Deflationism and the Gödel Phenomena[1]

NEIL TENNANT

Any (1-)consistent and sufficiently strong system of first-order formal arithmetic fails to decide some independent Gödel sentence. We examine consistent first-order extensions of such systems. Our purpose is to discover what is minimally required by way of such extension in order to be able to prove the Gödel sentence in a non-trivial fashion. The extended methods of formal proof must capture the essentials of the so-called 'semantical argument' for the truth of the Gödel sentence. We are concerned to show that the deflationist has at his disposal such extended methods—methods which make no use or mention of a truth-predicate.

This consideration leads us to reassess arguments recently advanced—one by Shapiro and another by Ketland—against the deflationist's account of truth. Their main point of agreement is this: they both adduce the Gödel phenomena as motivating a 'thick' notion of truth, rather than the deflationist's 'thin' notion. But the so-called 'semantical argument', which appears to involve a 'thick' notion of truth, does not really have to be semantical at all. It is, rather, a reflective argument. And the reflections upon a system that are contained therein are deflationarily licit, expressible without explicit use or mention of a truth-predicate. Thus it would appear that this anti-deflationist objection fails to establish that there has to be more to truth than mere conformity to the disquotational T-schema.

1. Introduction and background

Deflationism holds that the truth-predicate does no more, essentially, than furnish a stylistic variant for making or endorsing assertions referred to or quantified over. The truth-predicate may be used when one does not wish to go to the trouble of repeating the assertion in question ('What he said was true'). The truth-predicate may also be used when one does not have the exact form of words to hand ('The Pope's next assertion will be true'), or when one wishes to endorse some or all assertions in a potentially infinite list ('Any utterance by any Pope, now or in the future, is or will be true'). One can also use the truth-predicate to endorse an assertion anaphorically, when the sentence in question has been 'identified' only by means of some quantifier

expression. An important example of this final kind will be given presently.

Deflationism has its roots in Ramsey's contention that to assert that φ is true is to do no more than assert φ, unadorned. Truth is not a substantial property whose metaphysical essence could be laid bare. It has no essence; it is as variegated as the grammatical declaratives that would be its bearers. There would therefore appear to be no gap, on the deflationists' view, between claims that are true and assertions that are warranted; or, generally, between *truth*, on the one hand, and, on the other hand, grounds for assertion, or *proof*.

Now, no considerations can be more important and immediately relevant to the assessment of such a view than the Gödel phenomena. Consider, for example, the opening sentence of Michael Dummett's well-known paper 'The Philosophical Significance of Gödel's Theorem':[2]

> By Gödel's theorem there exists, for any intuitively correct formal system for elementary arithmetic, a statement [G] expressible in the system but not provable in it, *which not only is true but can be recognised by us to be true …* (Emphasis added.)

This is a paradigm example of the final kind mentioned above. Here we have a use of the truth-predicate in anaphoric application to a statement (or sentence) falling appropriately in the range of the quantifier expression 'there exists a statement expressible in the system but not provable in it'. One cannot actually assert the statement in question; for the exact form of words is not available in such a context.[3] So one can 'assert' it only by using the truth-predicate in apposition to the noun-phrase in the quantifier expression. (If one employed instead the clumsier 'such that' expression, then one could use the anaphoric pronoun 'it' instead of the noun-phrase: '… there exists a statement expressible in the system but not provable in it, such that it [i.e. the statement in question] is true.')

So, in so much as stating the philosophical crux of Gödel's theorem, Dummett has furnished the kind of use of the truth-predicate that any deflationist would wish to deconstruct. Yet deflationists appear not to have taken the Gödel phenomena into account. These phenomena receive no mention, for example, in the two standard post-Gödelian

---

[2] Dummett 1963.

[3] Of course, the exact form of words is available, in principle, by applying the constructive recipe implicit in the proof of Gödel's incompleteness theorem, so as to produce the undecidable sentence. The context in question here, however, is the philosophical summary of the situation as quoted from Dummett.

references (with a combined total of 159 pages) frequently cited in support of deflationism.[4]

An anti-deflationist case based on the Gödel phenomena has at last been made. One anti-deflationist argument, due to Stewart Shapiro, is based on what might be called the 'objection from non-conservativeness'.[5] A different argument, for an objection in the same spirit, has been raised independently by Jeffrey Ketland.[6] Although I hold no brief for deflationism as a fully satisfactory account of truth, it is nevertheless my purpose here to argue on behalf of deflationism by countering the objections that Shapiro and Ketland have raised. As we shall see presently, they deploy somewhat different arguments for the same anti-deflationist conclusion. Since I oppose that common conclusion, I shall be finding different points on which to disagree with them, within each of their respective arguments.

My counter to both Shapiro and Ketland will amount to this: the deflationist has 'philosophically modest' means for carrying out the so-called 'semantical' argument for the 'truth' of the Gödel sentence. Indeed, that argument can be directly and faithfully regimented in what can readily be seen to be deflationarily licit terms. The 'deep structure' of the reasoning is faithfully preserved on this regimentation. The deflationary version of the semantical argument does not involve any 'circumratiocinations' obscuring or deforming what we would common-sensically take to be the obvious deductive structure, or 'line of thought', of the argument. Once one appreciates this deflationarily licit presentation of the semantical argument, one realizes that to call it a 'semantical' argument is to misrepresent its suasive structure.

2. On the significance of the proposal to be made on behalf of the deflationist

A case will be made below for the following conclusion: the deflationist has properly deflationary means for attaining the insight that the undecidable Gödel sentence for any sound theory of arithmetic is one that ought to be asserted, rather than denied. The case for this conclusion terminates in section 9.

Now, why should this conclusion be important, philosophically? Because, to put it briefly, it flies in the face of established folklore in the

[4] See Field 1986 and Horwich 1990.

[5] Shapiro 1998.

[6] Ketland 1999.

philosophy of logic, mind and mathematics. The current received wisdom, to be encountered in almost every publication on the philosophical significance of Gödel's first incompleteness theorem, is that this celebrated metatheorem shows that 'truth somehow transcends proof'.

The orthodox contention, more precisely, is that despite the Gödelian incompleteness of our formal system *S*, we can come to appreciate that the undecidable Gödel sentence *G* for *S* is *true*. The system *S* (assuming it is consistent) does not furnish a proof for *G*, so we do not convince ourselves of the truth of *G* by working inside the system *S*. Rather, because of the special relationship that *G* bears to the system *S* (by being constructed after a diagonal method that can be glossed as involving 'reference to' proofs-in-*S*), we somehow reflect on that relationship, at a 'higher' level. We conduct the so-called 'semantical argument' (see below for details). But in so providing the semantical argument, this meta-argument[7] now concludes, *we make essential use of the concept of truth itself.*

I think it is fair to say that this account of our 'knowledge of the truth of *G*' is widespread and standard: so much so, indeed, that it is quite surprising that opponents of the deflationary view of truth have taken so long to marshall it in their cause against what they see as the deflationists' altogether too 'thin' notion of truth.

We shall consider here, by way of illustration, two influential examples of a substantialist account of our 'knowledge of the truth of *G*'.

Dummett provides the following context-setting preliminaries for the intuitive semantical argument by employing a particularly substantial or 'thick'-looking, because explicitly relational, predicate of truth-in-a-model. He writes (loc. cit., p. 186):

> A common explanation is as follows. Since [*G*] is neither provable nor refutable, there must be some models of the system in which it is true and others in which it is false. Since, therefore, [*G*] is not true in *all* models of the system, it follows that when we say that we can recognise [*G*] as true we must mean 'true in the *intended* model of the system.' We thus must have a quite definite idea of the kind of mathematical structure to which we intend to refer when we speak of the natural numbers; and it is by reference to this intuitive conception that we recognise the statement [*G*] to be true.

I say that these are 'context-setting preliminaries' for the semantical argument because the latter argument is not itself thereby given. Here

---

[7] Actually, since the semantical argument is often taken to be an argument conducted at the metalevel, the philosophical argument that adduces considerations about the role of truth in the semantical argument itself is being conducted at the meta-metalevel. Nothing important turns on this; I enter this observation only to defend myself against the captious.

now is the semantical argument itself, as Dummett sets it out later (loc. cit., p. 191):

> The statement [G] is of the form $\forall x A(x)$, where each one of the statements $A(0)$, $A(1)$, $A(2)$, … is true: since $A(x)$ is recursive, the notion of truth for these statements is unproblematic.[8] Since each of the statements $A(0)$, $A(1)$, $A(2)$, … is true in every model of the formal system, any model of the system in which [G] is false must be a non-standard model … whenever, for some predicate $B(x)$, we can recognise all of the statements $B(0)$, $B(1)$, $B(2)$, … as true in the standard model, then we can recognise that $\forall x B(x)$ is true in that model. This fact … we know on the strength of our clear intuitive conception of the structure of the model.

One can find many more examples of similar explanations, by other philosophers, of the insight employed by the semantical argument for the truth of the Gödel sentence G for the given system S.[9] The insight is frequently articulated with a somewhat grosser degree of logical analysis than is provided by Dummett. In the quote just given from Dummett, we see the analysis being given quite explicitly, in terms of the relation between a universal numerical quantification and all its numerical instances. But other influential sources have been less explicit than this. For example, in all of Kleene's classic *Introduction to Metamathematics*, revered for its philosophically subdued foundational precision, we find only the following version of the semantical argument (at p. 426):

> … if we suppose the number-theoretic formal system to be consistent, we can recognise that [G] is true by taking into view the structure of that system as a whole, though we cannot recognise the truth of [G] by use only of the principles of inference formalized within that system …

Likewise, in his much-cited 1961 essay 'Minds, Machines and Gödel', John Lucas gives a rather elliptical account of the semantical argu-

---

[8] Dummett means by this that it is unproblematic to hold the principle of bivalence for recursive, or decidable, statements of arithmetic. This is because all their quantifiers are bounded. Hence the determination of a truth-value for such a statement involves at most a finite search among the natural numbers, together with decision-making with respect to effective operations and relations on those finitely many numbers. By 'unproblematic' Dummett definitely did not mean anything like 'deflationarily licit'; such a reading of his 1963 paper would be completely anachronistic.

[9] We note in passing that Crispin Wright has raised doubts as to whether we really are able to 'see' that G is true. His reason for denying this cognitive achievement is that the semantical argument is only a conditional demonstration of G, since it relies on the assumption that the system S is sound—or at least is consistent. And this, according to Wright, is only a 'commitment' of ours, rather than something that can be assumed as a ground for a non-conditional demonstration of the truth of G. See Wright 1993 at pp. 334–7.

ment.[10] This account does not make explicit appeal to the quantificational structure of the Gödel sentence *G*. That quantificational structure would have to be exploited, however, in any faithful and fully detailed regimentation of the elliptical argument presented.[11] Lucas presents it as follows:

> Essentially, we consider the formula which says, in effect, 'This formula is unprovable-in-the-system'.[12] … the formula 'This formula is unprovable-in-the-system' is … unprovable-in-the-system. Further, if the formula 'This formula is unprovable-in-the-system' is unprovable-in-the-system, then it is true that that formula is unprovable-in-the-system, that is, 'This formula is unprovable-in-the-system' is true.

Of all the accounts of how one comes to see the truth of the Gödel sentence, Dummett's is the most sensitive to its actual logical structure. The reasoning, in essence, runs as follows.

Semantical argument for the truth of the Gödel sentence:

G is a universally quantified sentence (as it happens, one of Goldbach type, that is, a universal quantification of a primitive-recursive predicate). Every numerical instance of that predicate is provable in the system *S*. (This claim requires a subargument exploiting Gödel-numbering and the representability in *S* of recursive properties.) Proof in *S* guarantees *truth*. Hence every numerical instance of *G* is *true*. So, since *G* is simply the universal quantification over those numerical instances, it too must be *true*.

Throughout this reasoning, philosophers standardly take the concept of truth involved to be substantial, or 'thick'.[13] Dummett, as we have seen,

---

[10] Lucas 1961 p. 44.

[11] Other examples of this more elliptical version of the semantical argument can be found in the influential early book *Gödel's Proof* (Nagel and Newman 1958); and also in the later work of two influential popularizers, the first of whom, interestingly, opposes Lucas's conclusion that minds transcend machines, and the second of whom supports it. Both of them agree, however, on the structure of the semantical argument being more or less as Lucas presents it. See Hofstadter 1979, at pp. 271–2 and p. 448; and Penrose 1989, at p. 140.

[12] As we shall see below, the meaning of 'says, in effect' is to be explicated by appeal to the interdeducibility, within the system, of the sentence *G* in question with a syntactically distinct sentence that says, via coding and representability, that every number is not a code number of a proof of *G*. That is why *G* has the form $\forall x A(x)$, where $A(x)$ is a primitive recursive predicate.

[13] Indeed, Christopher Peacocke has even gone so far as to argue—erroneously, in my view—that the semantical argument can go through only on a realist (not just: substantialist) conception of truth. See his paper 'Proof and Truth', section 4, 'Gödel's Theorem: A Problem for Constructivism', (Peacocke 1993). For criticism of Peacocke's argument, see Wright, loc. cit., pp. 343–9. Compare also Michael Detlefsen's passing claim that 'the usual argument for the truth of [*G*] is an argument for its classical truth' (Detlefsen 1986 p. 163). The 'usual' argument to which Detlefsen

is no exception; and is in distinguished and influential company in this regard.

The writers quoted above were writing at a time when (despite the much earlier work of Ramsey) deflationism about truth was not a matter that pressed itself upon the consciousness of the philosophical community. The result, I think, is that it is fair—without holding these sophisticated thinkers to have been in any way naïve—to say that a certain interpretive dogma had managed to establish itself in response to the Gödel phenomena:

> **The substantialist dogma**
>
> The way in which the semantical argument establishes the truth of the Gödel sentence requires that the notion of truth be substantial.

3. The main aim of this paper

What I set out to show here is that this philosophical consensus that has built up around the Gödel phenomena can be seen, upon closer examination, to be mistaken. In our attempt to overcome the inevitable incompleteness of whatever formal system we are using, by deciding its Gödel sentence one way or the other, we do not need to invoke or deploy any 'thick' notion of truth. The Substantialist Dogma is false.

The technical results that are available to support this philosophical conclusion are of reasonable vintage;[14] yet somehow the philosophical community has neglected, thus far, to draw upon them (but see n. 46).

My main purpose in this paper will have been achieved if I can simply show that there is a 'deflationary way' of faithfully carrying out the so-called 'semantical argument' for the truth of the independent Gödel sentence. It has to avoid ellipses and glosses, and be highly regimented—yet also be, and manifestly so, a formal homologue of the semantical argument as it is properly understood. The paper is centrally occupied with drawing the correct and appropriate philosophical lesson from pre-existing foundational work by logicians. Because the technical details are rather delicate, we have to be very clear and precise about them, and present them in a way that makes them relevant to the philosophical considerations being entered.

---

was referring is what I have been calling the elliptical form of the argument. But by 'classical', in this context, I think Detlefsen meant to be stressing the correspondence aspect of truth—its substantiality—and not its bivalence.

[14] The relevant foundational results were first made available by Solomon Feferman, only one year before the paper of Dummett cited above. See Feferman 1962.

I am aware that I am proposing but one possible response among many that might issue from, or on behalf of, deflationists against those who think truth is 'thick'. Deflationism, in all its ascetic lack of trappings, is nevertheless a broad church. One leading deflationist, Hartry Field, has responded to the objection from non-conservativeness in a different way.[15] This response of my own might be disavowed by theorists with genuine as opposed to *ersatz* deflationary zeal. But I would hope that that would not stand in the way of a careful assessment of the potential 'deflationary way out' of the Gödelian predicament that is to be offered below.

4. The Gödel phenomena

Our discussion will be greatly facilitated by setting out formal details in a way that will make them amenable to the philosophically motivated innovations that will be suggested below. In particular, we need to have a clear understanding of the exact structure of the metaproofs involved in the Gödel phenomena. The expository part of what follows, and the form in which the details are cast, are therefore essential for our subsequent discussion.

Notation:

$\underline{n} =_{df}$ the numeral for natural number n.

An *S-refutation* of φ is understood as an *S*-proof of $\perp$ from φ as an assumption.

A system *S* of formal arithmetic is said to be *consistent* just in case for no sentence φ does *S* both prove and refute φ; or, equivalently, just in case there is no *S*-proof of $\perp$.

A system *S* of formal arithmetic is said to be *ω-consistent* just in case for no predicate $\psi(x)$ do there exist *S*-proofs of $\psi(\underline{0})$, $\psi(\underline{1})$, $\psi(\underline{2})$, … ; $\exists x \neg \psi(x)$.

This necessary and sufficient condition for ω-consistency of *S* can also be expressed as follows:

for every predicate $\psi(x)$, if there exist proofs in *S* of $\psi(\underline{0})$, $\psi(\underline{1})$, $\psi(\underline{2})$, … , then it is consistent with *S* to assume $\forall x \psi(x)$.

The 1-*consistency* of *S* is a special case of ω-consistency:

[15] Field 1999.

for every *primitive recursive* $\psi(x)$, if there exist proofs in S of $\psi(\underline{0})$, $\psi(\underline{1})$, $\psi(\underline{2})$, … , then it is consistent with S to assume $\forall x \psi(x)$.

We speak of S as a 'formal system of arithmetic', rather than a theory, to remind ourselves that S is (recursively) axiomatizable. All theorems of S are conclusions of S-proofs that use only S-axioms and the rules of first-order logic with identity.

Gödel's first incompleteness theorem tells us that for any sufficiently strong system S of formal arithmetic, an S-independent sentence G can be constructed. G is S-independent in the following sense:

if S is consistent, then G is not a theorem of S; and if S is 1-consistent, then $\neg G$ is not a theorem of S.[16]

Put another way, with emphasis on the proofs afforded by the formal system S, one can say, equivalently,

if S is consistent, then there is no S-proof of G; and if S is 1-consistent, then there is no S-refutation of G.[17]

The sentence G is usually called 'the Gödel sentence for S'. It is also often described as a 'self-referential' sentence.

The uniqueness of G, though, is only a relative matter. Certain choices concerning Gödel-numbering of expressions (the coding of syntactic items), and concerning other related matters, have to be settled one way or another in the course of G's construction. Different conventional decisions in these regards would lead to syntactically distinct Gödel sentences (in the language of formal arithmetic). We shall go along with established convention, however, in speaking of 'the' Gödel sentence for the system S.

$\#E =_{df}$ the code (Gödel-number) of syntactic item E.

Note that the effective coding operation # will embody the conventions referred to above.

$\overline{E} =_{df} \underline{\#E}$.

The *Representability Theorem*, provable for systems S of formal arithmetic that extend Robinson's finitely axiomatized theory Q (or even

---

[16] To be historically accurate, Gödel's original proof employed the assumption of ω-consistency rather than 1-consistency; but in that application it is really only 1-consistency that is needed.

[17] That it is necessary to assume the 1-consistency of S is evident from the fact that the (1-inconsistent) theory $PA + \neg Con_{PA}$ can refute its own consistency statement, hence refute its own Gödel sentence. I owe this observation to Steve Simpson.

just that infinitely axiomatizable subtheory of Q known as R),[18] states that

for all primitive recursive relations $\rho(x_1,\ldots,x_n)$, there exists a primitive recursive formula $R(x_1,\ldots,x_n)$, such that for all natural numbers $k_1,\ldots,k_n$, if $\rho(k_1,\ldots,k_n)$, then $R(\underline{k_1},\ldots,\underline{k_n})$ is provable-in-S; and if not-$\rho(k_1,\ldots,k_n)$, then $\neg R(\underline{k_1},\ldots,\underline{k_n})$ is provable-in-S.

A special instance of the representability theorem is:

there exists a primitive recursive formula $Proof_S(x,y)$, such that for all natural numbers n, m, if m codes an S-proof of the sentence coded by n, then $Proof_S(\underline{n},\underline{m})$ is provable-in-S; and if not-(m codes an S-proof of the sentence coded by n), then $\neg Proof_S(\underline{n},\underline{m})$ is provable-in-S.

We say that this primitive recursive formula $Proof_S(x,y)$ in the language of arithmetic 'numeralwise represents S-proof'.

For any sentence φ in the language of S, we have, as special cases of the foregoing instance of the Representability Theorem, the inferences

(α) m codes an S-proof of φ; therefore, there is an S-proof of $Proof_S(\overline{\varphi},\underline{m})$;

(β) not-(m codes an S-proof of φ); therefore, there is an S-proof of $\neg Proof_S(\overline{\varphi},\underline{m})$.

From the two-place representing formula $Proof_S(x,y)$ one can define a one-place provability predicate for S:

$Prov_S(x) =_{df} \exists y Proof_S(x,y)$.

The Gödel sentence G for a system S is so defined as to be interdeducible, within S, with the (distinct) sentence $\forall y \neg Proof_S(\overline{G},y)$ — the sentence that asserts, via the chosen coding, that G has no proof in S. The 'self-referential' character of G should therefore not be taken too literally.

The Gödel sentence is not syntactically identical with the statement of its own unprovability-in-S, even though such a conflation is often encouraged by informal talk of the Gödel sentence being 'self-referential'. Informally, G is thought to be 'saying of itself' that it is not provable-in-S.[19] But such 'saying of itself that it is not provable in S' as G achieves is achieved only by dint of the aforementioned *interdeducibil-*

---

[18] For details, see, e.g., Tarski 1953.

[19] We saw this in the quote from Lucas above; and both Hofstadter and Penrose follow suit.

ity of G with $\forall y \neg Proof_S(\overline{G},y)$.[20] In technical terms, G is a fixed point for the negation of the S-provability predicate. The upshot, however, is that we can rest assured that there are, in the system S,[21] two proofs that we shall abbreviate as

$$
\begin{array}{cc}
G & \forall y \neg Proof_S(\overline{G},y) \\
\Gamma & \Xi \\
\forall y \neg Proof_S(\overline{G},y) & G
\end{array}
$$

We shall use these proofs Γ and Ξ as constants in the metaproofs to be given below. The proof Γ establishes $\forall y \neg Proof_S(\overline{G},y)$ from G; while the proof Ξ establishes the converse inference. Both Γ and Ξ can be constructed in accordance with the rules of intuitionistic relevant logic.

An important sentence related to G is $\neg \exists y Proof_S(\overline{0=1},y)$. The latter asserts that there is no S-proof of the sentence $0=1$. This is conventionally taken as the consistency statement for S, abbreviated as Con_S. Another important Gödelian fact is that there is, in a rather weak subsystem S_w of S, an S_w-proof (Φ_1, say) of G from $\neg \exists y Proof_S(\overline{0=1},y)$ and an S_w-proof (Φ_2, say) of $\neg \exists y Proof_S(\overline{0=1},y)$ from G:[22]

$$
\begin{array}{cc}
\neg \exists y Proof_S(\overline{0=1},y) & G \\
\Phi_1 & \Phi_2 \\
G & \neg \exists y Proof_S(\overline{0=1},y)
\end{array}
$$

Now of course any reasonable formal system of arithmetic contains the axiom that 0 is initial, that is, that 0 is not a successor. Expressed inferentially, this amounts to the rule

$$
\frac{0=s(t)}{\bot}
$$

Since $1 =_{df} s(0)$, a special instance of this rule would be

$$
\frac{0=1}{\bot}
$$

---

[20] Robert Jeroslow has given a new self-referential formula denoted by the same term as denotes the arithmetized claim that its negation is provable. This, however, is not the standard choice of independent Gödel sentence when philosophers discuss these matters. See Jeroslow 1973.

[21] And, of course, in all extensions of S.

[22] This weak subsystem can be taken (and is usually taken) to be primitive recursive arithmetic (*PRA*). Another system that can serve in this regard is exponential function arithmetic (*EFA*), which is properly contained in *PRA*. It is an interesting question just how much weaker a 'natural' theory S_w of arithmetic will still suffice for proof of the interdeducibility of G and Con_S for any axiomatizable extension S of S_w. See Buss 1986, Ch. 7 and Cook 1975.

So, any system like *S* can show that 0=1 is absurd. The problem of consistency, however, is that 0=1 might itself be derivable in *S*, thereby revealing the whole system *S* to be absurd. We must bear in mind that the unprovability-in-*S* of the consistency of *S* (Gödel's second incompleteness theorem) is of course no obstacle to our being able to infer absurdity, within *S*, from 0=1.

We proceed below to examine consistent extensions *S** of any given consistent and sufficiently strong system *S* of formal arithmetic. Our eventual purpose will be to find what is minimally required in *S** in order to be able to prove the *S*-independent Gödel sentence *G* in a non-trivial fashion that is faithful to the essential deductive structure of the semantical argument.

5. What we should not assume for the proof of *G*

The extra methods in *S** that suffice to prove *G* will of course, by Gödel's first incompleteness theorem, not be available in *S* itself. The extra methods can be axiomatic or inferential. That is, we could, for *S**, either specify new axioms not available in *S*, or provide new rules of inference not available in *S* (or do both).

The semantical argument, despite its apparent brevity, has a non-trivial structure. We cannot, therefore, simply adopt *G* itself as a new axiom in order to obtain the system *S**, and then provide a one-line proof of *G* in *S**. This would violate the requirement that our regimentation (within *S**) be faithful to the structure of the informal reasoning. The semantical argument is designed to help one *understand why asserting* G *would be the right thing to do*, rather than denying *G* (or: refusing to assert *G* and refusing to deny *G*).

Moreover, in arriving at this understanding, we seek, on behalf of the deflationist, to avoid any use of a truth-predicate. If we succeed in doing so, then, *a fortiori*, we shall have avoided any use of a *substantial* notion of truth. That is, we shall have shown that such understanding can be attained by the deflationist. The reason why we set out to eschew the truth-predicate is to make it manifest that deflationist justice has been done to the semantical argument. Even if one could use just a modicum of non-creatively extending truth-talk, and, thereby, have our deflationist cake and eat it, we shall eschew it. We seek a deflationist high road, lest anyone think we have smuggled something essentially involving a substantial notion of truth through the back door.

Given this self-imposed constraint, we shall avoid any use of the principle of Σ_1^0-reflection for the system *S*. This is the principle that

says—using semantical vocabulary explicitly—that any S-provable Σ_1^0-sentence is true (in the standard model). In other words, any existential quantification of a bounded formula is provable in S only if it has a 'witness' among the standard natural numbers.

Another sufficient reason for avoiding appeal to Σ_1^0-reflection for S is that it is much stronger than what is barely needed. To see this, consider the following. It is well-known that the 1-consistency of any system S is equivalent to the principle of Σ_1^0-reflection for S. Moreover, the 1-consistency of S implies, but is not implied by, the consistency of S.[23] Hence the 1-consistency of S implies, but is not implied by, G.

Quite apart from its unnecessary strength and its explicit use of semantical vocabulary, the principle of Σ_1^0-reflection does not lend itself readily to faithful regimentation of the structure of reasoning in the semantical argument, which passes from (truth of) all numerical instances of the universally quantified sentence G to (the truth of) G itself.

We are requiring that our regimentation (within the extension S^* of S) of the semantical argument be faithful to its informal structure. This requirement also rules out making any use of Con_S—the arithmetical claim to be interpreted as the claim that S is consistent—as a new first principle (within S^*, but not in S). Con_S might at first glance seem attractive, both because it is of exactly the right logical strength (implying, and being implied by, G modulo S) and because it is free of any semantical vocabulary. The problem, however, with adopting Con_S as the new principle that will afford a proof of G is that the proof in question will be intractably long. This is in fact the proof called Φ_1 above. The argument for (the first half of) Gödel's first incompleteness theorem proceeds informally, at the metalevel, from the assumption of S's consistency to the conclusion that G is not provable in S. By arithmetizing the consistency assumption as the formal sentence Con_S and making the object-language sentence G itself the sought conclusion of the semantical argument (rather than the metalinguistic claim to the effect that G is not provable in S), the deductive route Φ_1 to be traversed becomes much more arduous. This, indeed, is what (in principle) needs to be established for the proof of Gödel's *second* incompleteness theorem. Gödel himself never even spelled out the proof Φ_1, but gave only persuasive indications as to how one might obtain it, if one were a sufficiently competent metamathematician. It was only later that Bernays provided the missing details.[24] Clearly, such an *arithmetized* argu-

---

[23] See C. Smorynski 1977 at p. 852.

[24] Hilbert and Bernays 1939.

ment Φ_1 from Con_S to G fails to meet the requirement that the regimentation of the semantical argument be faithful to its essential deductive structure.

6. Flirting with the truth-predicate

One way to specify new axioms is to extend the language L of S with new vocabulary; for then formulae not previously available (in the language L of S) will become available (in the language $L^\star$ of $S^\star$) as substituends of the axiom schema of induction. Thus new axioms will be generated for $S^\star$ that were not available in S.

An example of such linguistic extension would be the explicit adoption, in the language of $S^\star$, of a truth-predicate T for the system S. Now of course we have just nailed our flag to the deflationist's mast, and promised to have no truck with a truth-predicate. The reader should rest assured that we are not now relaxing that methodological resolve for a peek inside Pandora's box. We simply need to lift the lid in order to set out the problem that Shapiro poses for the deflationist.

A minimal requirement on a truth-predicate T affording an extension $S^\star$ of S would be that $S^\star$ license all inferences of the form

$$\frac{\varphi}{T(\overline{\varphi})} \qquad \frac{T(\overline{\varphi})}{\varphi}$$

where φ is a sentence of S

We shall call these the *T-inferences*. The first is the rule of *T*-introduction, or semantic ascent. The second is the rule of *T*-elimination, or disquotation. The *T*-inferences characterize truth only in a deflationary, non-substantial way. The *T*-inferences yield only a conservative extension of any theory to which they are added.[25] Indeed, Volker Halbach has shown that even Peano Arithmetic (*PA*) with full induction on formulas involving *T*, plus all 'uniform Tarskian biconditionals'—i.e. all instances of

$$\forall n_1 \ldots \forall n_k \left(T(\overline{\varphi(\underline{n_1}, \ldots, \underline{n_k})}) \leftrightarrow \varphi(n_1, \ldots, n_k) \right)$$

—is conservative over *PA*.[26] Stronger principles (than the mere instances of the *T*-scheme) governing the truth-predicate are needed in

---

[25] This observation is made also by Jeffrey Ketland, loc. cit., pp. 69–94. See his Theorem 1, p. 76. This is the same result as Lemma 2.4.2 of Feferman, (1991 at p. 14).

   The result is implicit in even earlier work of Harvey Friedman and Michael Sheard. See their paper (Friedman and Sheard 1987) at p. 17.

[26] For a model-theoretic argument, see Halbach 1996, Korollar 13.2. For a proof-theoretic argument, see Halbach 1999 Lemma 2.1.

S^* in order to ensure that S^* is stronger than S in the language of S. Something tantamount to axioms concerning *satisfaction* would appear to be needed.[27]

The conservativeness result tells us that in order to make the notion of truth substantial, further inference rules or axioms involving the truth-predicate, over and above the T-rules, would have to be adopted. Alternatively, one could adopt suitable Tarskian axioms or rules governing the *satisfaction* of open formulae, rather than the mere truth of closed formulae (sentences). This effects a *non-conservative* extension over Peano arithmetic. Ketland offers a result (his Theorem 2, loc. cit, p. 81) showing that the 'Tarski extension' (via satisfaction axioms) of any consistent extension of *PA* allows one to prove the claim that every theorem of *PA* is true. This is essentially Theorem 2.5.3 of Feferman's paper 'Reflecting on incompleteness' (loc. cit., p. 16). Given Gödel's second incompleteness theorem, it follows that the Tarski extension of such a theory is a *proper* extension.

7. Gödelian objections to deflationism

In this section we set out as clearly as possible the objections that Shapiro and Ketland have raised against deflationism, based on the Gödel phenomena.

According to the deflationist, the T-rules above are *all* that anyone could reasonably require of the notion of truth (which is to be expressed by T). Now as Shapiro observes,

> If truth … is not substantial—as the deflationist contends—then we should not need to invoke truth in order to establish any results not involving truth explicitly. (p. 497)

Thus, as Shapiro sees it, the deflationist will want to regard the T-inferences above as effecting only a conservative extension of the system S:

> … adding a truth predicate to the original theory [should] not allow us to prove anything in the original language that we could not prove before we added the truth predicate. (p. 497)

Conservativeness, says Shapiro, is 'essential to deflationism'. (p. 497) The requirement that S^* be a conservative extension of S can be stated formally as follows:

---

[27] We say 'something tantamount to', since in the arithmetical context, and in the notation of this paper, one can exploit the device of numerically quantifying into positions that are underlined within overlined expressions (as with the uniform Tarskian biconditionals just mentioned). I owe this observation to Volker Halbach.

Conservativeness: If $S^\star$ proves φ and φ is in L (the language of S), then S proves φ

But one claim that any theorist would want to make about truth, Shapiro contends, is that all S-theorems are true:

Soundness: If $\exists y Proof_S(\overline{\varphi}, y)$, then $T(\overline{\varphi})$

As Shapiro then shows, the deflationist who wanted to maintain the soundness of S in this fashion would face an uncomfortable aporia. The following three 'philosophical' claims (or rules) are jointly inconsistent with Gödel's first incompleteness theorem:

· All talk, in $S^\star$, of truth (of sentences of S) *conservatively* extends S

· $S^\star$ enables one to assert that all theorems of S are true

· $S^\star$ contains the T-rules, which are licit

Proof of the inconsistency: Consider the following proof in $S^\star$:

Suppose for *reductio* that $\exists y Proof_S(\overline{0=1}, y)$. Then by soundness of S, it follows that $T(\overline{0=1})$. By T-elimination, infer $0=1$. But 0 is initial. Contradiction. Hence by $\neg I$, $\neg \exists y Proof_S(\overline{0=1}, y)$.

So $S^\star$ proves $\neg \exists y Proof_S(\overline{0=1}, y)$. But $\neg \exists y Proof_S(\overline{0=1}, y)$ is in L. Thus by conservativeness, S proves $\neg \exists y Proof_S(\overline{0=1}, y)$. Hence, given Φ_1 above, S proves G. But this contradicts Gödel's first incompleteness theorem. *QED.*

Ketland shares Shapiro's view concerning the shortcomings of deflationism. He writes (loc. cit., p. 88)

… our ability to recognize the truth of Gödel sentences involves a theory of truth (Tarski's) which *significantly transcends the deflationary theories*. (Emphasis in original.)

Shapiro, as we have just seen, wanted to venture further. His aim was to have the deflationist committed to a contradiction (via the aporia above). That involved seeing the deflationist as saddled with a commitment to the soundness claim for S, in the form given above. Ketland merely sees the deflationist as unable to do what the substantialist can do — namely, vindicate our ability to see the truth of the Gödel sentence.

It is Shapiro, then, who mounts the potentially more dire objection — that the deflationist's central tenet will involve him in contradiction when he attempts to vindicate that same ability.

8. Meeting the Gödelian objections

8.1 Contra Ketland

Ketland, as we shall see below, errs by assuming without argument that Tarski's theory of truth is the *only* way that we can come to recognize the truth of Gödel sentences. (Here we mean, by this mention of truth: recognize that we *ought to assert* these sentences, rather than deny them.) The quote given from Ketland at the end of the previous section is acceptable when the word 'involves' is replaced by 'could be displayed by invoking':

> our ability to recognize the truth of Gödel sentences could be displayed by invoking a theory of truth (Tarski's) which significantly transcends the deflationary theories;

but it is unacceptable when 'involves' is replaced by 'can be displayed *only* by invoking':

> our ability to recognize the truth of Gödel sentences can be displayed *only* by invoking a theory of truth (Tarski's) which significantly transcends the deflationary theories.

As explained above, our main aim is—*pace* Ketland—to show precisely that there is *another* way to recognize the truth of Gödel sentences, which does not invoke anything like a Tarskian theory of truth significantly transcending the deflationary theories. If I succeed in this constructive aim, then that will be dispositive; for a deflationist rendering of the semantical argument for the truth of the Gödel sentence will refute Ketland's objection. That deflationist rendering will be given in section 9 below. We now turn to a more detailed examination of Shapiro's objection.

8.2 Contra Shapiro

Consider, then, the predicament that Shapiro thinks he has created for the deflationist by means of the aporia outlined above. The question that would obviously arise for the deflationist would be: What feature of the extended system $S^\star$ should we give up? The foregoing proof shows (*modulo* facts even more basic) that there cannot be a system $S^\star$ with the three major properties just listed. The basic facts, it is worth reminding ourselves, dictate the following:

> We cannot give up Gödel's theorem, since that is an established piece of mathematics. So too is the existence of the proof Φ_1. We cannot

deny that $\neg\exists yProof_S(\overline{0=1}, y)$ is in L, for that is obvious. We cannot give up the rule $\neg I$ of negation introduction, since that is essential to logic. We cannot give up the claim that 0 is initial, since that is a central and indispensable axiom of arithmetic.

So what about the three epistemologically less basic—more controversial, or philosophical—claims under adjudication? Of these, we cannot give up the disquotational T-elimination rule, since the admissibility of that rule is (half of) all that even the most austere deflationist wishes to claim about truth. Every party to the debate agrees on this; disagreements would arise only with respect to what *extra* principles governing truth might be deflationarily licit and/or needed in order to do justice to pre-truth-theoretic intuitions.

What, then, about giving up the requirement that our truth-theorizing should conservatively extend our theory of elementary arithmetic? Some, such as Shapiro, contend that is of the essence of the deflationist philosophical position that we are seeking to protect from outright refutation. Others, such as Field, contend that the deflationist can embrace non-conservativeness of truth-theorizing in this regard in good conscience.[28] Field's response to Shapiro's aporia is to give up the conservativeness assumption, and to disavow it as a commitment on the part of the deflationist.

For one who does regard the conservativeness assumption as essential to deflationism, there remains only one way out: give up the claim, in $S^\star$, that all S-theorems are true. Shapiro seems to think that this would be a terrible consequence for deflationism, revealing its inadequacy as a theory of truth. And Field, speaking for the deflationist, goes so far as to say (loc. cit., p. 538)

> … we certainly want to be able to prove inductively that all theorems of [S] are true on the basis of the truth of the axioms [of S] and the truth-preservtion of the rules of inference, and the [new induction] axioms … [involving 'true'] are obviously needed for such a derivation.

But would giving up the claim, in $S^\star$, that all S-theorems are true be such a terrible consequence for deflationism? And do we really need to agree with Field that we want to be able to *say* in $S^\star$ that all S-theorems

---

[28] The real source of that non-conservativeness, according to Field, lies with the notion of natural number, not with the notion of truth, when one accepts the indefinite extensibility of the composite concept 'true statement about the natural numbers' containing both those two notions as constituents. As he emphatically puts it (loc. cit., p. 539, fn. 12):

> It is something about our idea of natural numbers that makes it absurd to suppose that induction on the natural numbers might fail in a language expanded to include new predicates (whether truth predicates or predicates of any kind): nothing about truth is involved.

are true, and, moreover, be able not only to prove this in S⋆, but prove it *inductively*?

In reply, we would point out that the card-carrying deflationist (such as Horwich's minimalist) is not in a position, anyway, to assert that all *S*-theorems are true. He cannot *say* this in so many words. (Such is the lesson of the conservativeness result for the addition of the *T*-rules or the *T*-sentences.) *But* the deflationist can subscribe to, and indeed express this conviction (as to the soundness of *S*), in other, arguably satisfactory, ways.

The question that arises is: should we be in the business of stating soundness in a form that explicitly adverts to the preservation of truth? Perhaps all that Shapiro's aporia achieves is the insight that the answer to this last question should be negative. We are being asked to believe that the would-be S⋆ claim

> All *S*-theorems are true

is the only — or, if not the only, then at least the most desirable, or an obligatory — way to express our reflective conviction as to 'the soundness of *S*'. But is that the only way to express this conviction? On behalf of the deflationist, we venture to suggest not.

8.3 Reflection principles

Grover, Camp and Belnap, in their classic paper on the prosentential theory of truth,[29] remind us that there are other ways to express such reflective intuitions or convictions without resorting to the use of an explicit truth-predicate. One can express (within $S^\star$) the reflective insight that *S* is 'sound for *X*', where *X* is a particular fragment of *L*. Thus the soundness of *S* for *primitive recursive* sentences of *L* could be expressed, for example, by the following principle.

> (*pa*) If $\overline{\varphi}$ is a primitive recursive sentence and $\overline{\varphi}$ is provable-in-*S*, then φ

Note that the second conjunct of the antecedent ensures that φ is a sentence of *L*. Both conjuncts are sentences of *L*, using coding to make their syntactic claims. But the principle itself does not belong to the system *S*. Rather, the principle serves to *extend S*. In general, the language $L^\star$ of S⋆ extends the language *L* of *S*. Therefore any sentence of *L* can be *used* by the S⋆-theorist. That is why in this principle (*pa*) we have the bare conclusion φ, unadorned by any truth-predicate not in *L*.

---

[29] Grover, Camp and Belnap 1975.

The principle (*pa*) is but one of many possible reflection principles, all of them eschewing any use of a truth-predicate. As Feferman explains,[30]

> a reflection principle provides that the axioms of [the extending system] shall express a certain trust in the system of axioms [being extended]. (p. 261)

and

> By a *reflection principle* we understand a description of a procedure for adding to any set of axioms [*S*] certain new axioms whose validity follows from the validity of the axioms [*S*] and which formally express, *within the language of* [S], evident consequences of the assumption that all the theorems of [*S*] are valid. (p. 274; my emphasis)

A more recent explanation that Feferman offers is[31]

> [Reflection principles] are axiom schemata … which express, insofar as is possible without use of the formal notion of truth, that whatever is provable in *S* is true. (pp. 12–13)

One well-known reflection principle for a system *S*, which is equivalent to the 1-consistency of *S*, is that of Σ_1-reflection for *S*:[32]

$$(\Sigma_1^S) \text{ If } \overline{\varphi} \text{ is } \Sigma_1 \text{ and } \overline{\varphi} \text{ is provable-in-}S, \text{ then } \varphi.$$

Note that reflection principles allow one to 'lift off' the provability predicate, as one passes from left to right. In the converse direction, we have the corresponding kinds of *completeness* principle. For example, one such principle that we shall be invoking later is that of 'completeness for primitive recursive truths':

$$(ap) \text{ If } \overline{\varphi} \text{ is primitive recursive and } \varphi, \text{ then } \overline{\varphi} \text{ is provable-in-}S.$$

Note, however, that in talking about completeness for any class of 'truths', we are adverting to a principle that allows one to 'graft on' the provability predicate as one passes from left to right, and that the antecedent φ does not itself involve the truth-predicate. Hence all such principles, framed after this fashion, are deflationarily licit.[33]

Now of course we know from Tarski[34] that

---

[30] Feferman, 'Transfinite recursive progressions of axiomatic theories', loc. cit.

[31] Feferman, 'Reflecting on incompleteness', loc. cit.

[32] Note that we have dropped the superscript 0, in an attempt to make a notational distinction between this principle and its explicitly truth-wielding analogue of Σ_1^0-reflection, which was discussed above.

[33] In order to remind ourselves of this, we have devised the label '(*ap*)' for the last principle so as to abbreviate 'if *a*ssertable then *p*rovable'. By thinking of assertability within a system we avoid thinking of predicating truth.

> [Gödel sentences] possess the property that it can be established whether
> they are true or false on the basis of the metatheory of higher order having a
> correct definition of truth. (p. 274)

Thus, if one is able to deploy a formal notion of truth within a *metatheory* for the theory S — that is, within a system 'of higher order' than S — one can establish that the independent Gödel sentence G for S is true. In the light of the reflection principles, however, particular interest now attaches to the question whether, in passing from S to whatever system $S^\star$ might be able to prove the Gödel sentence G, any mention or use of a truth-predicate will actually be required.

8.4 Doing without the truth-predicate by using reflective extensions

Relative to our explicitly stated goal, is it *necessary* to adopt a truth-predicate and to have it feature explicitly in any *other S*-principles* than the (deflationary) *T*-rules given above?

It turns out that the answer is negative. But this is not because we need rules governing T other than the straightforward *T*-rules. Rather, it is because one does not need T at all. This negative answer confutes Shapiro's attempt to raise an insuperable difficulty for deflationism. (Shapiro, recall, sought to show how by using a truth-predicate one can (a) convince oneself that S is sound, hence (b) convince oneself that the statement of S's consistency is true, hence (c) convince oneself that the formerly independent Gödel sentence is actually true.)

In the paper just cited, Feferman investigates the relative strengths of the extensions of a system S that can be obtained by means of different reflection principles. Suppose S contains Peano arithmetic, and is ω-consistent. First, one obtains only a conservative, not a proper, extension[35] of S by adding all φ such that

$$\vdash_S Prov_S(\overline{\varphi}).$$

(Call this the 'self-monitoring extension' of S.)

Next, by Gödel's incompleteness theorem(s), one obtains one and the same *proper* extension of S by adding either G_S or Con_S. (Call this the 'consistency extension' of S.)

A yet stronger extension is obtained[36] by adding as axioms all sentences of the form

---

[34] Tarski 1956.

[35] Loc. cit., p. 274, Theorem 2.17(i).

[36] Loc. cit., p. 274, Theorem 2.17(ii).

$Prov_S(\overline{\varphi}) \rightarrow \varphi.$

(Call this the 'soundness extension' of *S*.) Feferman elsewhere considers the soundness extension as 'a means of expressing faith in the correctness of *S* without any new predicates at all.'[37] He writes that one

> variant of Gödel's doctrine is that the 'true reason' for the incompleteness phenomena is that though a formal system *S* may be informally recognized to be correct, we must adjoin formal expression of that recognition by means of a reflection principle in order to decide Gödel's undecidable statements.

The soundness extension, however, is *stronger* than the consistency extension because $S + Con_S$ cannot prove $Prov_S(\overline{\neg Con_S}) \rightarrow \neg Con_S$.

An even stronger extension[38] would be obtained by adding as axioms all sentences of the form

$\forall n \, Prov_S(\overline{\psi(\underline{n})}) \rightarrow \forall m \psi(m).$

(Call this the 'uniform extension' of *S*, by means of the 'uniform reflection principle'.)[39]

Equivalent uniform extensions are obtained[40] by using one or other of the following variants of the uniform reflection principle:

Add all sentences of the form $\forall m \psi(m)$ such that $\forall n \, Prov_S(\overline{\psi(\underline{n})})$;

or

Add all sentences of the form $\forall n \, (Prov_S(\overline{\psi(\underline{n})}) \rightarrow \psi(n)).$

What we intend to offer here, taking our cue from these reflection principles, are thoroughly deflationist means for expressing the soundness of *S*, and for obtaining grounds justifying both the assertion of the Gödel sentence and the assertion of the consistency of *S*. These means will not involve any recourse to a truth-predicate. They will, however, involve inferential extension of the system *S* itself. But then this is only to be expected, in the light of the Gödel phenomena.

Feferman's reflection principles admit of a nicely prosentential interpretation. In order to be able to obtain an $S^\star$-proof of *G*, it ought to be enough to furnish $S^\star$ with certain reflection principles of the same kind

[37] Feferman 1998 at p. 233.

[38] 'Reflecting on incompleteness', p. 276, Theorem 2.19(ii).

[39] Feferman informs me that, though the principle was unnamed in his 1962 paper, it has since come to be known as the 'uniform reflection principle'. The reader should not be misled by possible connotations concerning the presence of existential quantifiers, which mention of uniformity usually carries.

[40] Loc. cit., p. 276, Theorem 2.19(i).

as the 'prosentential' principle (*pa*) above, but with enough logical strength to carry out the desired proof of *G*. The consistency extension, of course, would be an uninformative hammer with which to crack the independent walnut. What we want, rather, is a reflection principle that will be more informative than the bald assertion of *G* itself, and that will allow one faithfully to reproduce the reasoning in the so-called 'semantical argument' for 'the truth of' *G*. Now of course such a reflection principle will have to produce a proper extension of *S* that is at least as strong as the consistency extension of *S*. We shall see, however, that *it need not be any stronger than the consistency extension.*

Now this may be puzzling, given that Feferman's other reflection principles (in particular, his uniform reflection principle) produce yet stronger extensions. A closer look, however, alerts one to the fact that in his uniform reflection principle, the predicate ψ is allowed to have arbitrary logical complexity. Feferman's uniform principle is therefore one of uniform *arithmetic* reflection. I propose, by contrast, a weaker uniform reflection principle, namely the principle of uniform *primitive recursive* reflection:[41]

($UR_{p.r.}$) Add to *S* all sentences of the form $\forall n(Prov_S(\overline{\psi(\underline{n})}) \rightarrow \forall m \psi(m))$, where ψ is primitive recursive.

It turns out that uniform primitive recursive reflection is just what is needed in order to be able to formalize faithfully the reasoning in the 'semantical argument'. For, it produces exactly the consistency extension—which is, after all, the minimum that will have to be produced anyway, since the independent Gödel sentence for *S* is equivalent, modulo *S*, to *S*'s consistency-statement. So ($UR_{p.r.}$) is of *exactly* the right logical strength—it is both necessary and sufficient unto the conclusion in support of which it is being adopted. Moreover, ($UR_{p.r.}$) intrudes itself at the appropriate point in our reconstruction of the semantical argument is such a way that all the inferential moves surrounding it are reconstructed in a pattern perfectly homologous to the structure of the informal reasoning itself. ($UR_{p.r.}$) not only has exactly the right logical strength unto the sought conclusion; it has, in addition, exactly the right logical form unto the desired deductive route to that conclusion. This is why I say that ($UR_{p.r.}$) is 'just what is needed' for faithful formalization of the semantical argument. It is not only the lightest hammer to crack the walnut, but also the one that allows the user to swing his arm in the familiar way.

---

[41] See Smorynski, 'The incompleteness theorems', loc. cit.; section 4.

Once we prove *G* in such a 'truth-predicate-free' system $S^\star$, the consistency of *S* follows by the *S*-proof Φ_2 above. So our subject will have reached the sought reflective conclusions that Shapiro insists he should be able to reach, but will have done so *without exploiting a truth-predicate on the way*, and *without using any machinery stronger than what is needed for the job*.

If Shapiro demands further that the deflationist do justice to the reflective intuition that *all* *S*-theorems are sound (and not just the primitive recursive ones), then we see no reason why we should not simply add, in $S^\star$, the principle

$$Prov_S(\overline{\varphi}) \rightarrow \varphi,$$

which produces the soundness extension. Löb's Theorem[42] ensures that this soundness principle could not be derivable in *S* without making *S* inconsistent. But here we are contemplating adopting the soundness principle in the extension $S^\star$ of *S*; and this averts that danger of inconsistency.

One can agree with Shapiro (loc. cit., p. 499) that the 'deflationist cannot say that all of the theorems of [*S*] are true'. But the deflationist can instead express (in $S^\star$) his willingness, via the soundness principle, to assert any theorem of *S*. The anti-deflationist desires to go one step further and embroider upon this same willingness by explicitly using a truth-predicate.

One is left wondering, however, whether what the deflationist is willing to do in this regard really falls short, in any epistemically unsatisfactory way, of what may reasonably be demanded of him. Why does one need to *say* at the metalevel what can be *shown* instead by adopting the inferential norm[43] expressed by the soundness principle at the metalevel? Part of the problem appears to be the felt need to adopt an explicit truth-predicate whereby one can *state* (or, as the deflationist would say: *over*state), by means of a single sentence in $L^\star$, the claim of *S*-soundness.

---

[42] Löb's Theorem states that no sufficiently strong and consistent system *S* can contain a 'proof-predicate' *P* for which, for all sentences φ, φ is *S*-deducible from $P(\overline{\varphi})$. For *P* to be a proof-predicate, the following conditions must hold for all sentences φ and ψ:

   (i) if φ is *S*-provable, then $P(\overline{\varphi})$ is *S*-provable;

   (ii) $P(\overline{\psi})$ is *S*-deducible from $P(\overline{\varphi \rightarrow \psi})$ and $P(\overline{\varphi})$; and

   (iii) $P(\overline{P(\overline{\varphi})})$ is *S*-deducible from $P(\overline{\varphi})$.

See Löb 1955. Obviously, given the predicate $Proof_S(x, y)$ representing the proof-relation, one can define the provability predicate $P(x)$ as $\exists y Proof_S(x, y)$.

[43] We call it an inferential norm since it is schematic in φ. It might be more useful to think of the principle as a rule of inference (in $S^\star$) rather than as an axiom.

This constrasts with the method of *showing*, not saying, that is adopted by the deflationist. When the deflationist adopts the soundness principle above, he is allowing that it may have infinitely many instances. If, instead of such a schematic principle, he were obliged to state the principle involved, he would be able to avail himself, *prima facie*, of one of two options:

(i) use an explicit truth-predicate, and say 'All *S*-theorems are true';

(ii) use a prosentential device, saying, instead (in the manner of Grover, Camp and Belnap), something like 'For every sentence that *S* proves, *tthat*'.

This is their sentential version of the schematic soundness principle, in a prosentential extension of ordinary English.[44] To those who might regard such linguistic innovation in English as bizarre, the prosententialist points out that philosophers have in the past gone so far as to recommend, on occasion, the revision of principles as fundamental as those of classical logic. That is, they have advocated changes in our *high-level theoretical commitments*. So it is well within the bounds of philosophical plausibility that the *expressive resources* of natural language concerning such deep matters as truth and sound reasoning may also need to be reformed and/or extended in certain ways, in order to reach a state of more perfect reflective equilibrium.

An anti-deflationist might object here that the conclusion of *S*-soundness was, in Shapiro's version of the justificatory or reflective process, arrived at *only after an argument* whose detailed steps have not been preserved by our deflationist who simply adopts the surrogate soundness principle above. That argument adverted to the truth of *S*-axioms and to the truth-preserving character of *S*-rules. To this objection a deflationist reply is at hand. Let us simply take surrogate rules (leading from premises about provability-in-*S* to simple assertions, not truth-predications, as conclusions) corresponding to each of the axioms and rules of inference whereby *S*-proofs are constructed. Then the soundness principle should turn out to be derivable in system *S**, and the deflationist will have mimicked the structure of Shapiro's justificatory process.

---

[44] Note the very important occurrence of the comma, separating off the quantifier phrase from the prosentence '*tthat*'. It would be a mistake to parse this claim in such a way as to make 'proves *tthat*' appear to be a grammatical constituent.

9. 'Meta-proofs' without the truth-predicate

We turn now to the technical matters needed in order to make good the philosophical claims entered above. We have already had to make mention of $Proof_S(x,y)$, the primitive recursive formula in the language of arithmetic that 'numeralwise represents S-proof' in the sense of the Representability Theorem.
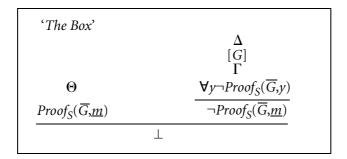
We also assume consistency of S, in the form of the following principle:

There is no S-proof of $\bot$.

Finally, we adopt the principle of *uniform primitive recursive reflection* ($UR_{p.r.}$) mentioned above.

Since the language of $S^\star$ at least contains that of S, and the stronger system $S^\star$ is taken to be a sound system of proof, any proof in $S^\star$ of a sentence φ in the language of S is a ground for asserting φ—even though φ might have no proof in the weaker system S. All that is needed, for the assertion of φ, is some proof of φ in what is recognized as some sound system of proof. In particular, if it turns out that there is a proof of φ in the sound system $S^\star$, then we are justified in asserting φ. We would thereby even have backing for the wordier claim that φ is true. For, from any philosophical perspective, sound proof suffices for truth. (Indeed, from an anti-realist perspective, the existence of some sound proof of φ is what the truth of φ consists in.) But the $S^\star$-proof itself that justifies the assertion of φ (or of φ's truth) need not make any use of an explicit truth-predicate—not even a truth-predicate for the language of the weaker system S.

The following S-proof-schema will recur in the $S^\star$-proofs to follow. It is at the heart of the phenomenon of the Gödel sentence for S 'saying of itself' that it is not provable-in-S. Once drawn into The Box, we will have violated the assumption that the system S is consistent.

We now provide a metaproof (that is, a proof in system S^*) of G. Note that we do not say 'showing that G is true'. The truth-predicate plays no role in the metaproof.

S^*-proof of G:

Suppose m codes an S-proof, say Δ, of G. By the representability inference (α) it follows that there is some S-proof, Θ say, of $Proof_S(\overline{G},\underline{m})$. Now Δ is an S-proof of G, from which in turn (via the S-proof Γ) one can deduce $\forall y \neg Proof_S(\overline{G},y)$. By $\forall$E we now have an S-proof of $\neg Proof_S(\overline{G},\underline{m})$; which, given Θ, puts us in The Box, violating S-consistency.

So m does not after all code any S-proof of G. By the representability inference (β) it follows that there is some S-proof of $\neg Proof_S(\overline{G},\underline{m})$. But m here is arbitrary. So for every n there is some S-proof of $\neg Proof_S(\overline{G},\underline{n})$. By $UR_{p.r.}$ it follows that $\forall y \neg Proof_S(\overline{G},y)$. Now, by the proof Ξ (which is in S, hence in S^*) G follows.

The foregoing S^*-proof justifies the assertion of G. The stronger system S^* contains methods for reflecting on the justificatory resources of the weaker system S. These methods can be seen at work in the application, in the proof just given, of various rules of inference that are available in S^* but not in S.

Note that we cut directly to the chase with our S^*-proof of G. We omitted the usual prolegomenon of showing that there is no S-proof or S-refutation of G. Some readers might harbour a lingering suspicion that perhaps there is need for use of an explicit truth-predicate in the proof of G's independence from S. To show that this is not so, we supply also the necessary metaproofs showing that there is no proof of G in S and that there is no refutation of G in S. In neither of these metaproofs is any use or mention made of a truth-predicate.

S^*-proof that G has no S-proof:

Suppose Δ is an S-proof of G. Then $\#(\Delta)$ ($=m$ say) codes an S-proof of G. Hence by (α) there is some proof, Θ say, of $Proof_S(\overline{G},\underline{m})$. Recall the S-proof Γ of $\forall y \neg Proof_S(\overline{G},y)$ from G. One step of $\forall$E yields an S-proof of $\neg Proof_S(\overline{G},\underline{m})$—which, given Θ, puts us in The Box, violating S-consistency.

So there cannot be any such S-proof Δ of G after all.

Note that we make no use of any assumptions or inferences involving an explicit truth-predicate. The S^*-proof just given, showing that G is

not provable in S, appeals to the first half (α) of representability of the proof relation, and to the consistency of the system S.

Another metaproof shows that there cannot be any refutation Ψ of G in the system S. This metaproof, too, makes no use of any assumptions or inferences involving an explicit truth-predicate. It appeals, however, to the Σ_1-reflection principle for S:

($\Sigma_1^{\,S}$) If $\overline{\varphi}$ is Σ_1 and $\overline{\varphi}$ is provable-in-S, then φ

and also to the completeness of S on primitive recursive statements (if *a*ssertable then *p*rovable):

(*ap*) If $\overline{\varphi}$ is primitive recursive and φ, then $\overline{\varphi}$ is provable-in-S.

Let Ω be the obvious intuitionistic S-proof of $\forall y \neg Proof_S(\overline{G},y)$ from $\neg \exists y Proof_S(\overline{G},y)$.

S^*-proof that G has no S-refutation:

Suppose Ψ is an S-refutation of G. Form the S-proof[45]

$$\frac{\overline{\neg \exists y Proof_S(\overline{G},y)}}{\begin{array}{c} \Omega \\ [\forall y \neg Proof_S(\overline{G},y)] \\ \Xi \\ [G] \\ \Psi \\ \bot \end{array}}$$
$$\overline{\exists y Proof_S(\overline{G},y)}$$

By Σ_1-reflection, $\exists y Proof_S(\overline{G},y)$. So suppose $Proof_S(\overline{G},\underline{m})$. (Note: this form of existential elimination, invoking a parametric *numeral* $\underline{m}$, presupposes Σ_1-reflection again.) Since $Proof_S(\overline{G},\underline{m})$ is primitive recursive, it follows by (*ap*) that there is an S-proof of $Proof_S(\overline{G},\underline{m})$. Call such a proof Θ. Suppose now for *reductio* that m codes an S-proof, say Δ, of G. Recall the S-proof Γ of $\forall y \neg Proof_S(\overline{G},y)$ from G. By one step of $\forall E$ we obtain an S-proof of $\neg Proof_S(\overline{G},\underline{m})$—which, given Θ, puts us in The Box, violating S-consistency.

So m does not code a proof of G after all. By (β) it follows that there is some S-proof, Θ say, of $\neg Proof_S(\overline{G},\underline{m})$. But once again, given Θ,

---

[45] Note that the last step of this proof is an application of classical reductio—or, more precisely, of Markov's Principle. But this step occurs within the object system, and is not taken at the meta-level.

this would violate *S*-consistency. So there cannot be any such *S*-refutation Ψ of *G* after all.

This metaproof explicitly appeals, in $S^{\star}$, to: Σ_1-reflection; consistency of *S*; the *S*-provability of assertable primitive recursive statements; and the second half, (β), of the representability of the relation of *S*-proof. Nowhere, however, is any appeal made to the behaviour of an explicit truth-predicate, not even one restricted to sentences of the weaker system *S*.

We have gone to great pains to prove our results directly, avoiding using any previously established result as a lemma for a further result. Now in the usual treatment of the Gödel phenomena, this is not the order of presentation. What is usually done is that one proves, assuming the consistency of *S*, that *G* is not provable in *S*. Then one proves, using Σ_1-reflection for *S* and the unprovability result, that *G* is not refutable in *S*. Then one uses the non-refutability result to show that *G* is true. This way of proceeding makes it look as though (the truth of) *G* can be established only by assuming Σ_1-reflection. This, however, is not the case. A more direct meta-proof of *G*, such as we have given above, can get by with only the principle of uniform *primitive recursive* reflection. This is of course implied by Σ_1-reflection.

The converse implication, however, appears to hold only *modulo* some *strengthening* of uniform primitive recursive reflection, and a further consistency assumption, which it is perhaps easy to grant. Let us strengthen our uniform reflection principle so that it applies not just to primitive recursive predicates, but also to Σ_1-predicates. (Thus its conclusion will be Π_2.) We may call the strengthened rule UR_{Σ_1}.

> *Theorem.* If the system ($S + UR_{\Sigma_1}$ for *S*) is consistent, then it implies Σ_1-reflection for *S*.

Naturally, any extended theory that we might countenance is intended to be consistent; so the use of UR_{Σ_1} would be tantamount to Σ_1-reflection. On its own, however, uniform *primitive recursive* reflection for *S* ($UR_{p.r.}$) does not imply Σ_1-reflection. Indeed, as remarked above, it can be shown that ($S + UR_{p.r.}$ for *S*) is equivalent to ($S + Con(S)$).

10. Conclusion

Deflationism—or at least its close cousin, prosententialism—does not come off at all badly in comparison with 'substantialism', the view that

the truth predicate expresses a substantial property. The objections to deflationism raised by Shapiro and by Ketland were designed to give substantialism an apparent edge over its rivals in the project of making sense of the Gödel phenomena. We have seen, however, upon a closer and more careful analysis, that the more modest view about truth can attain the various reflective insights—necessarily achievable only within an extended system—without recourse to any explicit use of a truth-predicate. We have argued that formulae involving the truth-predicate need not be regarded as obligatorily available substituends in the axiom schema of mathematical induction; and that the deflationarily licit *T*-inferences do not, anyway, vouchsafe a soundness result guaranteeing transmission of *substantial* truth. This deflationary stance enables one to avoid any damaging confrontation with the Gödel phenomena. We have also argued that deflationist or prosententialist strategies are available for 'reflecting' upon essentially incomplete formal systems in the process of continually extending them. This avoids any epistemic shortfall in comparison with substantialism.

Although we hold no brief for deflationism as an ultimately satisfactory theory of truth, our concern here has been to defend it against an interesting and ingenious attack on its philosophical credentials. For this defence of deflationism, the claim is not that a more substantial theory of truth is in any way incoherent. The claim is only that certain important accomplishments of such a theory are accessible from more modest beginnings. Whatever reasons one might have for eschewing a more substantial theory and holding to deflationism will have nothing to do with any alleged incoherence in the former. Rather, it will have to do with reasons of simplicity, economy of commitments, etc., that prompt one to reach one reflective equilibrium rather than another.[46]

Department of Philosophy   NEIL TENNANT
The Ohio State University
Columbus, Ohio 43210
USA

[46] Note added in proof: The only anticipation that we have found in the literature thus far of a prosententialist reading of the statement that there are true but unprovable sentences is in Grover 1990, p. 13, n.22. The present paper can be understood as a full culturing of that germinal insight, but at the level of argumentation as well as that of statement. (Unfortunately, Grover's paper came to our attention only after this paper was in proof.)

References

Anderson, Alan Ross (ed.) 1964: *Minds and Machines*. Englewood Cliffs: Prentice-Hall.

Barwise, J. (ed.) 1977: *Handbook of mathematical logic*. Amsterdam: North-Holland.

Buss, S. 1986: *Bounded Arithmetic*, Naples: Bibliopolis.

Cook, S. A. 1975: 'Feasibly Constructive Proofs and the Propositional Calculus'. *Proceedings of the Seventh ACM Symposium on Theory of Computing*, pp. 83–97.

Detlefsen, Michael 1986: *Hilbert's Program*. Dordrecht: Reidel.

Dummett, M.1963'The Philosophical Significance of Gödel's Theorem'. Reprinted in Dummett 1978 pp. 186–201.

——1978: *Truth and Other Enigmas*. London: Duckworth.

Dunn J. N. and A. Gupta (eds) 1990: *Truth or Consequences: Essays in Honor of Nuel Belnap*. Dordrecht: Kluwer.

Feferman, S. 1962: 'Transfinite recursive progressions of axiomatic theories'. *Journal of Symbolic Logic*, 27, pp. 259–316.

——1991: 'Reflecting on incompleteness'. *Journal of Symbolic Logic*, 56, pp. 1–49.

——1998: 'Infinity in Mathematics: Is Cantor Really Necessary? (Conclusion)', in *In the Light of Logic*. Oxford: Oxford University Press pp. 229-48.

Field, H. 1986: 'The Deflationary Conception of Truth', in MacDonald and Wright 1986, pp. 55–117.

——1999: 'Deflating the Conservativeness Argument'. *Journal of Philosophy,* 96, pp. 533–40

Friedman, Harvey and Michael Sheard 1987: 'An Axiomatic Approach to Self-Referential Truth'. *Annals of Pure and Applied Logic* vol. 33, pp. 1–21.

Grover, D., J. Camp and N. Belnap 1975: 'A Prosentential Theory of Truth'. *Philosophical Studies*, 27, pp. 73–125.

——1990: 'On Two Deflationary Truth Theories', in Dunn and Gupta 1990, pp. 1–17.

Halbach, V. 1996: *Axiomatische Wahrheitstheorien*. Berlin: Akademie Verlag.

——1999: 'Conservative Theories of Classical Truth', Studia Logica, 62, pp. 353–370.

Haldane, J. and C. Wright (eds) 1993: *Reality, Representation and Projection*. New York: Oxford University Press.

Hilbert, D. and P. Bernays 1939: *Grundlagen der Mathematik*, vol. I, Berlin: Springer, pp. 283–340.

Hofstadter, Douglas 1979: *Gödel, Escher, Bach: an Eternal Golden Braid*. Hassocks, Sussex: Harvester Press.

Horwich, P. 1990: *Truth*. Oxford: Blackwell.

Jeroslow, R. 1973: 'Redundancies in the Hilbert-Bernays Derivability Conditions for Gödel's Second Incompleteness Theorem'. *Journal of Symbolic Logic*, 38, pp. 359–67.

Ketland, J. 1999: 'Deflationism and Tarski's Paradise'. *Mind*, 108, 1999, pp. 69–94.

Löb, M 1955: 'Solution of a problem of Leon Henkin'. *The Journal of Symbolic Logic*, 20, pp. 115–118.

Lucas, J. R. 1961: 'Minds, Machines and Gödel', in Anderson 1964, pp. 43–59. Originally published in *Philosophy*, 36, 1961.

MacDonald, G. and C. Wright (eds) 1986: *Fact, Science and Morality: Essays on A. J. Ayer's Language, Truth and Logic*. Oxford: Blackwell.

Nagel, Ernest and James Newman 1958: *Gödel's Proof*. New York: New York University Press.

Peacocke, Christopher 1993: 'Proof and Truth' in Haldane and Wright, 1993, pp. 165–92.

Penrose, Roger 1989: *The Emperor's New Mind*. Oxford: Oxford University Press.

Shapiro, S. 1998: 'Proof and Truth: Through Thick and Thin'. *Journal of Philosophy*, 95, pp. 493–521.

Smorynski, C. 1977: 'The incompleteness theorems', in Barwise 1977. pp. 821–865.

Tarski, A. with A. Mostowski and R. M. Robinson 1953: *Undecidable Theories*. Amsterdam: North-Holland.

———1956: 'The concept of truth in formalized languages', in Woodger 1956, pp. 152–278.

Woodger, J. H. (trans. and ed.) 1956: *Logic, Semantics, Metamathematics: Papers by Alfred Tarski from 1922–1938*. Oxford: Clarendon Press.

Wright, Crispin 1993: 'About 'The Philosophical Significance of Gödel's Theorem: Some Issues', in *Realism, Meaning and Truth*, 2nd edn. Oxford: Blackwell, pp. 321–54.