Deflationism, Arithmetic, and the Argument from Conservativeness

Daniel Waxman
New York University
danielwaxman@gmail.com
Penultimate draft – please do not cite!

I Introduction

Deflationism about truth is one of the major contemporary answers to the question of the nature of truth. The best way of briefly introducing the view is to say that it stands in opposition to the well-known 'traditional' accounts of truth, which attempt to give a substantive analysis of what the truth of a sentence or proposition consists in.[1] According to the correspondence theory, for instance, it is whether a sentence corresponds to the world that determines its truth; and according to the coherence theory, it is its inclusion in some (perhaps privileged) coherent set of sentences. Unlike these theories and others like them, deflationism rejects the idea that truth plays any serious explanatory role in philosophy or elsewhere.

In this paper, I want to discuss and assess a particular line of attack that has been pressed against deflationists in recent years, originating in articles by Stewart Shapiro (1998) and Jeffrey Ketland (1999). These authors contend that the distinctive commitments of deflationism force a recognition that a formal theory of truth must be *conservative* over the truth-free theory to which they are added ("the base theory"). Loosely put, this requirement states that any sentence in the base language that does not already follow from the base theory also does not follow from the truth theory. (Note the slippery term "follows from", which is ambiguous between (at least) two distinct notions of logical consequence. This will be discussed very soon below.) But, they think, this gets the deflationist into trouble in the case of arithmetic:

[1] As far as possible I want to ignore the issue of what the bearers of truth are; in particular I will stay neutral between those who embrace propositions and those who prefer sentences (however these are typed). For purposes of brevity and readibility, I will use "sentences", with the proviso hereby made that I stand officially agnostic on this issue.

the Gödel sentence (G) for Peano Arithmetic (PA) is not a theorem of PA itself, but becomes a theorem once PA is extended by adding certain natural principles governing truth. The resulting theory of arithmetical truth is strong enough to prove the Gödel sentence and is therefore non-conservative over arithmetic itself. And this, it is argued, poses a serious problem for deflationists, because (according to its critics) deflationism is committed to truth theories being conservative over their bases.

In broad outline, the deflationist has three potential avenues of response. The most basic choice-point concerns the question: are deflationists committed to conservativeness in some sense or another? What I will call *rejectionist* responses simply answer no. *Compatibilist* responses are those that maintain that deflationism is both committed to and compatible with non-conservativeness. Such views have a further question to answer: in the sense that deflationists are committed to conservativeness, what is the relevant notion of logical consequence employed? *Proof-theoretic* compatibilists answer that it is proof-theoretic or deductive consequence. Responses of this sort contend that a deflationist can in good conscience accept a conservative theory of arithmetical truth, either by denying the need for G to be a theorem of the truth theory or by sketching a different – as it were, a truth-free – route to its derivation. *Semantic* compatibilists accept a conservativeness requirement, but insist that it is to be understood in terms of a semantic notion of logical consequence, according to which G *is* a genuine logical consequence of the axioms of arithmetic.[2] Responses of this sort necessarily involve going beyond the proof-theoretic resources of first-order predicate logic: as we know, by the completeness theorem for first-order arithmetic, semantic and proof-theoretic consequence coincide for such languages.

All of the present defenders of deflationism in the literature of which I am of aware are either rejectionists or proof-theoretic compatibilists. By contrast, semantic compatibilism has had no explicit defenders to date.[3] Although an attempt to force the deflationist into the acceptance of a rich notion of logical consequence was at the forefront of Shapiro's original paper, it has been for the most part ignored by those in the deflationist camp in favour of a more direct rebuttal. Volker Halbach goes so far as to call it 'more a trap than a way out' (Halbach 2001, p. 170). But a moment's consideration will show that this possibility is of more relevance than has been acknowledged. In a debate which hinges, among other things, on the nature of arithmetic, the question naturally arises: how is arithmetic to be understood? As I see the issue, the crux is whether we possess what is often called a categorical conception of the natural numbers – an understanding of arithmetic sufficiently rich to rule out the so-called non-standard models that arise for the first-order axiomatization – or whether arithmetic is understood in more axiomatic terms, as (merely) a formal theory in first-order logic. The issue of whether

[2]To forestall potential misunderstanding, 'semantic' here is used to mean, roughly, 'model-theoretic', as opposed to 'having to do with meaning'.

[3]Horwich (1990) is perhaps committed to such a view. See footnote 7 for further discussion.

we have a categorical conception of the natural numbers, and if so, how, is a huge one, and I will not be able to discuss it in depth here. Nevertheless, I shall argue that this distinction is crucial for finding our way sensibly through the conservativeness argument. Once we see things clearly, the deflationist can respond disjunctively as follows, without taking a stand on which of the disjuncts obtains. If we have a categorical conception of arithmetic, then this fact alone provides us with reason to countenance a stronger-than-proof-theoretic notion of logical consequence, and the deflationist in particular is put at no disadvantage by doing so. But if we do not, then the reasons for requiring the derivation of the Gödel sentence lapse. Either way, the deflationist is in the clear.

By the end, I hope to have accomplished three main goals. The first is to have clarified the debate, and in particular to have dispelled some of the confusion surrounding the way in which arithmetic and various notions of logical consequence figure in it. The second is to have explained why deflationists need not fear notions of semantic consequence that are stronger than proof-theoretic consequence – or, at least, no more than anyone else. The third is to have shown, through the viability of the disjunctive strategy mentioned above, that the form of conservativeness that deflationists should endorse is sensitive to the underlying theory of arithmetic. Far from being an unwelcome consequence, this is as it should be: for deflationists, truth is explanatorily inert, and so it should come as no surprise that one's conception of arithmetic will have a major impact on how one's resulting theory of arithmetical truth will turn out.

The plan of the rest of the paper is as follows. Section 2 briefly looks at the diversity of views that go under the heading of deflationism and draws out some preliminary constraints that such views impose on formal theories of truth. Section 3 introduces and summarizes some technical material concerning formal theories of arithmetical truth, considers some of the different ways one might expand PA via the addition of a truth predicate, and states results about the conservativeness of these theories over PA itself. After these preliminaries, Section 4 presents in detail the non-conservativeness objection to deflationism. Section 5 discusses several of the most influential deflationist responses, and (while finding much of worth), argues that none is fully satisfying. I then go on in Section 6 to argue for the disjunctive response I advocate on behalf of the deflationist. Finally, Section 7 contains some concluding remarks about the most plausible forms of deflationism going forward.

II Deflationism and Formal Theories of Truth

The immediate difficulty faced by any attempt to say something informative about deflationism, or assess any argument presented against it, is that it comes in many varieties. As I introduced it earlier, deflationism is the rejection of the thesis that there is any philosophi-

cally illuminating theory of truth. But this basic idea can and has been fleshed out in several different ways. Here I distinguish four strands of deflationism and attempt to draw out some (preliminary) implications that they have for formal theories of truth.

The first strand of deflationism is *metaphysical*: it claims that truth is a metaphysically thin, or deflated, or disunified property; or more radically still, that there is no single property that true sentences have in common. Although this kind of terminology is perhaps suggestive, it is very difficult to see how to develop the idea due to the murkiness of the notions involved. To the extent that the metaphysical formulation is appealed to by the proponents of the argument from conservativeness against deflationism, it is used in an unanalyzed and wholly intuitive way. As I will later argue, proponents of the argument are better served by bypassing the metaphysical formulation altogether.

Secondly, there is a *conceptual/semantic* strand of deflationism, concerning the *concept* of truth or the *meaning* of the predicate "is true."[4] The core claim of this variety is that the conceptual/semantic content of truth is *thin*: it is, in some sense to be elaborated, given by nothing more than the T-Schema:

(T-Schema) $T(\ulcorner \phi \urcorner) \leftrightarrow \phi$

Different varieties of conceptual/semantic deflationism carry out this elaboration in different ways. Here are two of the most influential examples. On Paul Horwich's *minimalist* view (Horwich 1990), it is constitutive of an English speaker's understanding of 'is true' that he is disposed to accept the (infinitely many) English instances of the T-Schema.[5] The concept of truth is then defined derivatively as the common constituent of belief states expressed in uses of the word by those who understand it. For Horwich, that is the sense in which all there is to truth is given by the T-Schema.[6]

As a result of his minimalist commitment that our understanding of truth is underwritten by the instances of the T-Schema, Horwich faces what has come to be known as the generality problem. The problem is this: the bare instances of the T-Schema seem to be too weak for deflationist purposes. In particular, they do not entail principles such as the law of excluded middle $\forall \phi T(\phi) \vee T(\neg \phi)$ and the compositional laws, for instance for conjunction:

$$\forall \phi \forall \psi (T(\phi \& \psi) \leftrightarrow (T(\phi) \& T(\psi)))$$

[4] Although one might have reason to distinguish these two claims – the semantic and the conceptual – for our purposes, it would introduce irrelevant complications. We can take the claim indifferently to concern the *content* of truth, without worrying here whether content is to be understood linguistically or mentally.

[5] Strictly speaking, Horwich's view is stated in terms of propositions rather than sentences, but for our purposes nothing turns on this subtlety.

[6] I use corner-brackets as a naming device. See footnote 13 for more on notation, and Section 3.2 for a brief discussion of the implications of various choices available here.

that, intuitively, a theory of truth should be able to prove. Consequently, there is a real question as to how minimalists are able to earn themselves the right to endorse such generalizations.[7]

Hartry Field's position is different. For Field, the centrality of the T-Schema is given by the fact that for any speaker who understands ϕ, there is a cognitive equivalence between ϕ and $T(\ulcorner\phi\urcorner)$, with the notion of cognitive equivalence cashed out in broadly inferentialist terms. However, to overcome the generality problem, Field wants to go further than merely accepting the instances of the T-Schema. His preferred way of doing this is by working directly with a *schematic* version of the T-Schema. The idea is to exploit our facility with schematic reasoning directly – as it were, with schematic variables entering into the object language rather than the meta language. In short, schematic variables are introduced into the language and rules of inference are introduced that allow one to substitute sentences for schematic variables and to infer $\forall x(Sent(x) \rightarrow P(x))$ from the schema $P(\ulcorner\Psi\urcorner)$.[8] This allows one to derive generalizations such as the compositional laws from the schematic version of the T-Schema. I will have a little more to say about this schematic approach in Section 6.

A third thesis of deflationism is that the positive role played by truth is solely an *expressive* or *generalizing* one. This claim is often combined with the other varieties of deflationism above, but is also occasionally taken to constitute the primary claim of deflationism itself. In particular, it is held that the primary purpose of the notion of truth is to allow us to indirectly endorse sentences to which we are able obliquely to refer but whose content we are unable or unwilling fully to specify (e.g. 'what the Pope said yesterday is true' or 'the Gödel sentence for PA is true') and to endorse simultaneously whole classes of sentences, including infinite sets (e.g. 'everything the Pope says is true' or 'all of the theorems of PA are true.') The idea that the notion of truth serves primarily an expressive/generalizing role is a natural fit with the conceptual/semantic varieties of deflationism. For it provides an elegant explanation of the – otherwise seemingly recalcitrant – fact that truth-talk is a well-entrenched and indeed communicatively essential part of natural language. If the predicate 'is true' is governed by principles as apparently trivial as the T-Schema, one might reasonably wonder why we see fit to employ it at all, let alone to the extent we do; and the expressive/generalizing use of

[7] See Gupta (1993) for an early critique of deflationism on these grounds (although, as Hartry Field pointed out to me, the relevant results were already known to Tarski), as well as Section 3 below. This problem becomes yet more pronounced when it is claimed by minimalists such as Horwich that the primary *purpose* of (the notion of) truth is to establish generalizations (see the third deflationist thesis below); for in that case, the minimalist's truth predicate is unable adequately to carry out the very task for which it is designed. Horwich is of course aware of the issue, and to settle it proposes a rule of inference analogous to the ω-rule (see Section 5) which will allow the troublesome generalizations to be obtained. However, this commits him to the acceptance of a notion of semantic consequence stronger than proof-theoretic. Although Horwich never to my knowledge discusses the argument from conservativeness, this suggests that he may be friendly to what I have called a semantic compatibilist view, for, once the ω-rule is added as a rule of inference, G becomes a consequence of the axioms of PA.

[8] Details of the schematic approach can be found in appendix to Field (2001, Ch. 4).

"true" provides an explanation of these facts that is at least plausible on its face.

Fourthly and finally, there is a thesis concerning the *explanatory* role of truth. On this understanding, deflationism is the claim that truth does not play a central explanatory role in our best theories of the world, including but not limited to the philosophies of mind and language, intentional psychology and semantics.[9] It is thus opposed to views that hold that truth plays a more robust explanatory role in our understanding of the world. Barry Loewer gives a helpful list of some reasonably prevalent philosophical doctrines that would, if correct, conflict with deflationism so construed: that

> truth is a substantive value and goal for belief and assertion, that understanding a language and grasping thoughts consists in knowledge of truth conditions (and other semantic properties and relations), that the difference between factual and non-factual discourse is explicated in terms of semantic notions, and that semantic properties (and relations) of thoughts and other intentional mental events and states enter into rationalizing and causal explanations of actions. (2005, p. 59)

Needless to say, this list is not exhaustive, and indeed the argument from conservativeness can be viewed as attemping to add an additional entry: that an inflationary notion of truth allows us, furthermore, to explain the fact stated by the Gödel sentence for PA.

In what follows, the form of deflationism I am interested in is the view constituted by the conjunction of a version of the semantic/conceptual claim, the expressive/generalizing claim, and the explanatory claim. In particular, I take Field's disquotationalism and Horwich's minimalism both to be paradigmatic deflationary views in the sense I will consider.

Let us now ask: what requirements does deflationism, so understood, place upon a formal theory of truth? I take the foregoing discussion to motivate at least the following three conditions, each corresponding to one of the strands of the view. It must (1) include or allow the derivation of every instance of the T-Schema[10]; (2) allow the expression and derivation of generalizations, including generalizations involving the notion of truth; all while (3) ensuring that truth plays no essential explanatory role that does not derive from its expressive/generalizing capacities. Conditions (2) and (3) are perhaps less than fully clear; but they will receive further discussion as we proceed.

[9] As Hartry Field has emphasized, there is a crucial caveat: if truth figures essentially in an explanation *due only to its expressive/generalizing role*, that is perfectly acceptable in the eyes of most deflationists.

[10] At least (to avoid confrontation with the Liar and other semantic paradoxes), for sentences stated in the truth-free fragment of the language.

III Some Technical Preliminaries

3.1 Conservativeness

For the sake of definiteness, I take a theory to be simply a set of sentences understood as its axioms – not, as the notion is sometimes used, a set of sentences closed under logical consequence. Let $\mathcal{T}$ be a theory stated in a language $\mathcal{L}$ (the 'base language'). Now consider an extended language $\mathcal{L}^+$, and let $\mathcal{T}^+$ be a theory extending $\mathcal{T}$ and stated in $\mathcal{L}^+$. The general notion of a conservative extension is defined as follows:

> $\mathcal{T}^+$ is a *conservative extension* of $\mathcal{T}$ if, for any sentence ϕ in the base language $\mathcal{L}$, if ϕ follows from $\mathcal{T}^+$ then ϕ follows from $\mathcal{T}$ alone.

The definition above contains the ambiguous term 'follows from', indicating that a species of logical consequence is in play. So we can disambiguate between (at least) the following two species of conservativeness:

> $\mathcal{T}^+$ is a *deductively (or proof-theoretic) conservative extension* of $\mathcal{T}$ if, for any sentence ϕ in the base language $\mathcal{L}$, if ϕ is deducible (in a given proof system) from $\mathcal{T}^+$ then ϕ is deducible from $\mathcal{T}$ alone.

> $\mathcal{T}^+$ is a *semantically conservative extension* of $\mathcal{T}$ if, for any sentence ϕ in the base language $\mathcal{L}$, if ϕ is true in all models of $\mathcal{T}^+$ then ϕ is true in all models of $\mathcal{T}$ alone.

Due to the soundness and completeness of (all of the natural) deductive systems for first-order logic, these two notions are coextensive in the case where such a deductive system is used and where $\mathcal{T}$ is a first-order theory. But as we will later be considering notions of semantic consequence in which they come apart, the distinction between the two species of conservativeness is well worth keeping in mind in what follows.

3.2 Some Ways of Adding a Truth Predicate to PA

Let $\mathcal{L}$ be the usual language of arithmetic, and let PA be the theory consisting of the usual first-order axioms of Peano arithmetic (stated, of course, in $\mathcal{L}$). PA is an infinitely axiomatized theory, in particular containing every instance of the induction schema for formulas ϕ containing vocabulary in $\mathcal{L}$:

(Induction Schema) $(\phi(0)\&\forall y(\phi(y) \rightarrow \phi(Sy))) \rightarrow \forall x\phi(x)$

The question we are interested in is: how might we extend PA to obtain a theory of arithmetical truth? As we shall see, there are several ways in which this can be done.

Syntactically, the natural approach is to move to a new language $\mathcal{L}^+ = \mathcal{L} \cup \{T\}$ whose new one place predicate T is to be understood as truth. To predicate truth of sentences we require a means of naming them, and we do this in the usual way by assuming a Gödel-numbering so that $\ulcorner \phi \urcorner$ denotes the Gödel-number of ϕ and E_n is the expression whose Gödel-number is n. Then $T(n)$ is to be understood as 'E_n is true'.

All of the theories we examine are extensions of PA, that is, every axiom of PA will be one of their axioms. Furthermore are all *materially adequate*, in Tarski's sense, that is, they entail each instance of the T-Schema:

(T-Schema) $T(\ulcorner \phi \urcorner) \leftrightarrow \phi$

where ϕ is a sentence stated in $\mathcal{L}$.[11]

Here then are four theories to consider:[12]

1. PA_D is the theory whose axioms are those of PA together with every instance of the T-Schema for sentences stated in $\mathcal{L}$. This is perhaps the most direct way of obtaining a materially adequate truth theory for PA – in essence, by adding the minimally required T-sentences by brute force.

2. PA_D^+ is the theory whose axioms are those of PA_D together with the instances of the induction schema required to extend it in the following way. Whereas PA includes an instance of the induction schema for every sentence stated in the truth-free language $\mathcal{L}$, PA_D^+ includes also an instance for every sentence *stated in the extended language* $\mathcal{L} \cup \{T\}$. In other words, PA_D^+ extends induction to deal with cases including the truth predicate.

   Philosophical remark: PA_D^+ and PA_D might be thought of as formal implementations of Horwichian minimalism about truth (see Section 2): someone who believes that "all there is to say" about truth is given by the instances of the T-Schema will plausibly adopt one of these two theories (the choice between them depending on whether they wish to extend arithmetical induction to sentences containing T).

3. PA_T is obtained by adding to PA Tarskian compositional clauses for the logical connectives. There are several equivalent ways in which this can be done (some of which

[11] Note the restriction to sentences stated in $\mathcal{L}$: we cannot, on pain of serious engagement with the paradoxes, require that our theories satisfy the T-Schema where ϕ is allowed to contain the truth predicate.

[12] The names of the different theories are adopted from Shapiro (2002).

first define satisfaction, and then define truth in terms of that). Here is a representative set of such axioms:[13]

- $\forall s \forall t(T(s = t) \leftrightarrow s^\circ = t^\circ)$
- $\forall \phi(T(\neg\phi) \leftrightarrow \neg T(\phi))$
- $\forall \phi \forall \psi(T(\phi \& \psi) \leftrightarrow T(\phi) \& T(\psi))$
- $\forall v \forall \phi(T(\forall v \phi) \leftrightarrow \forall t T(\phi[t/v]))$

4. PA^+_T is the same as PA_T except that it also extends induction to include sentences containing T.

Let us turn now to a examination of the strength of the four theories discussed above.

PA_D and PA^+_D are both conservative extensions of PA.[14]However, both have difficulty proving generalizations concerning truth. For instance, although each theory can prove, *for each sentence* ϕ, $T(\ulcorner\phi\urcorner) \vee T(\ulcorner\neg\phi\urcorner)$, neither can prove the universal generalization of this fact:

$$\forall \phi T(\phi) \vee T(\neg\phi)$$

This is a consequence of the compactness of first-order logic: for the generalization to follow, it would have to follow from finitely many of the axioms; but it clearly does not follow from any finite set of T-sentences.

PA_T is also a conservative extension of PA.[15] Unlike PA_D and PA^+_D, PA_T is very good at proving generalizations; for instance, it is able to prove bivalence, as well as analogous principles for other connectives.[16]

Not all expressible generalizations fare so well, however. Consider the following principle:

[13]An explanation of the notation (which allows for greater perspicuity at the cost of suppressing the details of the coding): Quantification into sentence position, for instance of the form $\forall\phi...$ is to be understood as shorthand for $\forall x(Sent(x) \rightarrow ...)$ where $Sent(x)$ is a complex arithmetical formula satisfied by a number n if and only if n is the Gödel number of a sentence of $\mathcal{L}$. (Which arithmetical formula it is will depend on fine-grained – and for our purposes, irrelevant – details of the Gödel numbering.) Similarly for quantification into term position. I use the symbol $^\circ$ as an abbreviation for the evaluation function for terms. Lastly, instances of the equality sign and logical connectives within the scope of T are shorthand for the application of functions representing the effect of applying those connectives. So, for instance, the axiom for negation above becomes when written out more fully: $\forall x(Sent(x) \rightarrow (T(\dot{\neg}x) \leftrightarrow \neg T(x))$ where $\dot{\neg}$ represents the function that takes a sentence to its negation. See Halbach (2011, p. 65) for more details.

[14]For PA_D and PA^+_D, here is a quick sketch of the proof. Suppose we can derive ϕ. The derivation of ϕ consists of a finite number of sentences; so there must be an upper bound on the complexity of these sentences – suppose it is Σ_n. Now, it is possible within PA to define a partial truth predicate T_n for Σ_n-sentences. We can thereby obtain a PA-proof of ϕ by replacing the instances of the T-Schema used in the original derivation with instances of the T_n-schema (together with explicit proofs of the T_n-sentences appealed to). For a detailed proof see Halbach (2011).

[15]See Halbach (2011, p. 100) for a proof theoretic argument to this effect.

[16]Thanks to Jeffrey Ketland for pointing out a mistaken claim about PA_T in an earlier draft.

(T-Bew) $\forall\phi(Bew_{PA}(\phi) \rightarrow T(\phi))$ – all the theorems of PA are true.

Although all the *instances* of T-Bew are provable in all four theories we are examining, the general principle itself is not provable in PA_D and PA_D^+ (by a simple compactness argument) and in PA_T. By contrast, PA_T^+ *does* prove T-Bew, and as a result is able to formalize the following argument, resulting in its proof-theoretic non-conservativeness over PA.[17] To see this, note that (since the RHS is an instance of T-Bew) that:

$$PA_T^+ \vdash Bew_{PA}(\ulcorner 0 = 1 \urcorner) \rightarrow T(\ulcorner 0 = 1 \urcorner).$$

From the T-Schema (whose instances are derivable in PA_T^+) we have also:

$$PA_T^+ \vdash T(\ulcorner 0 = 1 \urcorner) \leftrightarrow 0 = 1,$$

and combining the two, we get:

$$PA_T^+ \vdash Bew_{PA}(\ulcorner 0 = 1 \urcorner) \rightarrow 0 = 1.$$

But of course PA proves that $0 \neq 1$, as does its extension PA_T^+. Contraposing and applying modus ponens, we obtain

$$PA_T^+ \vdash \neg Bew_{PA}(\ulcorner 0 = 1 \urcorner).$$

But $\neg Bew_{PA}(\ulcorner 0 = 1 \urcorner)$ is just Con_{PA}, the statement that PA is consistent, and we know by Gödel's Second Incompleteness Theorem that (if PA is consistent, which I henceforth assume without comment) $PA \nvdash Con_{PA}$. Therefore PA_T^+ is a non-conservative extension of PA.

This completes the brief review of the technical situation. The following table succinctly summarizes the crucial results.

	Truth axioms	Extends induction?	Proves T-Bew?	Conservative over PA?
PA_D	T-instances	No	No	Yes
PA_D^+	T-instances	Yes	No	Yes
PA_T	Compositional	No	No	Yes
PA_T^+	Compositional	Yes	Yes	No

IV The Argument From Conservativeness

The argument from conservativeness can be viewed as operating in three steps. The first step is that deflationism imposes the constraint that truth-theories be conservative, in at least some sense of the notion. The second step contends that any adequate theory of arithmetical truth

[17] This argument is in all essentials the "semantic argument" discussed in more depth in Section 4.2.

will be *proof-theoretically* non-conservative. The third step argues that for deflationists, the relevant notion of consequence figuring in the conservativeness constraint is proof-theoretic, and that deflationists cannot avail themselves of a stronger, semantic notion. In the taxonomy introduced in Section 1, those deflationists who resist the first step are rejectionsts; those who accept the first step and resist the second are proof-theoretic compatibilists; and those who accept the first and second steps while resisting the third are semantic compatibilists. I will discuss each step in turn.

4.1 Deflationism Requires Conservativeness

Here is the core of Shapiro's argument for a conservativeness requirement. (Note that nothing in this passage suggests that conservativeness need be understood in proof-theoretic terms):

> I submit that in one form or another, conservativeness is essential to deflationism. Suppose, for example, that Karl correctly holds a theory B in a language that cannot express truth. He adds a truth predicate to the language and extends B to a theory B' using only axioms essential to truth. Assume that B' is not conservative over B. Then there is a sentence b in the original language (so that b does not contain the truth predicate) such that b is a consequence of B' but not a consequence of B. That is, it is logically possible for the axioms of B to be true and yet b false, but it is not logically possible for the axioms of B' to be true and b false. This undermines the central deflationist theme that truth is insubstantial. Before Karl moved to B', $\neg\phi$ was possible. The move from B to B' added semantic content sufficient to rule out the falsity of b. But by hypothesis, all that was added in B' were principles essential to truth. Thus, those principles have substantial semantic content. (1998, p. 497)

Ketland, along similar lines:

> One might suggest that these [conservativeness results] illustrate a kind of 'analyticity' or 'contentlessness' that deflationary theories of truth exhibit. Adding them 'adds nothing". Indeed, it is these metalogical properties that are closely connected to the idea that the deflationary truth theories illustrate the 'redundancy' or 'non-substantiality' of truth. Indeed, one might go further: if truth is non-substantial as deflationists claim then the theory of truth should be conservative. Roughly: non-substantiality $\equiv$ conservativeness. (1999, p. 79)

A preliminary point to make is that such arguments cannot be right as they stand. Halbach (2001) has pointed out that any adequate truth theory – including the pure minimalist theory

(consisting of the addition of the instances of the T-Schema) – will not be conservative over pure logic. For take some logical truth ϕ. Any adequate truth theory will prove the instances of the T-Schema for ϕ and $\neg\phi$, and therefore will prove both $T(\ulcorner\phi\urcorner)$ and $\neg T(\ulcorner\neg\phi\urcorner)$. It follows that $\exists x \exists y(x \neq y)$, a conclusion that does not follow from pure logic alone.

However, an appropriate and non-ad-hoc modification of the conservativeness claim is easy to come by. The essence of the problem raised by Halbach is that a theory of truth is committed to the existence of *bearers* of truth. As mentioned earlier, it is controversial what exactly truth bearers are. But that there must be truth bearers of one kind or another is not seriously up for debate; and whatever they are, a formal theory of truth must include names for them. This requirement is satisfied by the inclusion of a theory of syntax, containing a name for each potential truth bearer stateable in the base language. Typically, syntax is simply arithmetized: a scheme of Gödel-numbering is implemented, setting up a correspondence between natural numbers and potential truth bearers. Setting up a Gödel-numbering means that the theory of syntax is taken just to be PA; but this is not the only possible way to proceed. Equally well, one might choose to name potential truth bearers with strings of symbols, in which case the theory of syntax would be a theory of strings and concatenation. However this is done, the obvious amendment of the conservativeness requirement is that the truth theory be conservative over the base theory *plus the theory of syntax*. The conservativeness claim must, and can reasonably, be amended in this way.[18]

4.2 Adequacy Requires Proof-Theoretic Non-Conservativeness

Why do Ketland and Shapiro believe that any adequate truth theory will be (proof-theoretically) non-conservative? The answer concerns our ability to recognize the truth of the Gödel sentence for arithmetic. Here are the opening lines of an essay by Dummett:

> 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. (1963, p. 186)

Dummett here adds an epistemic claim that is not strictly contained in the formal mathematical result: that *G can be recognised by us to be true.* A natural question to ask is: how does this recognition proceed? As orthodoxy has it, it goes via an argument using the concept of arithmetical truth. Here again is Dummett on that argument:

> By hypothesis the axioms of the system are intuitively recognized as being true, and the rules of inference of the system as being correct... Hence we may establish

[18] Shapiro (2002) accepts this amendment to his original statement of the conservativeness requirement. See Heck (Forthcoming) for a helpful discussion of alternative theories of syntax.

> by an inductive argument on the lengths of formal proofs that each proof in the system has a true conclusion, and by another inductive argument on the number of logical constants in a statement that no statement is both true and false; concluding from this that the system is consistent [a statement which is provably equivalent in the system to *G*]. (1963, p. 195)

Let us call this the *semantic argument for G*.[19] Shapiro puts it as follows:

> once our subject has taken on the truth predicate, and he notices that all the axioms of [PA] are true and that the rules of inference preserve truth, he concludes that every theorem of [PA] is true. He also knows (from [the T-Schema]) that "$0 = 1$" is not true, and so "$0 = 1$" is not a theorem of [PA]. So our subject concludes that [PA] is consistent ([Con_{PA}]) and that [*G*] is true. *The defect of [any weaker theory] as a theory of arithmetic truth is that it cannot reproduce this simple, informal reasoning*. (1998, p. 499, emphasis added)

Ketland (1991, p. 91) similarly places heavy weight on the ability of PA^+_T to *prove* *G*, and thereby allow us to come to recognize its truth. For otherwise – without a proof from a theory of arithmetical truth – he thinks it is difficult, if not impossible, to explain why *G* is true.

It is worth distinguishing two desiderata for a deflationist theory of arithmetical truth that are not clearly separated by Shapiro and Ketland, one weak and one strong. The weak desideratum is that the deflationist must be able to derive the (truth of the) Gödel sentence by giving *some* proof or other from (deflationistically acceptable) principles concerning arithmetic and truth. The strong desideratum is that it must be possible to formalize the semantic argument itself in the deflationist's theory of arithmetical truth. As we will see, Tennant's response (discussed in Section 5) is an example of a deflationist response that satisfies the weak desideratum but not the strong one.

As was mentioned in Section 3.2, PA_D and PA^+_D fail to yield a great many generalizations which intuitively it seems an adequate theory of truth should yield. However, PA_T *does* yield many such generalizations, in particular the ones arising via the interaction of the connectives with the truth predicate. The compositional principles it includes are responsible for this additional power (although, as we have seen, they are not powerful enough to yield T-Bew). It might then be thought that PA_T is an attractive middle-ground for the deflationist to inhabit – it is strong enough to account for many, though not all generalizations involving truth, and yet weak enough to still be proof-theoretically conservative over PA. It is worth mentioning

[19] In effect, the proof in Section 3 of *G* is just the formalization of this argument in a sufficiently powerful theory of truth – it is precisely this result that leads to the non-conservativeness of PA^+_T over PA.

Shapiro's response to this move: he wants to deny such an option to the deflationist, for he thinks that the position is an unstable one:

> whether one is a deflationist or not, there is no good reason to demur from the extension of the induction scheme to the new language. There is no reason to demur from $[PA^+_T]$. Informally the induction principle is that for any well-defined property (or predicate), if it holds of 0 and is closed under the successor function, then it holds of all natural numbers. It does not matter if the property can be characterized in the original, first-order theory. (1998 p. 500)

In support of this claim, he cites Dummett: 'It is part of the concept of natural number, as we now understand it, that induction with respect to any well-defined property is a ground for asserting all natural numbers to have that property' (2007, p. 337). As we shall later see, the idea that induction must be extended to any property will prove to be a watershed between two different understandings of arithmetic.

4.3 The Move to Stronger Notions of Logical Consequence

Assuming the success of the argument so far, what room for manoeuver is left for the deflationist? The only escape route is the acceptance of a notion of logical consequence (or a notion of what is 'implicit in' a theory) that goes beyond its proof theoretic consequences – what I earlier called semantic compatibilism. For if the deflationist can appeal to such a notion, the possibility arises of arguing that the deductive consequences of PA^+_T in the language of arithmetic – including G – were 'all along', so to speak, consequences of our best truth-free theory of arithmetic. The sting would then be taken from the conservativeness argument: for although the theory of arithmetical truth would not be proof-theoretically conservative over arithmetic, it would nevertheless be semantically conservative for the relevant species of semantic consequence, thereby satisfying the intuition behind the conservativeness requirement.

Shapiro (1998), however, thinks that this option is a 'thin reed' for the deflationist to take. His discussion focuses primarily on two difficulties. Firstly, all of the available options for adding logical resources to obtain a strengthened notion of semantic consequence appear to have substantial logical and mathematical content. Take second-order semantic consequence under the standard semantics, for instance. Many mathematical structures – for example, the natural numbers and the real numbers – possess categorical second-order axiomatizations; even the universe of sets (about a rich a structure as is imaginable) is characterized 'quasi-categorically', in the sense that any two models of second-order Zermelo-Fraenkel set theory with the Axiom of Choice are either isomorphic or one is isomorphic to an initial segment of the other. As a result, truth in all of these structures can be defined in terms

of second-order consequence. The second-order approach in particular faces the additional worry that second-order logic involves an expansion in one's ontology, with some (most famously, Quine) going so far as to attribute to it all of the ontology of set theory. Secondly, any notion of consequence strong enough to deliver the set of arithmetical truths is bound, by Gödel's results, to be non-effective. This poses an epistemological problem: if there is no effective procedure by which we can recognize the validity of an inference, it is not clear how we can be said to truly grasp the notion of consequence at stake. At the very least, Shapiro thinks, these are obstacles which render this approach extremely unattractive.

V Deflationist Responses to the Argument

5.1 Can the conservativeness requirement be denied?

Is there any reasonable scope for denying that a deflationist theory of truth must be conservative? Shapiro and Ketland both emphasize that the conservativeness requirement draws upon what I have termed the metaphysical understanding of deflationism: that 'truth is insubstantial', or some similar claim. This is not indisputable: the formulation of the metaphysical construal is somewhat murky, and in particular, very little is said in support of the crucial transition from 'truth is insubstantial' to 'truth theories must be conservative'. Nevertheless, the transition has considerable intuitive force, for it seems extremely uncomfortable to maintain that truth is an insubstantial or non-robust property if the addition of truth principles leads one to rule out what were previously considered to be live possibilities concerning a (truth-free) subject-matter. Perhaps the best way of understanding the transition is as a proposed explication: the informal notion of metaphysical insubstantiality is to be (possibly partially) explicated in terms of the formal criterion of conservativeness. It is striking, and a mark in favour of the plausibility of this understanding, that the conservativeness requirement has attracted considerable support among deflationists themselves.

It might be thought that an analogous argument for the conservativeness requirement could be made, appealing not to the deflationist's committment to the metaphysical insubstantiality of truth, but rather to its explanatory inertness. On reflection I am doubtful that such an argument succeeds.[20]

Suppose B is our base theory and B^+ results from B by adding appropriate truth principles. Let *explanatory conservativeness* then be the claim that

> for any ϕ in the base language, if ϕ is explained by B^+ then ϕ is explained by B.

[20] Thanks to an anonymous reviewer for pressing me to clarify an earlier version of this argument.

An argument could be run against the deflationist if it could be shown that (i) PA^+_T explains G, (ii) PA fails to explain G, and (iii) the deflationist is committed to explanatory conservativeness.

I think that (i) is arguably true. Although not all proofs are explanatory, the semantic argument for G very plausibly is.[21] So, since the semantic argument can be run in PA^+_T, we have every reason to believe that PA^+_T thereby explains G.

I am much more doubtful of (ii). The best reason I can think of for endorsing (ii) appeals to the claim that proof-theoretic derivability is a necessary condition for mathematical/logical explanation. But – to anticipate issues that arise in Section 5 – if we countenance a notion of logical consequence richer than proof-theoretic derivability, this condition will appear unmotivated.

Nevertheless, even if (i) and (ii) are granted for the sake of argument, the Gödel sentence nevertheless fails to cause a problem for the deflationist because (iii) is false. Recall from §2 that the deflationist finds utility in the notion of truth precisely because it plays a generalizing and expressive role. At best, then, explanatory conservativeness is a constraint to which the deflationist is committed only when dealing with explanations *in which the role of truth does not derive solely from its expressive/generalizing capacities*. As Field puts it, 'any use of 'true' in explanations which derives solely from its role as a device of generalization should be perfectly acceptable' (1999, p. 537). But the role played by truth in the semantic argument is precisely a generalizing one: as Shapiro acknowledges, the 'centerpiece of the explanation' is the ability of truth to 'make the generalization' that 'every theorem of PA is true' (1998, p. 506). So I am sceptical that considerations of explanatory inertness can be marshalled in service of an argument against the deflationist here.

5.2 Interlude – Two Conceptions of Arithmetic

Before we continue discussing how the deflationist can respond to the argument from conservativeness, we need to get clear on a more basic issue first. What do we mean when we talk about 'arithmetic'? There is an ambiguity here, underscored by two different conceptions of the subject-matter. The first conception is broadly model-theoretic in nature: for lack of a better term, let us call it the *categorical* conception of arithmetic. It holds that arithmetic is a subject about a particular mathematical structure – the natural numbers, $\mathbb{N}$, the elements of which are obtained by beginning with 0 and interating the successor operation finitely many times. By contrast, the other conception is broadly proof-theoretic in nature: let us call it the *axiomatic* conception of arithmetic. It holds that our best understanding of arithmetic consists in (and is exhausted by) the proof-theoretic consequences of a particular set of ax-

[21] See Shapiro (1998, p. 505) for additional argument to this effect.

ioms, namely first-order PA. As is well known, these two conceptions of arithmetic come apart. First-order PA has so-called non-standard models, models which are not isomorphic to $\mathbb{N}$.[22] For the categorical theorist, such models are to be ruled out as *unintended interpretations* of the axioms, in contrast to $\mathbb{N}$ which is often called the intended interpretation. On the categorical conception, arithmetic is not fully adequately captured by its axiomatization in first-order PA, or indeed any first-order axiomatization, for any such first-order theory will admit non-standard models. If the axiomatic conception is viewed through a model-theoretic lens, arithmetic will appear to have a diversity of admissible models, none of which has any claim to being privileged as 'intended' over and above the others.

It is a fundamental question in the philosophy of mathematics as to which of the conceptions is most defensible.[23] Although there is a natural tendency amongst mathematicians to take something like the categorical conception for granted, for a variety of reasons some philosophers have been sceptical of our ability to 'intend' a unique interpretation. Those who have tried to sustain a categorical conception have often sought to do so by endowing their logic with additional resources of sufficient strength to identify a single model (or, at least, a class of isomorphic models). There are many ways in which this can be done, if one is willing only to countenance the required resources. Here are a few possibilities.

Firstly, one might appeal to second-order quantification with the "standard" semantics.[24]The second-order Peano axioms differ from the first-order axioms only in that they replace the (infinitely many instances of the) induction schema with a single induction *axiom* $\forall X(X(0)\&\forall y(X(y) \rightarrow X(Sy))) \rightarrow \forall x X(x)$. Non-standard models then do not arise, for second-order PA is a categorical theory – any two models are isomorphic. The existence of non-standard models can be explained by noting that the first-order induction schema allows induction only over predicates definable in the language of arithmetic; and since there are subsets of $\mathbb{N}$ that are not so definable, the induction schema is bound therefore to 'miss' genuine instances of induction, and thereby 'over-generate' models of arithmetic. Alternatively, if one has scruples about appealing to full second-order quantification – perhaps for reasons of ontological commitment, or epistemic worries about the intractibility of the second-order consequence relation – there are other options. One option is adopting the logic L_{Q_0}, involving the addition of an

[22]One way to see this uses the Lowenheim Skolem theorem: any first-order theory with infinite models has models of every infinite cardinality, so there are models of PA of cardinality $> \aleph_0$. Another way is to expand the language of arithmetic by adding a new constant symbol c and consider the theory obtained by adding to PA the set of sentences $\{s_n : n \in \mathbb{N}\}$ where s_n is $c > \bar{n}$. Every finite subset of this theory has a model – if k is the largest number such that s_k is in our subset, interpret c as $k + 1$. So, by the compactness theorem, that theory has a model. But although this model satisfies PA, it also contains a non-standard number which is so to speak "infinitely distant" from 0.

[23]Not only in the philosophy of arithmetic: similar questions arise in the philosophy of set theory also.

[24]The standard semantics take the second-order quantifiers to range over all sub-collections of the domain, as opposed to Henkin semantics, which takes them to range over a distinct (and possibly truncated) 'second-order' domain.

additional cardinality quantifier whose interpretation is stipulated to be 'there exist at most finitely many__'. The consequence relation for L_{Q_0} is non-effective, but not as rich as full second-order consequence (for instance, L_{Q_0} does not characterize the real numbers up to isomorphism).[25] In the same vicinity is ω-logic, which licenses the inference from (the infinitely many premisses) $\phi(0), \phi(1), \phi(2)...$ to $\forall n \phi(n)$. A further alternative – which will perhaps be attractive to deflationists who emphasize the use of the notion of truth in expressing infinite sets of sentences – is the adoption of an infinitary logic such as $L_{\omega_1,\omega}$, which allows the construction of infinite conjunctions and disjunctions. As with second-order logic, these alternatives characterize the natural numbers uniquely up to isomorphism, and so non-standard models do not arise.

Accepting any of these logical resources will commit one to accepting a semantic consequence relation according to which G – and indeed all other arithmetical truths – are genuine semantic consequences of arithmetic. This is one reason for some philosophers' unease with the categorical conception: on one plausible interpretation of Gödel's theorems, the set of arithmetical truths is not recursively enumerable, and so it is questioned whether we can have any grasp of a non-effective consequence relation that generates all such truths.

5.3 Reflection Principles to the Rescue?

Tennant (2002) takes what I have called a proof-theoretic compatibilist view: he accepts that deflationism is committed to the proof-theoretic conservativeness of an arithmetical truth-theory over PA, but thinks that this will pose no threat to deflationism. In fact, he thinks that the deflationist position involves the claim that truth is not a genuine property, and that this motivates deflationists to refuse to extend induction to expressions containing the truth predicate, the intuitive idea being that we are required to extend induction only to genuine properties. So in particular, Tennant denies that a deflationist theory of truth need have the (non-conservativeness inducing) implication T-Bew, the generalization to the effect that all the theorems of PA are true. But since he accepts that we do have reason to recognize the truth of G, he offers, as he must, an alternative explanation of how this is possible. To cut a long story short, his favoured approach employs the adoption of (a rule of inference corresponding to) the following reflection principle for primitive recursive formulae:

PR-Reflection $\forall n Bew_{PA}(\ulcorner \Psi(n) \urcorner) \rightarrow \forall n \Psi(n)$ (where Ψ is primitive recursive).

PR-Reflection is (provably within PA) equivalent to G and Con_{PA}, which leads Tennant to conclude that it is 'just what is needed in order to be able to formalize faithfully the reasoning

[25] See Shapiro (2001).

in the [semantic argument]'.[26] For the Tennant-style deflationist, the appeal of explaining our justification for recognizing the truth of G via this principle is clear: the move allows him to accept a conservative *truth-theory*, PA_D, while, so to speak, pinning the blame for non-conservativeness on the reflection principle – a principle which (it is claimed) does not need to appeal to considerations of truth for its motivation.[27]

Ketland criticizes Tennant on this point:

> *if* Tarski's theory of truth [i.e. PA_T^+] provides at least *one* way of 'recognizing the truth of Gödel sentences', then this fact alone contradicts deflationism (for a deflationary theory of truth should be conservative). The entirely different assumption, which I did not make, that this is 'the only way' is irrelevant. (2005, p. 83)

Ketland seems here to say that the mere *existence* of a non-conservative theory of arithmetical truth that proves G poses a threat to deflationism. But this line of thought, however, only goes through on the premiss – which is the crux of the current dispute, and which I take it that Tennant will deny – that the relevant theory of arithmetical truth (in this case, PA_T^+) is properly understood as *deflationary*. The alternative, reflection-principle-based explanation of the truth of G is precisely an attempt to undermine such a premiss by arguing that the deflationist need not be committed to PA_T^+; so this attempted short way with the Tennant-style deflationist begs the question and fails to pose a real threat. As far as I can tell, there is no simple outright refutation of Tennant's strategy. Rather, its attractiveness rests on the answers to two questions: Firstly, to what extent can PR-Reflection be plausibly motivated independently of the truth-theoretic considerations it is supposed to sidestep? And secondly, how problematic is the insistence that the truth-theory fail to prove T-Bew?

I do not think the first issue poses much of a problem, although I have little to say in its support over and above what Tennant already says. In brief, the idea is that our acceptance of reflection principles is licensed by our acceptance of the theory itself; such principles are simply formalizations of our commitment to accept whatever follows from the theory. It may be possible to argue that acceptance of PR-Reflection is tacitly underwritten by some truth-theoretic considerations, but I do not see how such an argument would run.

[26] Tennant makes much of the fact that PA_T^+ proves, not only Con_{PA}, but $Con_{PA+ConPA}$, $Con_{PA+ConPA+ConConPA}$, and so on, while adding (the rule form) of PR-Reflection only enables the proof of Con_{PA}. So, he contends, the truth-theorist is guilty of using a stronger tool than is required. However, I do not find this line of thought convincing in support of Tennant's proposal. For that proposal relies on the restriction of the reflection principle to only primitive recursive sentences, and no independent motivation for such a restriction is given. If *unrestricted* reflection is accepted and iterated, then Tennant's theory too will be substantially stronger than required. See Feferman (1962), Section 5.

[27] Since Tennant makes clear that he himself is not a deflationist, there are, strictly speaking, no Tennant-style deflationists.

The second issue is, in my view, far more damaging. As has been emphasized, a central thread running through deflationism emphasizes the *expressive utility* of the notion of truth in allowing us to state and prove generalizations. Field (1999, p. 536) on this point writes: 'it is quite uncontroversial that the notion of truth can be used to make generalizations, and that these generalizations can be important to giving rise to commitments not involving the notion of truth.' But now, the sentence $\forall\phi(Bew_{PA}(\phi) \rightarrow T(\phi))$ – T-Bew – seems like precisely such a generalization. It is not as if the Tennant-style deflationist will want to *deny* that all the theorems of PA are true, in the sense of wanting a truth-theory from which its negation follows; for even on the most austere deflationist theories, i.e. those based on adding the bare instances of the T-schema, it is provable that *each individual theorem* is true.[28] Such an austere stance carries with it a sense of refusing to live up to one's commitments; and what is worse, these commitments are precisely generalizations of the kind that the notion of truth is supposed to help the deflationist formulate and derive. Consequently, Tennant's approach fails to satisfy what in Section 4.2 I called the strong desideratum for a theory of truth: although he is arguably able to explain the truth of the Gödel sentence, he is not able to carry out the semantic argument for *G* – intuitively, a cogent argument, and one that a theory of arithmetical truth ought to be able to capture. So, at the very least, Tennant's strategy incurs substantial costs, and motivates the exploration of alternatives.

5.4 First-Order Deflationism

I want now to examine Jody Azzouni's (1999) defence of what he calls the first-order deflationist – essentially, a deflationist who accepts what we have called the axiomatic conception of arithmetic, and whose understanding is consequently insufficient to rule out the eligibility of non-standard models.[29] Like Tennant, Azzouni is a proof-theoretic compatibilist: he wants to abstain from extending the induction schema to sentences containing the truth predicate and accept proof-theoretic conservativeness, which rules out the possibility of proving T-Bew. He comes down on the side of accepting what we have called PA_T. To fend off attacks based on the inability of this theory to prove certain generalizations, he states that such generalizations are not 'essential to truth.' I am not sure how convincing this line of thought is, due to the difficulty in fleshing out what is really 'essential to truth' and what is merely extraneous. For instance, Azzouni accepts the need for a theory of truth to prove generalizations

[28] Why does PR-Reflection fail to establish T-Bew even in the presence of the T-Schema? Two reasons. Firstly, note that PR-Reflection is restricted to primitive recursive sentences; T-Bew, by contrast, is a claim about *all* sentences, not just those that are primitive recursive (or of the form $\forall n\Psi(n)$ where Ψ is primitive recursive). Secondly, even if PA_D + PR-Reflection yielded all the instances of T-Bew, the further step of collecting them up into the universal generalization – T-Bew itself – would remain to be established, and as mentioned in Section 3.2 the bare instances of the T-Schema are bad at establishing generalizations concerning truth. For a rigorous proof, see Smorynski (1977).

[29] Perhaps a better name would be a 'proof-theoretic' deflationist.

such as the compositional axiom for conjunction, and it is difficult to see a principled reason why *that* kind of generalization is 'essential to truth' while others like T-Bew are not.[30]

But Azzouni's grounds for refusing to extend induction are more interesting. He notes that the intuitive considerations that Shapiro adduces in support of extending induction apply only to the *standard model* – that is, they are cogent only if one has the categorical conception of arithmetic in mind. Recall Shapiro's claim that 'the induction principle is that for any well-defined property (or predicate), if it holds of 0 and is closed under the successor function, then it holds of all natural numbers. It does not matter if the property can be characterized in the original, first-order theory'. But these considerations hardly apply to those – like proponents of the axiomatic view – who profess to acknowledge only first-order-definable properties in inductions. Proponents of such a view will indeed regard the extension of induction as ruling out perfectly legitimate models of PA. For this reason, Azzouni thinks, the first-order deflationist can escape the burden of Shapiro's strictures; and he seems to imply that only by going first-order can the deflationist so escape.

Although this response is inventive, it nevertheless has the undesirable property of committing the deflationist to a particular conception of arithmetic: the axiomatic conception. For although this may in the final analysis be the only defensible conception, it is far from obvious that this is so. Many will contend that it is simply a datum that we *do* have an intended interpretation of theories of arithmetic – *the natural numbers* – and that we *can* make sense of the distinction between standard and non-standard models. To underscore just one of the many implications of Azzouni's view, it follows that *G* is not an arithmetical truth altogether, for (even if PA is consistent) there are some non-standard models in which *G* is false! If deflationism requires an axiomatic conception of arithmetic, then we are faced with a judgement of relative plausibility: what is more plausible, that we are able to grasp the standard model of arithmetic, or that deflationism is correct? At best, deflationism is on highly controversial ground, and at worst it is landed with an impoverished and inadequate understanding of arithmetic.

Despite these critical remarks, I think that Azzouni's observation – that the extension of induction to formulas containing the truth predicate implictly appeals to considerations motivated only by the standard model – is insightful, and underscores a fundamental ambiguity in the debate about how arithmetic is to be understood. And while I do not think that being forced into adopting the axiomatic view is necessarily an attractive solution for the deflationist, attending to the ambiguity will allow the deflationist satisfyingly to respond to the non-conservativeness objection. In that connection, let us turn to the response mooted in Field (1999).

[30] On similar grounds, Shapiro (2002) thinks that the inability of the first-order deflationist to prove T-Bew is in itself a *reductio* of that position.

5.5 Diagnosing the Source of Non-Conservativeness

Like Azzouni and Tennant, Field seems to embrace the proof-theoretic conservativeness requirement. But unlike both authors, Field advocates the extension of induction to include instances involving the truth predicate. As he puts it:

> people who commit themselves to the arithmetic induction schema mean to be committing themselves not only to its instances in the language they have but to all instances in any legitimate expansion, including one with a truth predicate. (1999, p. 539)

As mentioned, one advantage of adopting a theory like PA^+_T that extends induction is that it enables the proof of generalizations such as T-Bew. This is of course not news to Field, who emphasizes that it is of a piece with a deflationist understanding of the notion of truth:

> the main point of having the notion of truth, many deflationists say, is that it allows us to make fertile generalizations we could not otherwise make; where by a fertile generalization I mean one that has an impact on claims not involving the notion of truth. (1999, p. 533)

This combination of views would appear to put Field directly in the crosshairs of the non-conservativeness argument. His response is distinctive and suggestive. In essence, he claims that the extended induction axioms derive from *facts about the natural numbers*, rather than *facts about truth*. As he puts it, what the extended induction instances involving the truth predicate depend on is 'a fact about the natural numbers, namely, that they are linearly ordered with each element having only finitely many predecessors.' (Field 1999, p. 538). In this regard, extending induction to include sentences involving the truth predicate is no different from extending induction by including any other new predicate defined in one's language: in both cases the extension is licensed by facts about the nature of the natural numbers, and in neither case is there any reason to expect the extension to behave conservatively. The strategy is, in short, to diagnose non-conservativeness as stemming from the 'arithmetic' part of 'arithmetical truth'. We might put it this way: the move from PA to PA_T is justified on truth-theoretic grounds, while the move from PA_T to PA^+_T – the move that introduces non-conservativeness – is justified on number-theoretic grounds; so it is arithmetic, and not truth, that is to blame for non-conservativeness. It is important to note that Field has in mind a categorical conception of arithmetic: the number-theoretic grounds that Field appeals to in extending induction are ones that assume the falsity of the axiomatic conception of arithmetic. The giveaway is the appeal to 'finitely many predecessors', a notion that cannot fully be captured using only

first-order proof-theoretic resources.[31] Field assumes that our understanding of the natural numbers transcends their axiomatization in first-order PA, and that it is on the basis of this understanding we are justified in extending induction.[32]

As attractive as it is, I believe that there is a serious tension in Field's view. In Section 1 I distinguished between proof-theoretic compatibilists – those deflationists who accept a proof-theoretic conservativeness constraint for truth theories – semantic compatibilists – those who accept a conservativeness constraint formulated in terms of some stronger-than-proof-theoretic notion of semantic consequence – and rejectionists – those who deny any kind of conservativeness requirement. The question is: which type does Field's response fall under? On the one hand, he seems to accept proof-theoretic conservativeness as a constraint: as he puts it, 'there is no need to disagree with Shapiro when he says conservativeness is essential to deflationism'. But the problem with this reading is that Field does in fact accept a non-conservative theory of arithmetical truth. Of course he tries to deflect the blame away from truth and on to arithmetic, but this deflection is an *excuse* for non-conservativeness, not a *denial* of it. So, it seems, Field's response is best interpreted as being outright rejectionist. For the reasons discussed in Section 5.1, I think that rejectionism is an unattractive way for the deflationist to go.

That being said, I think that Field's discussion contains an important insight: namely, that the additional strength of the theory obtained by extending induction can be attributed to arithmetic, and not truth. Unlike Field, and as I will argue in the next section, I think this insight can be used – to the extent that we have reason to adopt (as Field does) a categorical conception of arithmetic – to motivate not a rejectionist view, but a semantic compatibilist one.

VI The Disjunctive Strategy

I believe that the deflationist is in a position to mount a compelling response to the argument from non-conservativeness and to Shapiro's charge that a deflationist must move to a non-effective notion of consequence. The response stands neutral on the question of whether we have most reason to adopt a categorical conception of arithmetic or an axiomatic conception, and proceeds disjunctively depending on how this question is ultimately best resolved. The crucial idea is that on either disjunct, the view is going to end up being a compatibilist one, for both disjuncts are going to end up accepting *some* conservativeness requirement or

[31] Naturally it can be captured in first-order set theory, but the problem of non-standard interpretations for set theory is just as acute.

[32] This is not quite an accurate representation of Field's position, since he accepts a theory in which the T-Schema is used schematically in the object language; but his theory is close enough to PA^{+}_{T} that this sloppiness is justified on expositional grounds.

other. But the notion of logical consequence in which the relevant conservativeness requirement is stated – proof-theoretic or semantic – will depend on the operative understanding of arithmetic. Let me explain each disjunct in turn.

Suppose, on the one hand, that an axiomatic conception of arithmetic is most defensible. On this conception, our understanding of the subject matter of arithmetic is constituted by our understanding of a canonical first-order axiomatisation of arithmetic (let us continue to assume that this is given by the axioms of first-order PA). I take this to mean that, as far as sentences in the language of arithmetic are concerned, the axiomatic theorist is committed to accepting all and only those that are deductive consequences of PA. Our question, then, is: what truth theory should a deflationist who adheres to such a conception adopt? My answer is that such a deflationist has a principled reason to accept PA_T (or even perhaps something weaker still), and in particular to demur from accepting PA_T^+. The reason for resisting the move to PA_T^+ more or less falls out of the axiomatic characterization of arithmetic: extending induction to cover sentences containing T allows the derivation of G – a sentence that is (on this view) not licensed by the background understanding of arithmetic, since it is not derivable from the axioms.[33] Against the background of an axiomatic conception of arithmetic, then, the deflationist can and should accept a proof-theoretic conservativeness requirement with a clear conscience; for whichever of the theories of arithmetical truth weaker than PA_T^+ is accepted, it will be proof-theoretically conservative over the relevant base.[34]

But suppose, on the other hand, we can be shown to be working with a categorical conception of arithmetic, one that cannot be captured in first-order terms. If that is the case, then we must somehow be in possession of resources that enable us to rule out non-standard models as being non-standard. However these resources are best characterized, they will induce a notion of a consequence of arithmetic that goes beyond merely what can be derived from PA in a given formal proof system. But if we think such a notion of consequence is in good standing and is implicit in our conception of arithmetic, then surely *it*, and not proof-theoretic consequence, is what should figure in the relevant conservativeness requirement. And if conservativeness is understood in this way – in terms of an enriched notion of consequence – then PA_T^+ will indeed be a conservative extension, for G (and in fact, for reasons of categoricity, all other true sentences in the language of arithmetic) will be consequences of the (suitably enriched) base theory. In other words, I am recommending that on this disjunct, the defla-

[33] The situation will be one in which the axiomatic theorist is commited to accepting the disjunction $G \vee \neg G$ – as an instance of the law of excluded middle, it clearly follows from PA – but not committed to accepting either of the disjuncts. In general, I take it that there is nothing incoherent about such a stance.

[34] What about the objection that *any* adequate theory of truth for a theory T will allow the proof of the reflection principle 'all theorems of T are true'? (An objection of this form is implicit in Ketland (1999) and was pressed upon me by Jeffrey Ketland in correspondence.) I do not deny that this is an attractive feature for a theory of truth to possess; nevertheless, I believe it presupposes the falsity of the axiomatic conception of arithmetic, for it requires the acceptance of truths in the language of arithmetic that do not follow (in the relevant sense) from its axioms.

tionist accept PA^+_T (or something stronger still) and embrace *semantic* compatibilism.[35]The proof-theoretic result on which the anti-deflationist argument rests will come to appear of dubious relevance, for – given a categorical conception of arithmetic – formal derivations will be seen as a legitimate but non-exhaustive way of drawing out the genuine consequences of the axioms. There is no reason for it to be at all objectionable that the addition of a theory of truth is non-conservative when this unduly weak notion of consequence is considered. In short: it is the *conception of arithmetic as categorical, as specifying a unique mathematical structure* that is responsible for proof-theoretic non-conservativeness, not anything to do with truth.

Naturally, the deflationist taking this line needs to have something to say in reply to the reasons why Shapiro thinks that this option– that of adopting a strengthened notion of consequence – is a "thin reed" for the deflationist to take. Recall that there were two complaints: the first concerning the substantive ontological character of the additional logical resources, and the second concerning our ability to epistemically grasp the resources required to induce a non-effective consequence relation. I think the reply here ought to run basically as follows. If Shapiro's objections are cogent, then they are objections to the very possibility of a categorical conception of arithmetic. If the logical resources required for a categorical conception of arithmetic are ontologically problematic, or we are somehow epistemologically unable to grasp them, then that is a problem for *anyone* who maintains that arithmetic has an intended interpretation. But if that somewhat pessimistic diagnosis turns out to be correct, still no problem for deflationism has been disclosed: the deflationist can (like everyone else) coherently sidestep the worries by adopting an axiomatic conception of arithmetic. I claimed earlier that the cost of doing this was that it forecloses upon the possibility of a categorical

[35]The reason for the qualification that the categorical deflationist might want to move to a truth theory *stronger* than PA^+_T is simply that, whatever additional logical resources one appeals to in order to attain categoricity, one will presumably want one's eventual theory of truth to apply to sentences formulated in terms of them. (Thanks to an anonymous reviewer for raising this point and the need to address it.) And here of course PA^+_T is insufficient, for it is only a theory of truth for sentences in the *first*-order language of arithmetic, not for sentences in the newly-enriched language of arithmetic (however this enrichment is carried out). Let me take the second-order formulation of arithmetic as a representative example and mention two ways in which a truth theory can be obtained. (Both of these extensions are more naturally formulated in terms of satisfaction rather than truth, but as mentioned in §3.2, this amounts only to a cosmetic difference). Firstly, one could work against a set-theoretic background and follow Shapiro (1991, p. 72) in generalizing the notion of a variable assignment to include values of second-order variables and adding the obvious satisfaction clauses (e.g. s satisfies $\forall X \phi$ iff, for every variable assignment s' that differs from s only (at most) in its assignment to X, s' satisfies ϕ). Truth can then be defined as usual as satisfaction by all variable assignments. Or – perhaps philosophically cleaner for a committed second-order theorist – one could follow Rayo (1999, p. 7) in employing second-order quantification and handling variable assignments by taking them to be the values of second-order variables obeying certain conditions.

To be sure, either one of these approaches will require logical or mathematical machinery beyond first-order PA: the Shapiro-style approach appeals to enough set theory to assign sets or sets of n-tuples of individuals as the values of second-order variables, and the Rayo- and Uzquiano-style approach appeals to second-order quantification. But this should be neither surprising nor objectionable in the present dialectical context. It should not be surprising that a truth theory for second-order arithmetic will require resources going beyond those involved in giving a truth theory for first-order arithmetic. And appealing to such resources is legitimate, for extending one's truth theory in this way will still result in a conservative extension over PA (in the enriched sense of conservativeness that, I have argued, the deflationist on this disjunct of the response ought to adopt).

conception of arithmetic; but if the option of a categorical conception is already foreclosed upon due to (what are by hypothesis) sound objections against the notion of consequence that it involves, then the first-order deflationist position begins to look more attractive.

Someone taking Shapiro's line might respond in the following way: it is not that the additional logical resources in question are problematic in themselves, but rather, they are problematic *specifically when pressed into service by the deflationist*. However, I see no reason to think that this is the case, and no such reason can to my knowledge be found in the literature. Here I will just give some prima facie arguments in support of this claim. Take the ontological/mathematical worry first. Deflationism is in the first instance a thesis about truth, not about the abstract ontology implicated by or the substantive mathematical character of certain logical resources. If there is a problem with the ontology or mathematical character of second-order logic (or whatever other option is adopted), then this is surely a problem for everyone, not just deflationists. Shapiro tentatively advances a contrary line:

> There are second-order categorical characterizations of just about every major mathematical structure, including the natural numbers, the real numbers, the first inaccessible rank of the set-theoretic hierarchy, and beyond, well into the hierarchy of large cardinals. Therefore, truth for each of those theories can be reduced to second-order logical consequence... a critic of deflationism might respond to the second-order maneuver by arguing that the deflationist is hiding the robustness of truth in the second-order consequence relation. If the consequence relation is itself robust, then the contemplated maneuver fails to show that truth is thin and has no nature. (1998, p. 510)

However, this line of thought begs the question against the categorical deflationist.[36] That truth in arithmetic and set theory can be defined in terms of second-order consequence is a firm technical result, not in dispute. But what *is* in dispute is whether truth in these theories is robust, that is, deflationistically unacceptable. To hold that the deflationist 'is hiding the robustness of truth' in the consequence relation *presupposes* that truth in the relevant domains is indeed robust – a presupposition that simply begs the question in the present context. What the critic of deflationism needs here are independent reasons that the deflationist in particular has committments that block the adoption of the second-order consequence relation, and I am not aware that any such reasons can be given. Of course, this is not conclusive. But I believe the burden is on the anti-deflationist: *what specifically about deflationism* makes the mathematical character or ontological committments of second-order consequence (or

[36] This is not a criticism of Shapiro, for it is clear from the text that he has reservations about the argument he presents.

the alternatives mooted above) problematic? In what way does it build in a committment to inflationary conceptions of truth?

The epistemological worry fares similarly. The question is again: why is the *deflationist* in a uniquely bad position in accounting for our grasp of non-effective consequence relations? It is true that such consequence relations are highly computationally intractable, and there are very pressing questions about how epistemic agents like ourselves can grasp them (if indeed we can). I am not in possession of an account of how we are able to grasp such relations, so I cannot say for sure what the answer will look like. But if it is to pose a special problem to deflationism, it must presumably be argued that our grasp of such relations depends on a robust conception of truth; and it strikes me as difficult to construct an argument to this effect. How would conceiving of truth as correspondence to the facts, or coherence, or pragmatically, or whatever, help here? None of these conceptions of truth appear to somehow allow us to transcend the computational and epistemic limitations with which we are faced, or make it any easier to grasp a non-effective consequence relation; or at least, if they do, I cannot see how. I do not know how to refute the idea that the deflationist faces a special problem here, without providing a concrete proposal as to how the problem can be solved. But if the anti-deflationist wishes to take this line, surely he faces the burden of explaining how inflationary truth is able to provide a solution.

Let me briefly sum up the discussion of this section. The deflationist can adopt a conciliatory, disjunctive response to the argument from conservativeness, hedging his bets with respect to the question of how arithmetic ought to be understood. In advance of resolution of this question, there are two paths open to the deflationist: (i) to adopt a categorical conception of arithmetic, accept PA^{+}_{T} (or perhaps a stronger theory) as a theory of arithmetical truth, accept a strengthened notion of semantic consequence, and accept a conservativeness constraint formulated in terms of that notion; or (ii) to adopt an axiomatic conception of arithmetic, accept PA_T (or perhaps some other truth theory that is proof-theoretically conservative over PA), and accept a proof-theoretic conservativeness constraint. Which option is more attractive will depend on the tenability of categorical conceptions of arithmetic; but either way, the deflationist is in the clear.

VII Concluding Remarks: Where Does Deflationism Go From Here?

Where does this all leave deflationism? As I have argued, the deflationist who believes that our understanding of arithmetic is sufficient to rule out non-standard models has independent reason to accept logical resources stronger than proof-theoretic logic. Return again to Field's proposal of employing schematic principles and reasoning as a means of deriving the

compositional principles and solving the generality problem – this is as attractive a version of expanding one's logical resources as any. It has been charged that schematic reasoning of this form is tantamount to admitting a non-effective consequence relation, for the schematic version of induction allows the derivation (with the other axioms of PA) of all arithmetical truths whatsoever.[37] But there is a straightforward sense in which this non-effectiveness can be seen to derive from the 'indefinite extensibility' of the language (as it might be put): if we restrict the stock of predicates to a *fixed* set, the consequence relation becomes once again effective. This, I think, goes a very long way in assuaging the epistemological worries that Shapiro raises, for it is no longer as if we are faced with unrecognizable and unverifiable claims of the validity of certain inferences (as, arguably, say, full second-order logic would commit us to).

What about ontological worries, to the effect that accepting schematic reasoning commits us to recognizing an ontology of sets? The question of ontological commitment of schemata is even less resolved than the analogous question for full second-order logic (which is itself an issue of considerable controversy).[38] At the very least, it is not *clear* that employing schemata commits one to recognizing sets or anything comparably problematic, and there is a long tradition (including Quine himself) of using schemata precisely as a means of avoiding full second-order logic.

I do not mean to here pin my colours to the mast of schematic reasoning; but I do think that at the very least it provides a promising way for the deflationist to reconcile the desire for a categorical conception of the natural numbers with the need for an epistemically tractable notion of consequence.[39]

References

Azzouni, Jody 1999: 'Comments on Shapiro'. *Journal of Philosophy*, 10, pp. 541–544

Barwise, John 1977: *Handbook of Mathematical Logic*. Amsterdam: North Holland

Dummett, Michael 1963: 'The Philosophical Significance of Gödel's Theorem'. In Dummett 1978, pp. 186–214.

——1978: *Truth and Other Enigmas*. Cambridge: Harvard University Press

[37] The system is equivalent to a subsytem of second-order logic allowing the expression of Π^1_0 sentences. See Shapiro (2002) for discussion.

[38] See McGee (1997) for an influential paper advocating the thesis that schemata avoid certain undesirable commitments of second-order logic; and see Pedersen and Rossberg (2010) for an opposing argument to the effect that they are on the same footing.

[39] Many thanks to Ralf Bader, Cian Dorr, Hartry Field, Yu Guo, Paul Horwich, Jeffrey Ketland, Penn Lawrence, Harvey Lederman, Noel Swanson, Jared Warren, Crispin Wright, and Mike Zhao for helpful discussion and comments. Thanks are also due to two anonymous reviewers and the editor of Mind for valuable comments and feedback.

——1994: 'Reply to Crispin Wright'. In McGuinness and Oliveri (eds) 1994, pp. 329–38

Feferman, Solomon 1962: 'Transfinite Recursive Progressions of Axiomatic Theories'. *Journal of Symbolic Logic*, 27, pp. 259–316.

Field, Hartry 1999: 'Deflating the Conservativeness Argument'. *Journal of Philosophy*, 10, pp. 533–540.

——2001: *Truth and the Absence of Fact*. Oxford: Clarendon Press.

Gabbay, Dov and Franz Guenthner 2001: *Handbook of Philosophcal Logic, Volume 1 (2nd edition)*. Dordtrecht: Kluwer.

Gupta, Anil 1993: 'A Critique of Deflationism'. *Philosophical Topics*, 1, pp. 57–81.

Halbach, Volker 2001: 'How Innocent is Deflationism?' *Synthese*, 1, pp. 167–194.

——2011: *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.

Heck, Richard Forthcoming: 'The Logical Strength of Compositional Principles'. *Notre Dame Journal of Formal Logic*.

Horwich, Paul 1990: *Truth*. Oxford: Oxford University Press.

Horsten, Leon and Volker Halbach (eds) 2002: *Principles of Truth*. Frankfurt: Hänsel-Hohenhausen

Kaye, Richard 1991: *Models of Peano Arithmetic*. Oxford: Clarendon Press.

Ketland, Jeffrey 1999: 'Deflationism and Tarski's Paradise'. *Mind*, 429, pp. 69–94.

——2005. 'Deflationism and the Gödel Phenomena: Reply to Tennant'. *Mind*, 453, pp. 75–88.

Loewer, Barry 2005: 'On Field's "Truth and the Absence of Fact": Comment'. *Philosophical Studies*, 1, pp. 59–70.

McGee, Vann 1997: 'How We Learn Mathematical Language'. *Philosophical Review*, 1, pp. 35–68.

McGuinness, Brian and Gianluigi Oliveri (eds) 1994: *The Philosophy of Michael Dummett*. Dordtrecht: Kluwer.

Pedersen, Nikolaj and Rossberg, Marcus 2010: 'Open-endedness, Schemas and Ontological Commitment'. *Noûs*, 2, pp. 329–339.

Quine, Willard V.O. 1986: *Philosophy of Logic*. Cambridge: Harvard University Press.

Rayo, Agustin and Gabriel Uzquiano 1999: 'Toward a Theory of Second-order Consequence'. *Notre Dame Journal of Formal Logic*, 40(3), pp. 315–325.

Shapiro, Stewart 1991: *Foundations Without Foundationalism: A Case for Second-Order Logic*. Oxford: Oxford University Press.

——1998: 'Proof and Truth: Through Thick and Thin'. *Journal of Philosophy*, 10, pp. 493–521.

——2001: 'Systems Between First- and Second-Order Logic'. In Gabbay and Guenthner (eds) 2001, pp. 131–188.

——2002: 'Deflation and Conservation'. In Horsten and Halbach (eds) 2002, pp. 103–128.

Smorynski, Craig 1977: 'The Incompleteness Theorems'. In Barwise 1977, pp. 821–865.

Tennant, Neil 2002: 'Deflationism and the Gödel Phenomena'. *Mind*, 443, pp. 551–582.