in a broadly similar way, see Skyrms [1984].

loop: if one tries to work out whether it is true, one is sent to consider Gottlob's utterance—one isn't sent back to consider Bertrand's. It thus seems plausible that Bertrand's utterance, unlike Gottlob's, is simply true.

Note that the idea is *not* that 'true' is context sensitive (in the way that 'yours' is, for example), expressing different properties in Gottlob's and Bertrand's utterances. After all, there do not appear to be two properties, or two contents, to express here. The idea is rather that the explanation of the fact that Gottlob's utterance is not true, while Bertrand's is, is *simply* that the former is a loop, while the latter isn't.

And a similar approach to Liar *sentences* would also seem to be natural. Thus, if c denotes $\neg T(c)$, then this sentence would be neither true nor false (like Gottlob's utterance). But if b is a distinct name of the same sentence (i.e. $\neg T(c)$), then $\neg T(b)$ would simply be true (like Bertrand's utterance). For these sentences seem to be 'structurally' just like the utterances: $\neg T(c)$ is a 'loop' (saying of itself that it is untrue) while $\neg T(b)$ is not. One would thus have exceptions to classical logic: because $\neg T(b)$ and $b = c$ will be true (the latter is a simple identity, after all), but $\neg T(c)$ will not be. However, these exceptions flow naturally from what would seem to be a plausible treatment of Liar and related sentences. Gaifman [2000a] shows how to develop an approach to truth on which Liar tokens and sentences, and related tokens and sentences, are handled along these lines. Note that, although on this approach there will be exceptions to classical logic in the sense that there will be classically valid arguments with true premises but untrue conclusions, there will certainly *not* be exceptions in the sense of there being such arguments with true premises and *false* conclusions.

If we want an approach to provability on which we can prove that analogous sentences cannot be proved then a similar approach would seem to be natural. On such an approach, if our base theory T proves $c = \text{`}\neg\mathrm{Pr}(c)\text{'}$, for example, then $\neg\mathrm{Pr}(c)$ will not be provable (in the extended theory concerned with its own provability). However, if b is a distinct name with $\vdash_T b = c$, then $\neg\mathrm{Pr}(b)$ will be provable. We would thus have exceptions to classical logic in the sense that certain sentences will be provable (in our theory), even though certain classical consequences of them will not be. As I said in the introduction, the idea is not that classical logic fails to be truth-preserving in such cases. It is simply that the most natural way of satisfying certain desiderata (such as being able to prove that provability-liars cannot be proved) would seem to be by allowing such exceptions to classical logic. Further, as in the truth case, we will have exceptions to classical logic only in the sense of there being classical consequences of things that we can prove that we cannot prove; we will never be able to refute, i.e. prove the negations of, such classical consequences.

How might one actually give such a theory? Proofs that use the rule (R1) first prove a sentence A, and then, on this basis, prove $\mathrm{Pr}(\text{`}A\text{'})$. Similarly, proofs that use (R2) first prove $\neg B$, and then prove $\neg\mathrm{Pr}(\text{`}B\text{'})$. The process by which we will prove that provability-liars cannot be proved must of course be different. The natural way to conceive of this is as follows. (Here $\neg\mathrm{Pr}(c)$ is such that $\vdash_T c = \text{`}\neg\mathrm{Pr}(c)\text{'}$.) One first 'classifies' $\neg\mathrm{Pr}(c)$ as unprovable, and then, on this basis, proves $\neg\mathrm{Pr}(\text{`}\neg\mathrm{Pr}(c)\text{'})$. The first step is constituted by a line in the proof consisting not of a sentence, but a pair of a sentence together with the letter 'e', i.e. $\langle\neg\mathrm{Pr}(c), \mathrm{e}\rangle$. ('e' is for 'exception', since $\neg\mathrm{Pr}(c)$ will lead to exceptions to classical logic in the sense described.) The reason for having this line $\langle\neg\mathrm{Pr}(c), \mathrm{e}\rangle$, rather than proving $\neg\mathrm{Pr}(\text{`}\neg\mathrm{Pr}(c)\text{'})$ more directly, is that it will be useful to keep track of which sentences have been classified as unprovable in this way (because we do not want to prove them, even once we have proved things that have them as classical consequences).

3.1 A Non-Classical Approach

I will now give a simple version of this sort of approach. For the rest of the paper I make assumptions 1–3 (i.e. all of the assumptions of §2 except those of §2.5 on restricted self-reference). And I adopt the following definition.

Definition 6. A sentence A of $\mathcal{L}$ is a *loop* if either $A \leftrightarrow \mathrm{Pr}(\text{`}A\text{'})$ or $A \leftrightarrow \neg\mathrm{Pr}(\text{`}A\text{'})$ is a classical consequence of T.

That is, on the approach to be given it will be such sentences that are classified as unprovable (or 'exceptions'). The aim of this definition is just to illustrate the general idea behind this sort of approach. Other approaches of this sort would of course define loops in more sophisticated ways.

Definition 7. $\Delta(T)$ is the theory with language $\mathcal{L}$, axioms the members of T, (A1), (A6) and (A7), and rules (RE), (R1) and (R3).

(A6) $\forall x(\mathrm{Triv}(x) \rightarrow \exists y \in x \neg\mathrm{Pr}(y))$

(A7) $\langle A, \mathrm{e}\rangle$ if A is a loop.

(RE) $\alpha_1, \ldots, \alpha_n \,/\, B$ if $n \geq 0$, B is a classical consequence of $\mathrm{SENT}(\mathcal{L}) \cap \{\alpha_1, \ldots, \alpha_n\}$, and $\langle B, \mathrm{e}\rangle$ is not one of $\alpha_1, \ldots, \alpha_n$.

(R3) $\langle A, \mathrm{e}\rangle \,/\, \neg\mathrm{Pr}(\text{`}A\text{'})$

I will use W for $\Delta(T)$. I should say something to explain this choice of axioms and rules. Firstly, I should stress that a line of the form $\langle A, \mathrm{e}\rangle$ is simply a way of marking the fact that A has been classified as unprovable. Such a pair is not a formula of the language, and so cannot be combined with connectives, for example, to form larger formulas such as $\langle A, \mathrm{e}\rangle \wedge B$. Similarly, such pairs do not belong to Th_W (which is the set of *sentences* that are provable in W).

Secondly, we do not need (R2) simply because (A6) allows us to do without it. (If $\vdash_W \neg A$, then $\vdash_W \mathrm{Pr}(\text{`}\neg A\text{'})$, by (R1), and so by (A6) and assumption 3 we will have $\vdash_W \neg\mathrm{Pr}(\text{`}A\text{'})$.)

Thirdly, we have (A6) rather than (A2), i.e.

$$\forall x \forall y(\mathrm{Conseq}(x, y) \wedge \forall z \in x \mathrm{Pr}(z) \rightarrow \mathrm{Pr}(y)),$$

because the latter will not be true on this approach (given the exceptions to classical logic).[21]

As before, we want to prove that W is p-sound. To this end, let $\alpha_1, \ldots, \alpha_n$ be some proof in W.

Definition 8. Let $m \in \mathbb{N}$ with $0 \leq m \leq n$. $X_m = \{\alpha_i : i \leq m$ and α_i is a sentence$\}$. $Y_m = \{A :$ for some $i \leq m$, $\alpha_i = \langle A, \mathrm{e}\rangle\}$.

By the *C-supervaluationist scheme* I mean that which restricts attention to classical interpretations in which Pr is assigned a classically consistent set of sentences of $\mathcal{L}$.

Lemma 3. *For any $m \leq n$,*

(a) *X_m contains no loops;*

(b) *X_m is classically consistent;*

(c) *if $m < n$, then every member of X_{m+1} is true in $M + \langle X_m, Y_m\rangle$ under the C-supervaluationist scheme.*

Proof. Induction on m. If $m = 0$ then $X_m = \varnothing$, and so (a) and (b) are trivial. For (c): $X_{m+1} = X_1$ contains at most a single sentence of $T \cup \{(\mathrm{A1}), (\mathrm{A6})\}$, and it is easy to see

[21] An alternative version of the approach would have an additional new predicate letter E (for 'exception'). We could then have:

$$\forall x \forall y(\mathrm{Conseq}(x, y) \wedge \forall z \in x \mathrm{Pr}(z) \wedge \neg E(y) \rightarrow \mathrm{Pr}(y)).$$

that any such sentence will be true in $M + \langle X_0, Y_0 \rangle$ (under the C-supervaluationist scheme). So let $m = r + 1$. For (a) suppose that A is a loop with $A \in X_m$. There are two cases, corresponding to those of definition 6. Suppose first that $A \leftrightarrow \mathrm{Pr}(\text{`}A\text{'})$ is a classical consequence of $T$. By (c) of the inductive hypothesis, $A$ is true in $M + \langle X_r, Y_r \rangle$. But this requires $A \in X_r$, contradicting (a) of the hypothesis. Suppose now that $A \leftrightarrow \neg\mathrm{Pr}(\text{`}A\text{'})$ is a classical consequence of T. By (c) of the inductive hypothesis we again have that A is true in $M + \langle X_r, Y_r \rangle$, and thus that $\neg\mathrm{Pr}(\text{`}A\text{'})$ is. This requires that either $X_r \cup \{A\}$ is inconsistent, or $A \in Y_r$. But $X_r \cup \{A\}$ cannot be inconsistent because every member of $X_r \cup \{A\}$ is true in $M + \langle X_r, Y_r \rangle$ (by (c) of the inductive hypothesis). And we cannot have $A \in Y_r$ since this would require A (i.e. α_m) to be either an axiom or derived using (R1) or (R3). But it is easy to see that none of the axioms or sentences that can be derived in this way are such that $A \leftrightarrow \neg\mathrm{Pr}(\text{`}A\text{'})$ is a classical consequence of T. (b) is clear by (c) of the inductive hypothesis. And (c) is clear by the inductive hypothesis together with inspection of the axioms and rules. □

Theorem 5. *W is p-sound.*

Proof. Let $A \in \mathrm{Th}_W$, and let $\beta_1, \ldots, \beta_k$ be a proof of A in W. Let X_k and Y_k be as in definition 8 (but in terms of this proof rather than $\alpha_1, \ldots, \alpha_n$). By (c) of lemma 3, A is true in $M + \langle X_k, Y_k \rangle$. We thus have $M_W \vDash A$ as long as Th_W is classically consistent, and $\mathrm{Th}_W \cap Y_k = \varnothing$. The latter follows from (a) of lemma 3. For the former it is sufficient that if $\gamma_1, \ldots, \gamma_p$ and $\delta_1, \ldots, \delta_q$ are proofs in W, then so is $\gamma_1, \ldots, \gamma_p, \delta_1, \ldots, \delta_q$ (since we would then be done by (b) of lemma 3). But this is clear by inspection of the rules together with (a) of lemma 3. □

W also satisfies (II). It satisfies the right-to-left direction of (III), and the left-to-right direction except for the case when A is a loop such that $A \leftrightarrow \neg\mathrm{Pr}(\text{`}A\text{'})$ is a classical consequence of T. (Failure of this direction of (III) in this case is of course natural and desirable, since we would not want to prove $\neg A$.)

Loops are treated as outlined above. For example, suppose $\vdash_T c = \text{`}\neg\mathrm{Pr}(c)\text{'}$. Then $\vdash_W \neg\mathrm{Pr}(\text{`}\neg\mathrm{Pr}(c)\text{'})$, $\vdash_W c = \text{`}\neg\mathrm{Pr}(c)\text{'}$ but $\nvdash_W \neg\mathrm{Pr}(c)$. Similarly, the following will all be provable in $W$: $\neg\neg\neg\mathrm{Pr}(c)$, $\neg\mathrm{Pr}(c) \wedge A$ (for any $A$ with $\vdash_W A$), and $\neg\mathrm{Pr}(\text{`}\bot\text{'}) \rightarrow \neg\mathrm{Pr}(c)$. We will thus have exceptions not just to the indiscernibility of identicals, but also to double negation elimination, conjunction elimination, modus ponens, and so on. For a slightly different example, suppose that B is $\exists x(x = b \wedge \neg\mathrm{Pr}(x))$, and $\vdash_T b = \text{`}B\text{'}$. Then $\vdash_T B \leftrightarrow \neg\mathrm{Pr}(\text{`}B\text{'})$, and so $\vdash_W \neg\mathrm{Pr}(\text{`}B\text{'})$ and then $\vdash_W b = b \wedge \neg\mathrm{Pr}(b)$. We will thus have an exception to existential generalization; an exception to universal instantiation can be produced

similarly. Although exceptions to all of these rules are perhaps unfamiliar, they are an almost immediate consequence of what would seem to be a natural treatment of provability-liars. After all, we cannot prove $\neg\mathrm{Pr}(c)$ without violating soundness, but there is no comparable obstacle to proving any of the sentences just mentioned.

We also of course have

$$\vdash_W \forall x \forall y(\mathrm{Neg}(x, y) \rightarrow \neg\mathrm{Pr}(x) \vee \neg\mathrm{Pr}(y)).$$

W would thus seem to be a natural extension of T that can prove a range of things about its own provability, including its consistency, and the fact that provability-liars cannot be proved.

There is one further point about W that should be noted, which is that although neither (D2) nor (D3) will hold in general, the instances of these used in the argument of Gödel's second theorem (note 18) *are* satisfied.[22] This suggests that a version of this sort of approach might be possible that would yield theories that can prove not only their consistency, but also conditionals corresponding to—i.e. stating—their rules (in the way in which those in (D2) correspond to modus ponens, and those in (D3) correspond to (R1)). Consideration of such approaches must be left for another time, however.

Conclusion

The above framework and results also suggest a range of more general questions: some technical, some philosophical. The most obvious technical question is: what are the possible p-sound extensions of T? Another is: are there any natural closure properties of these theories?[23] (It is easy to see that they are not closed under unions, intersections or taking a subtheory.) On the philosophical side, the most obvious question would seem to be as follows. Gödel's second incompleteness theorem has been appealed to in a striking variety of philosophical debates, on issues

[22] More precisely, if G is such that $G \leftrightarrow \neg\mathrm{Pr}(\text{`}G\text{'})$ is a classical consequence of T, then the instances of (D2) and (D3) used in that argument are satisfied. These are the following.

$$\vdash_W \mathrm{Pr}(\text{`}G\text{'}) \wedge (\mathrm{Pr}(\text{`}G \rightarrow (\mathrm{Pr}(\text{`}G\text{'}) \rightarrow \bot)\text{'}) \rightarrow \mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}G\text{'}) \rightarrow \bot\text{'}))$$

$$\vdash_W \mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}G\text{'})\text{'}) \wedge (\mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}G\text{'}) \rightarrow \bot\text{'}) \rightarrow \mathrm{Pr}(\text{`}\bot\text{'}))$$

$$\vdash_W \mathrm{Pr}(\text{`}G\text{'}) \rightarrow \mathrm{Pr}(\text{`}\mathrm{Pr}(\text{`}G\text{'})\text{'})$$

[23] Thanks here to a referee for this journal.

from Hilbert's programme to mechanism. We have seen that some of the most basic and widespread philosophical claims made on behalf of this theorem are false. It is thus natural to ask: what light does this shed on such broader uses of the result?

But these questions, too, must wait for future work. What I hope to have established here is simply this: theories can 'self-reflect' to a far greater degree than is commonly supposed.[24]

References

Artemov, S. N. and L. D. Beklemishev. 2005. Provability Logic. In D. M. Gabbay and F. Guenthner (eds), *Handbook of Philosophical Logic, 2nd Edition: Volume 13*: 189–360. Dordrecht: Springer.

Detlefsen, M. 1986. *Hilbert's Program: An Essay on Mathematical Instrumentalism*. Dordrecht: D. Reidel.

Feferman, S. 1990. Introductory Note to 'Some Remarks on the Undecidability Results'. In K. Gödel, *Collected Works: Volume II, Publications 1938–1974*, S. Feferman *et al.* (eds): 281–87. Oxford: Oxford University Press.

Gaifman, H. 2000a. Pointers to Propositions. In A. Chapuis and A. Gupta (eds), *Circularity, Definition, and Truth*: 79–121. New Delhi: Indian Council of Philosophical Research.

—2000b. What Gödel's Incompleteness Result Does and Does Not Show. *Journal of Philosophy* 97: 46–70.

Gupta, A. 1982. Truth and Paradox. *Journal of Philosophical Logic* 11: 1–60.

Gupta, A. and N. Belnap. 1993. *The Revision Theory of Truth*. Cambridge, MA: MIT Press.

Halbach, V., H. Leitgeb and P. Welch. 2003. Possible-Worlds Semantics for Modal Notions Conceived as Predicates. *Journal of Philosophical Logic* 32: 179–223.

Herzberger, H. G. 1982. Notes on Naive Semantics. *Journal of Philosophical Logic* 11: 61–102.

[24]For comments and discussion, I am grateful to Andrew Bacon, George Bealer, Phillip Bricker, Heidi Lockwood, Vann McGee, Agustín Rayo, Marcus Rossberg, Josh Schechter, Zoltán Szabó, Timothy Williamson, the members of the Yale philosophy department, audiences at McGill University, University of Albany, SUNY and the University of Connecticut and an referee for this journal. I am particularly grateful to a second referee for detailed and very useful comments.

Hughes, G. E. and M. J. Cresswell. 1996. *A New Introduction to Modal Logic*. London: Routledge.

Japaridze, G. and D. de Jongh. 1998. The Logic of Provability. In S. R. Buss (ed.), *Handbook of Proof Theory*: 475–546. Amsterdam: Elsevier.

Kaplan, D. and R. Montague. 1960. A Paradox Regained. *Notre Dame Journal of Formal Logic* 1: 79–90.

Kreisel, G. 1953. On a Problem of Henkin's. *Indagationes Mathematicae (Proceedings)* 56: 405–6.

Kripke, S. 1975. Outline of a Theory of Truth. *Journal of Philosophy* 72: 690–716.

Lewis, D. 1983. New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61: 343–77.

Maudlin, T. 2004. *Truth and Paradox: Solving the Riddles*. Oxford: Clarendon Press.

Raatikainen, P. 2015. Gödel's Incompleteness Theorems. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2015 Edition)*. URL = <http://plato.stanford.edu/archives/spr2015/entries/goedel-incompleteness/>.

des Rivières, J. and H. J. Levesque. 1988. The Consistency of Syntactical Treatments of Knowledge (How to Compile Quantificational Modal Logics into Classical FOL). *Computational Intelligence* 4: 31–41.

Skyrms, B. 1984. Intensional Aspects of Self-Reference. In R. L. Martin (ed.), *Recent Essays on Truth and the Liar Paradox*: 119–31. Oxford: Clarendon Press.

Smith, P. 2013. *An Introduction to Gödel's Theorems*. Second edition. Cambridge: Cambridge University Press.

Visser, A. 1989. Peano's Smart Children: A Provability Logical Study of Systems with Built-In Consistency. *Notre Dame Journal of Formal Logic* 30: 161–96.