The innocence of truth

Cezary CIEŚLIŃSKI[1]

ABSTRACT

One of the popular explications of the deflationary tenet of 'thinness' of truth is the conservativeness demand: the declaration that a deflationary truth theory should be conservative over its base. This paper contains a critical discussion and assessment of this demand. We ask and answer the question of whether conservativity forms a part of deflationary doctrines.

1 Introduction

The claim that the notion of truth is innocent or 'metaphysically thin' has been traditionally attributed to philosophers defending a deflationary view on truth. Various explications of this position have been proposed in the literature. One of them, attributed to Horsten (1995), Shapiro (1998) and Ketland (1999), is that an adequate theory of truth for a given language conservatively extends a base theory of syntax for this language. This proposition has been debated quite extensively in recent literature. In particular, two interpretations of the phrase 'conservatively extends' have been proposed: a syntactic and a semantic (or model theoretic) interpretation, each contending for the status of a demand to be imposed on a satisfactory (deflationary) theory of truth.

As it happens, conservativity claims were put forward not by the deflationists themselves, but by their *opponents*. It is the critics (not the deflationists) who insisted that conservativity is a good explication or a commitment of the deflationary standpoint.[2] It seems, however, that since then the conservativity requirement has taken a life of its own: the merits and demerits of conservative truth theories have been debated quite independently of the real connection between the requirement in question and the deflationary tenets.[3]

[1] Institute of Philosophy, University of Warsaw, Poland. Email: c.cieslinski@uw.edu.pl.

[2] In effect, the critics claim that: (1) deflationary truth theories should be conservative, but (2) they cannot be! In this paper we concentrate exclusively on (1), leaving (2) aside. As for (2), let us just mention in passing that, according to Shapiro and Ketland, adding the truth axioms to a base theory *B* should permit us to prove a strong version of the reflection principle; namely, the formal analogue of the sentence "all theorems of *B* are true" should become provable, with the point being that such an adequacy condition excludes conservativity. This adequacy condition is also endorsed by Hannes Leitgeb (2007). We do not purport to analyse arguments for (2) here.

[3] E.g. Halbach (2011) writes: "Here I simply take it for granted that at least nowadays some authors take it that some conservativity claim forms an integral part of deflationist doctrines. At any rate, even independently of the discussion about deflationism, the question of whether truth theories are conservative over their base theories does bear philosophical significance." (p. 313) For further discussion, see also (Tennant, 2005) and (Cieśliński, 2010).

It is the aim of the present paper to fill in this gap: to consider *how exactly* conservativity is related to deflationism. Instead of taking conservativity for granted in a move characteristic of a large bulk of recent literature treating the conservativity proposal as just one more variant of deflationism, I investigate the question of whether attributing conservativity claims to deflationists was legitimate in the first place. In what sense, if any, do such claims form a part of deflationist doctrines? Is conservativity implied or supported by more traditional deflationary views? This is the principal question which will be handled here.

2 Deflationists on the innocence of truth

What is the basis for attributing claims of 'thinness of truth' to the deflationists? The aim of this section is to gather some samples from the deflationary literature, which are typically forwarded as justification for such an attribution. More precisely, we are to present below a selection of quotes from deflationary philosophers, providing textual basis for ascribing to (some) deflationists the following tenets:

(1) truth is not a property,
(2) truth is unlike other properties in that it has no nature,
(3) truth is a logical concept,
(4) truth is insubstantial,
(5) truth plays no role (or a very limited one) in explanations and justifications of non-semantic facts.

It should be acknowledged that the logical relations between (1)-(5) are far from obvious. The situation is even aggravated by the fact that sometimes these views are presented more as slogans than as well-argued and precise claims. Nevertheless, we find it useful to start with such a presentation. Before imposing demands on deflationary truth theories, it is advisable to examine what the deflationists were actually saying!

Starting with (1), here is a quote from (Ayer, 1935):

> If I say that it is true that Shakespeare wrote *Hamlet*, or that the proposition "Shakespeare wrote *Hamlet*" is true, I am saying no more than that Shakespeare wrote *Hamlet*. Similarly, if I say that it is false that Shakespeare wrote the *Iliad*, I am saying no more than that Shakespeare did not write the *Iliad*. And this shows that the words 'true' and 'false' are not used to stand for anything, but function in the sentence merely as assertion and negation signs. That is to say, truth and falsehood are not genuine concepts. Consequently there can be no logical problem concerning the nature of truth.

As we see, according to Ayer, there is no (genuine) concept of truth and there is also nothing that the word "true" stands for. In particular, "true" doesn't stand for a property of anything.[4]

Not all deflationists share Ayer's opinion. In particular, Paul Horwich disagrees:

> It is not part of the minimalist conception to maintain that truth is not a property. On the contrary, 'is true' is a perfectly good English predicate - and (leaving aside nominalistic concerns about the

[4] Other classical references are (Strawson, 1949) and (Grover, 1992): both authors claim that calling a sentence true does not involve ascribing the property of being true to this sentence.

> very notion of 'property') one might well take this to be a conclusive criterion for standing for a property of some sort. What the minimalist wishes to emphasize, however, is that truth is not a complex or naturalistic property but a property of some other kind. (Horwich 1999, p. 37)

Rejecting (1), Horwich explicitly embraces claim (2), stressing that truth is a property of a special kind: one without an underlying nature.

> Unlike most other properties, being true is unsusceptible to conceptual or scientific analysis. No wonder that its 'underlying nature' has so stubbornly resisted philosophical elaboration; for there is simply no such thing. (Horwich 1999, p. 5)

The lack of an 'underlying nature' is explained by Horwich in the following way:

> truth is entirely captured by the initial triviality, so that in fact nothing could be more mundane and less puzzling than the concept of truth (Horwich 1999, p. ix)

Here the 'initial triviality' is Tarski's famous T-schema: φ is true if and only if φ. Horwich's point here is that nothing deeper than that is needed to account for the way in which we are using the truth predicate. In fact, the search for a deeper theory is a misguided idea.

According to Horwich, "the truth predicate exists solely for the sake of a certain logical need" (ibid., p. 2), namely, the need to express generalisations.[5] Other contemporary deflationists have, in addition, called truth quite explicitly a logical notion:

> I would say that the most fundamental notion of truth is a purely logical notion applicable only to sentences we understand, and serving solely as a device of generalization; all other notions of truth for example, those applicable to other languages are to be explained in terms of this, plus relatively modest auxiliary notions (Field 1999, p. 534)

As for tenet (4) – the one concerning the insubstantiality of truth – Horwich writes:

> The claim that truth is not a complex or naturalistic property – that it is 'unreal' or 'insubstantial', in the sense advocated by minimalism – must not be confused with the idea that truths are unreal, or, in other words, that no sentence, statement, or belief is ever true. (Horwich 1999, p. 52)

Shortly afterwards he refers to the substantiality of truth in terms of "the association of the truth predicate with some beefed-up, highly esteemed metaphysical or epistemological property" (see p. 53). Additionally, in the Postscript to the second edition of his book, Horwich characterises a 'substantive' property as "the sort of property for which there might well be a constitution theory of the form '*x* is true = *x* is *F*'" (ibid., p. 143). On the same page, he also stresses that "no reductive theory of truth is likely to be correct".

To sum up: as I understand, there are two elements to Horwich's interpretation of the phrase 'insubstantiality of truth'. *First*, Horwich denies both the need and the possibility of defining truth as a naturalistic, metaphysical or epistemological property *F*.[6] *Second*, in spite of this, truth

[5] For example, instead of asserting separately the instances of the law of excluded middle (there are infinitely many of them!), we can assert a single sentence with the truth predicate: 'All the instances of the law of excluded middle are true'. Cf. (Horwich 1999, p. 4).

[6] The expressions "naturalistic", "metaphysical" and "epistemological" are left undefined by Horwich, although some examples are given in the following passage: "Amongst the products of this traditional point of view there is the correspondence theory (*x* is true iff *x* corresponds to a fact), the coherence theory (*x* is true iff *x* is a member of a coherent set of beliefs), the verificationist theory (*x* is true iff *x* is provable, or verifiable in ideal conditions), and the pragmatist theory (*x* is true iff *x* is useful to believe).

is not a mysterious notion. On the contrary, it can be easily and adequately characterised by means of a principle which is both *very simple* and *epistemologically basic* – namely, by Tarski's T-schema. This is what I take to be the content of Horwich's notion of insubstantiality.[7]

As for tenet (5), let us start with the following quote from a deflationary philosopher:

> On this issue, contemporary deflationists are in broad agreement: the function of truth talk is wholly expressive, thus never explanatory. As a device for semantic assent, the truth predicate allows us to endorse or reject sentences (or propositions) that we cannot simply assert, adding significantly to the expressive resources of our language. Of course, proponents of traditional theories of truth do not deny any of this. What makes deflationary views deflationary is their insistence that the importance of truth talk is exhausted by its expressive function. (Williams 1999, p. 547)

The key contrast to be observed in the quoted fragment is that made between explanation and expressiveness. It is stressed that a given notion (the notion of truth in particular) can be powerful in one respect but not in the other; that it can have a large expressive power without possessing any explanatory value.
Other deflationists were more cautious in this respect. Thus, Horwich wrote:

> truth does indeed enter into explanatory principles, but their validity may be understood from within the minimal theory. (Horwich 1999, p. 45)

In considering explanatory principles like "the truth of scientific theories accounts for their empirical success", Horwich treats them as generalisations of concrete observations of the following type:

> The theory that nothing goes faster than light works well because nothing goes faster than light.

Equivalently, one could say:

> The theory that nothing goes faster than light works well because it is true.

Yet in this context, Horwich adds:

> No further explanatory depth is achieved by putting the matter in terms of truth. Nonetheless, use of the truth predicate in this sort of context will often have a point. What it gives us is a certain economy of expression, and the capacity to make such explanatory claims even when we don't explicitly know what the theory is, or when we wish to generalize, e.g.
>
> True theories yield accurate predictions.

But nothing of this sort has ever survived serious scrutiny - which comes as no surprise to the deflationist, who denies that there is any prospect of an explicit definition or reductive analysis of truth, even a very approximate one" (Horwich 1999, pp. 120-121).

[7] However, some other deflationists did not place much weight on the idea of insubstantiality of truth. Here is a disparaging remark from Hartry Field: "[Putnam's and Wright's] discussions are directed at a version of deflationism which says that truth is not a property (Putnam) or not a "substantial property" (Wright), and *I'm not clear enough as to what that is supposed to mean*" (Field 1994, p. 265; emphasis mine).

> But these are precisely the features of truth that are central to the minimalist conception. Clearly they can provide no reason to go beyond it. (Horwich 1999, p. 49)

In effect, according to Horwich, the notion of truth functions in explanations in exactly the same manner as in other contexts. Whenever it is needed, if at all, it is because it permits us to express generalisations, which would otherwise remain inexpressible. The recourse to truth brings no explanatory depth.

3 Introducing conservativity: definitions and intuitions

We start with defining two notions of conservativity below, which will be evaluated as tools for the deflationist to explicate his position. One notion is syntactic: a conservative extension does not prove new theorems of the base language. The second is semantic and concerns the possibility of expanding models of the base theory.

Definition. *Let T_1 and T_2 be theories in languages L_1 and L_2 (with $L_1 \subseteq L_2$). Then:*

(a) *T_2 is syntactically conservative over T_1 iff $T_1 \subseteq T_2$ and $\forall \psi \in L_1 [T_2 \vdash \psi \rightarrow T_1 \vdash \psi]$.*
(b) *T_2 is semantically conservative over T_1 iff every model M of T_1 can be expanded to a model of T_2 (i.e. interpretations for new expressions of L_2 can be provided in M in such a way as to make T_2 true in the expansion of M).*

The two notions of conservativeness do not coincide. Semantic conservativeness is a stricter notion: it gives, via completeness theorem, the syntactic version[8], but the opposite implication does not hold. Examples of truth theories which are syntactically but not semantically conservative over their base theories will be given below.

Both notions are invoked by Shapiro (1998). Shapiro's motivation for accepting the conservativeness demand is succinctly formulated in the following fragment:

> How thin can the notion of arithmetic truth be, if by invoking it we can learn more about the natural numbers? (Shapiro 1998, p. 499)

In the next step the notion of conservativity is introduced. A representative fragment from Shapiro's paper runs as follows:

> 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 Φ in the original language (so that Φ does not contain the truth predicate) such that Φ 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 Φ false, but it is not logically possible for the axioms of B' to be true and Φ false. This undermines the central deflationist theme that truth is in-substantial. (Shapiro 1998, p. 497)

In the quoted passage, the claim of insubstantiality of truth is explicated in terms of syntactic conservativeness (even though the first sentence concerns just 'one form or another' of the

[8] That is, if T_2 is semantically conservative over T_1, then it is also syntactically conservative. Otherwise with $T_2 \vdash \psi$ and ψ being unprovable in T_1, there would exist a model of $T_1 + \neg\psi$. Obviously such a model could not be expanded to a model of T_2, contradicting the semantic conservativity assumption.

conservativeness demand).[9] A syntactically non-conservative truth theory permits Karl to 'learn more' about natural numbers, since it may lead him to accept a sentence Φ which, without the truth axioms, would remain unprovable.

On the other hand, what sort of motivation could stand behind the semantic conservativeness requirement? Imagine that Karl accepts a base theory B which admits a model M. Assume that he adds a truth predicate and extends B to a theory B', which is not semantically conservative over B – in particular, it excludes M. In other words, before introducing the notion of truth, M was possible, but afterwards M is out of the question. The truth predicate, as introduced by Karl, admits no interpretation in such a world. How thin can the notion of truth be if by invoking it we eliminate some previously possible interpretations of our base theory?[10]

For a more vivid illustration, imagine that Karl is an arithmetician, inhabiting some (possibly nonstandard) world M. All arithmetical sentences accepted by Karl are, as it happens, true in M. Then one day Karl has an excellent idea: he extends his language with the truth predicate and accepts new truth axioms (with perhaps a typical, deflationary motivation of enlarging the expressive power of his language). What may happen is that – unbeknownst to Karl – the truth predicate introduced via these axioms has no interpretation in his world. How thin can the notion of arithmetical truth be if, just by invoking it, Karl can end up with a theory with no interpretation in the world he inhabits? This is another formulation of semantic conservativeness intuition.

Viewed in these terms, both conservativity demands – semantic and syntactic – seem to be on a par. The intuitions standing behind them look very similar. Initially one might suspect that the deflationist's choice boils down to a simple acceptance or rejection of both of them. However, in the section following we will observe that this symmetry breaks down. There are important differences in philosophical argumentation for conservativity in both cases.

In order to appreciate what is at stake, in the next section we will review some basic non/conservativity results for truth theories.

4 Non/conservative truth theories

Accepting conservativity demands severely restricts the deflationist's choice of truth theories. In this section we illustrate these restrictions with some examples.

Here is the first observation to be made. If the innocence of truth is to be identified with semantic conservativity, then truth theories with full induction for the extended language (containing the truth predicate) are never innocent. In effect, even quite simple theories would be beyond the deflationist's reach.

The problem already starts at the level of very basic, disquotational conceptions of typed truth. Define TB^- as the result of extending PA with all the arithmetical substitutions of the local disquotation schema,[11] i.e.:

[9] At least, that is, if Karl's theory is first order. Of course, the consequence relation mentioned by Shapiro may be semantic, but in the first order case it will coincide with the syntactic one.

[10] Cf. (Strollo 2013, 529-533), especially footnote 44. Incidentally, we may also note the same effect in an example given by Shapiro. By adopting a syntactically non-conservative extension, Karl eliminates some models of the base theory; namely, all models in which Φ is false. In both cases introducing the truth predicate results in narrowing down the scope of possible interpretations of Karl's theory.

[11] The expression 'local disquotation schema' is typically used in the literature in reference to the schema of Tarskian biconditionals for sentences, exactly as in the definition of TB^-. In contrast, the expression "global disquotation schema" (or "the schema of uniform disquotation") refers to a more general case with quantifiers, i.e. to the schema "$\forall x_1 \dots x_n [T(\ulcorner \phi(x_1 \dots x_n) \urcorner) \equiv \phi(x_1 \dots x_n)]$".

$$TB^{-} = PA \cup \{T(\ulcorner \varphi \urcorner) \equiv \varphi: \varphi \text{ is a sentence of the language of PA}\}.$$

By an easy model-theoretic argument, TB^{-} is semantically conservative over PA. However, adding extended induction changes the situation quite drastically. Define TB as the extension of TB^{-} containing all the axioms of induction for the extended language. Although TB is still syntactically conservative over PA,[12] it doesn't satisfy the semantic conservativity condition.

Fact. TB is not semantically conservative over PA.[13]

The final effect is that an even purely disquotational, typed notion of truth combined with extended induction becomes inaccessible to the deflationist accepting the semantic conservativity demand.[14]

It is worth stressing in this context that non-conservativeness phenomena are not associated solely with compositional truth theories.[15] This point stands in contrast to the views expressed by some authors; for example, Ketland wrote:

> it seems that it is the compositionality of the principles governing truth which explains non-conservativeness, as the disquotational truth theory does remain conservative when induction and other schemes are extended. (Ketland 2010, p. 427)

and also:

> it seems to me that the compositional truth theory lies at the root of non-conservativeness, and if the conservativeness condition is correct, then compositional truth is non-deflationary or 'substantial' (Ketland 2010, p. 435)

Indeed, it is easy when concentrating solely on syntactic conservativity to overlook the insight about the model theoretic strength of extended induction. From this perspective, there is simply no difference even between TB^{-} and some fully inductive theories with uniform disquotation.[16] Even quite independently of any discussion about deflationism, it is worth stressing that model theoretic considerations can provide a useful measure for comparing truth theoretic (as opposed to arithmetical) strength of theories.[17]

---

[12] The simplest argument, to my knowledge, consists in observing that in a given proof of an arithmetical sentence in TB, all occurrences of '*T*' can be replaced by a suitable arithmetically definable, partial truth predicate. For the details here, see e.g. (Halbach 2011, 55-56).

[13] This observation is due to Fredrik Engström. For the proof, see (Strollo 2013) or (Cieśliński 2015).

[14] The observation can be generalised to the untyped theories of disquotational truth proposed in the literature. Since they contain TB, they are not semantically conservative over PA. For a discussion of untyped disquotational theories, see (Halbach 2009) and (Cieśliński 2011).

[15] It is a well known fact that some disquotational (non-compositional), untyped theories are not syntactically conservative over PA, with the theory PUTB being a famous example (see Halbach 2009). The impact of the present remarks is that some non-conservativity phenomena are also visible in the case of typed theories.

[16] The schema of uniform disquotation has a form "$\forall x_1 \dots x_n [T(\ulcorner \phi(x_1 \dots x_n) \urcorner) \equiv \phi(x_1 \dots x_n)]$". Adding to PA all arithmetical substitutions of this schema together with extended induction produces a theory denoted as UTB ("uniform Tarski biconditionals").

[17] Comparing TB with UTB provides a useful example. On the one hand, it is known that only recursively saturated nonstandard models of PA can be expanded to models of UTB. On the other hand, this is not

The second observation is that syntactic (and even semantic!) conservativeness can be squared with compositionality, although typically the price consists in sacrificing extended induction. For syntactic conservativity, a classical example is the typed compositional theory CT^-. Its axioms are the standard Tarski-style clauses which characterise the notion of truth for arithmetical sentences. We have the usual condition for atomic sentences (an expression of the form $\ulcorner t = s \urcorner$ is true iff the values of the terms t and s are the same); other axioms state that truth commutes with the logical connectives and quantifiers (e.g. the conjunction of arbitrary arithmetical sentences is true if and only if both conjuncts are true). It has been shown that CT^- is syntactically conservative over PA.[18] However, this is not a conservative extension of PA in the semantic sense.[19] On the other hand, the theory CT, obtained by adding full extended induction to CT^-, is arithmetically stronger than PA. By an easy folklore result, it is possible to formalise in CT a natural argument for the consistency of PA, which starts with the observation that all axioms of PA are true and proceeds via the fact that truth is closed under provability. The use of induction for formulas of the extended language is quite crucial in this reasoning.

An example illustrating the possibility of squaring semantic conservativity with compositionality is provided by the theory KF^- ('Kripke-Feferman'). The untyped compositional truth theory KF has been proposed as an axiomatisation of the Kripkean approach to self-referential truth.[20] We denote as KF^- the theory with the same truth axioms but with arithmetical induction only. It is known that KF^- is semantically conservative over PA (on the other hand, full KF is arithmetically much stronger).[21] In effect, even though the most basic typed compositional theory CT^- is not semantically conservative, with a different choice of compositional axioms not only syntactic but semantic conservativity can also be attained. However, in both cases the price to be paid lies in the sacrifice of the extended induction.

Finally, let's observe that squaring semantic conservativity with extended induction, while tricky, is to some extent possible. Provability of all arithmetical substitutions of the T-schema seems to be a minimal condition for an adequate theory of truth. If a given theory does not satisfy this demand, one could even wonder whether there is a good reason to call it a theory of truth, and not of something else. We have already seen that with full extended induction we lose semantic conservativity. However, some weak variants of extended induction preserve conservativity.[22] This gives the deflationist some hope: instead of rejecting out of hand extended induction as a part of his theory of truth,[23] he could try to find a persuasive defence of a particular version of induction incorporated into his theory. But let us also observe that an appeal to semantic conservativity cannot be a crucial part of such a defence, unless we have in

---

true about TB: there are nonstandard models of Peano arithmetic which are not recursively saturated but which can be expanded to models of TB (an unpublished result by M. Łełyk and B. Wcisło).

[18] This observation is a corollary to a classical theorem by Kotlarski, Krajewski, and Lachlan (1981), who showed that a satisfaction class can be constructed in an arbitrary countable, recursively saturated model of Peano arithmetic. Originally the construction has been carried out for relational arithmetic; see (Kaye 1991) for a proof for the language with function symbols. More recently, Enayat and Visser (2015) constructed a much simpler conservativeness proof.

[19] Only recursively saturated models of arithmetic can be expanded to models of CT^-. The result is due to Lachlan; see (Kaye 1991, Theorem 15.5, p. 228).

[20] For the list of axioms, we refer the reader to (Halbach 2011, p. 201).

[21] See (Cantini 1989, Proposition 5.8 and Corollary 5.9).

[22] See e.g. (Fischer, 2009), where a compositional, semantically conservative truth theory PT^- is introduced. The theory contains a weak form of extended induction for the (so called) total arithmetical formulas.

[23] Such a manoeuvre is quite difficult to justify; see the discussion in (Horsten 2011, p. 83ff).

advance a decent philosophical justification of the semantic conservativity demand. Well, do we have it?

5 Semantic conservativity – a philosophical assessment

Model theoretic conservativity has only recently started to be taken seriously as a desirable property of a truth theory.[24] To my knowledge, very few authors have tried to provide a careful argument for such demand for a theory of truth. The aim of the present section is to consider such arguments. Why should we demand semantic conservativity? In a recent paper, A. Strollo claims that such a demand is indeed justified:

> When we are interested in the deflationary metaphysics of truth […] what should really matter is whether every model of the base could be expanded to a model of the base theory plus a truth theory. (Strollo 2013, p. 530)

Otherwise, in his opinion

> truth would exhibit extralinguistic effects. It would affect the things the language talks about and not just our way of speaking of them. Not only would we operate at a linguistic level […], we would also need to intervene into the domain by changing and shaping it. In this sense, and in open contrast with the deflationist claim, the property of truth would enter reality as a robust ingredient. (ibid., p. 530)

However, the problem with this mode of thinking is that no link has been established with what the deflationists were actually saying. One may *decide*, of course, to define 'robust' in such a way that semantically non-conservative truth will emerge as robust. However, the question still remains whether the "contrast with the deflationist claim" is real or illusory. In effect, it is still unclear why it should matter to the deflationist whether his theory excludes some models.

Strollo himself doesn't answer this question. He remarks only that "good reasons had been put forward" for conservativeness demand, referring the reader to (Shapiro, 1998). We will return to Shapiro's argumentation later on, for the moment let us just note that in the present context this is hardly satisfactory. Shapiro's primary approach was epistemological (recall his "how thin can the notion of arithmetic truth be, if by invoking it *we can learn more* about the natural numbers") and it is far from clear how, if at all, it can be applied in the course of arguing for a particular conception of "deflationary metaphysics of truth".

As I see it, the semantic conservativity requirement encounters two major problems. So far I've stressed just the first of them; namely, that no sound textual basis has been given for imposing such a demand on truth theories proposed by the deflationists. This in itself doesn't have to be fatal. One could still propose semantic conservativity as the explanation of robustness *consistent* with what the deflationists were actually saying, while arguing that, for certain reasons, it's the best possible explanation doing justice to their words. There is,

[24] See (McGee, 2006) and (Strollo, 2013); cf. also (Fischer 2009, p. 814): "For example the theory PT^- has not been discussed although it seems to be a possible candidate for deflationism and it has some nice features other theories lack, like finite axiomatizability, conservativity over PA, and noninterpretability in PA. The investigation of these different minimal theories could help to establish criteria of adequacy for a deflationist theory of truth and clarify claims like truth is 'not substantial', the truth predicate has only an 'expressive' function and no 'explanatory' one." The conservativity mentioned in the quoted fragment is model-theoretic.

however, also the second and more serious difficulty: the requirement doesn't fit well with other deflationary doctrines.

Contemporary deflationists favour an axiomatic approach to truth. The notion of truth is to be characterised by means of simple axioms (e.g. disquotational ones), playing a role of epistemologically basic meaning postulates. In addition, the deflationists also put forward a negative claim: they claim that no other notion of truth is needed (that is, other than the one characterised by their axioms) Any concept of truth which goes beyond this is ill-conceived at worse and at best not needed.[25]

In order to illustrate the problem with semantic conservativity, consider the case of Peano arithmetic. Imagine that we extend it with truth axioms of our choice and claim that these axioms characterise the notion of arithmetical truth in such a way that no other notion of arithmetical truth is needed. Now, why should it matter whether our theory is semantically conservative over PA? Or to put it differently: why should the model-theoretic notion of truth have this sort of importance?

To dispel possible misunderstandings, let me emphasise that I'm ready to grant the deflationist free and full access to model theoretic tools and resources. The questioning of set theory – with model theory viewed as a fragment of it – is not, after all, a part of deflationary doctrines. It is the 'heavy' notion of truth, not classical mathematics, that is the focus of deflationary criticism! I take it as uncontroversial that the deflationist may use the model-theoretic apparatus. In effect the question is not *whether* he is permitted to use it (of course he is!) but *how* he may use it.[26]

In a nutshell, my answer to the last question runs as follows: the deflationist may freely use model theory as a technical tool. He can use models in completeness or conservativeness proofs, just as he can engage (if he wants) in other sorts of set theoretic investigations. But here comes the crucial limitation: what he cannot do is to describe arithmetical truth – truth *simpliciter* – as truth in some chosen (intended) model of arithmetic, while treating the last notion as indispensable and primary. He claims, after all, that arithmetical truth *simpliciter* is fully characterised by nothing other than his basic truth axioms! Moreover, he claims that this is the only notion of arithmetical truth which we require, and it is exactly these claims that would be compromised by the identification of truth *simpliciter* with truth in the intended model of arithmetic.

As an illustration of this danger, consider the following passage from McGee (2006), containing a plea for a condition stronger than syntactic conservativity requirement:

> [Syntactic] conservativity is too weak because it permits us to accept theories that are plainly incompatible with the meanings of the arithmetical terms. (McGee, 2006)

McGee asks us to consider an extension of PA built in a language with one additional predicate symbol "F". New axioms characterising F are F(0), F(1) … etc. (for each numeral), together with the formula "$\exists x \neg F(x)$". The resulting theory is syntactically conservative over PA (even if we extend induction to cover also formulas containing "F"). However:

[25] Cf. (Horwich, 1999). According to Horwich, truth theory should be axiomatised by nothing more than T-equivalences. On page 10 he claims that the traditional, inflationary approaches to truth "do not typically impugn the correctness of the equivalence schema […] but question its completeness. They deny that it tells us about the essential nature of truth, and so they inflate it with additional content in ways that, I will argue, are, at best, unnecessary and, at worst, mistaken."

[26] I assume here that the deflationist is concerned *solely* with the notion of truth simpliciter, and not with the notion of truth under an interpretation (or truth in a model). For more about this assumption, see the final paragraphs of this section.

> Even though this theory doesn't reveal its mendacity by entailing explicit falsehoods, we surely shouldn't accept it, since there is no way to partition the numbers into Fs and non-Fs so as to make the theory true. (McGee, 2006)

The expanded theory, although syntactically conservative over PA (and therefore consistent), is ω-inconsistent. The symbol F cannot be interpreted in such a way as to make the new axioms true, while preserving the usual meaning of arithmetical expressions. In other words, it is not possible to interpret the expanded theory in the intended model of arithmetic. Before adding F, the intended interpretation was possible, but now it has become out of the question. Since the intended interpretation is desirable, a theory which excludes it should be deemed inadequate.

The first remark to be made is that considerations of this sort are far from enough to justify the semantic conservativity demand for truth theories. Why should we require semantic conservativity if it is only the intended interpretation that matters? In other words, why should we demand the admissibility of *all* models, and not just the intended one? Even if we accept that the truth axioms added to a given base theory shouldn't exclude the intended model of the base theory, the question remains - why shouldn't they also exclude other models? As we have seen, the semantic conservativity demand eliminates such (syntactically conservative) extensions of PA as TB and CT^{-}. Nevertheless, the intended model of arithmetic *can* be expanded to models of these theories. Why should such theories be eliminated if it is only some deviant models which are made inadmissible by them?

It is, however, the second problem which is really damning, at least from the deflationist's perspective. With this way of thinking, what we care about is the intended interpretation – the notion of truth in the intended model. It is exactly this interpretation which shouldn't be excluded by our truth axioms. In effect, a notion of truth other than the one characterised by the axioms (namely, the notion of truth in the intended model), turns out to be needed, if only to justify the conservativity requirement imposed on the theory of truth. In other words, the problem is that in considerations of this sort, the notion of truth in the intended model is treated as our primary concept of truth *simpliciter*. That is why the deflationists would be extremely ill advised to engage in such argumentation.

Can we do better than that? Is there an argument for semantic conservativity which does not make the notion of the intended model primary? Well, I don't think so.

There are indeed some options to be considered. Instead of using the notion of the intended model, one could build an argument which takes *scepticism* about this notion as a starting point. In effect, one would postulate semantic conservativity exactly because the notion of the intended model is found problematic. The sceptic asks which model of arithmetic is to be singled out as the intended one and how it is to be done. The concerns begin with the observation that, as users of a given arithmetical theory (say, Peano Arithmetic), we are unable to differentiate between models. Our deductive apparatus or even our use of arithmetical concepts in science, does not fix uniquely a model which we could call "intended". An appeal to stronger theories, employing second order logic and guaranteeing categoricity, might not satisfy the sceptic either. He could claim, for instance, that the notion of a full power set of an infinite set, assumed in such an argumentation, is much more dubious than our idea of a natural number.[27] In view of this claim, the sceptic about the intended model could urge us to accept a

[27] See for example (Halbach and Horsten 2005, p. 176): "However, any kind of second-order approach will make use of the power set of the set of natural numbers. This power set, we submit, is far more problematic than the notion of the natural number itself. For the independence phenomena revealed by Gödel and Cohen suggest that the notion of the power set of the natural numbers may be inherently indeterminate or essentially relative." A similar opinion is expressed by Gaifman: "The absoluteness of

theory of truth which excludes no models, and that is, in effect, the semantic conservativity demand.

However, the above remarks are rather vague and they still leave unclear the exact shape of the argument supporting the semantic conservativity demand. One possible line of reasoning might run as follows:

1. Arithmetical truth is truth in some model (or a class of models) of Peano Arithmetic, corresponding to a fragment of real world.
2. We have no way of recognising models which do not correspond to a fragment of the real world.
3. A theory which excludes some models risks excluding the model corresponding to the real world; in effect there is a risk that it will not play the role of the theory of arithmetical truth *simpliciter*.
4. Therefore, all models should be treated on a par. A theory which does not satisfy the semantic conservativity requirement is not to be adopted.

Unfortunately, from the deflationist's perspective, this reasoning is again very problematic. The main difficulty is the same as before: the argument employs a notion of truth which goes beyond deflationary axiomatic characterization. Arithmetical truth is presented in premise 1 as "truth in a model"; indeed, as truth in some rather special model, which could (why not?) be called the intended one. The argumentation requires, in effect, that the notion of the intended model makes sense and that the notion of truth in the intended model makes sense as well (scepticism is only to be declared in regard to our possibilities of recognising such a model). It seems again that a separate notion of truth *simpliciter* is used to justify a demand for theories devised to characterise (self-sufficiently!) such a notion. The deflationist who declares other notions of truth as useless or meaningless should have no truck with such argumentation.

There is still the last (and rather desperate) move to consider. It consists in declaring from the start that the notion of the intended model is incomprehensible and that all models are on a par (alternatively, declaring all of them to be intended). In some contexts, such a move is indeed a natural one. For example, our first order logic is valid in every domain, with no domain being privileged over any other. In effect, it is natural to claim that each interpretation of first order logic is as intended as an arbitrary other one. Similarly, consider a theory of groups with the usual axioms of associativity, identity and inverses. It can reasonably be claimed that no interpretation of these axioms should be considered devious, i.e. all groups are on an equal footing, each of them is as intended as any other. The idea now would be that arithmetic should be treated in a similar manner. There is no 'intended' model of arithmetic, just as there is no 'intended group', which would determine the truth value of sentences independent from the axioms of group theory. Hence, the model theoretic conservativity of truth theory becomes a natural demand.

However, this approach is again very problematic and adopting it would commit the deflationist to a quite far reaching and dubious philosophical standpoint. Consider a sentence Con_{PA}, expressing (under natural reading) the consistency of Peano Arithmetic. It is known that this sentence is not provable in PA, unless PA is inconsistent. In effect, Peano Arithmetic has

---

the concept [of natural number] can be secured, if we help ourselves to the full (standard) power set of some given infinite set […] But this is highly unsatisfactory, for it bases the concept of natural numbers on the much more problematic shaky concept of the full power set. It is […] like establishing the credibility of a person through the evidence of a much less credible character witness." (Gaifman 2003, 15-16)

models which make $\neg Con_{PA}$ true. We believe, however, that Peano Arithmetic is in fact consistent. But, since this is what Con_{PA} expresses, we also believe that models which make $\neg Con_{PA}$ true are *wrong* – they do not represent correctly the arithmetical (or proof theoretical) facts. In this sense, they are not on a par with models satisfying Con_{PA}. It is at this point where the analogy with group theory breaks down. For a typical example, consider the condition stating that the operation in a group is commutative. It is known that such a condition is independent of the axioms of group theory – there are commutative (Abelian) groups, there are also non-commutative ones. However, neither of these types are "real world groups"; it makes no sense to say that the operation characterised by group theoretical axioms is *in fact* commutative. There are simply two types of groups, neither of which are "wrong" about some algebraic facts. The case of PA is quite different in this respect.

So far I have discussed the semantic conservativity demand within the arithmetical framework. I treated the arithmetic as a model case for the deflationary claims to be tested and I assumed that the deflationary position concerns truth simpliciter, and not something else (e.g. it's not an attempt to deflate parts of set theory). What happens, one might ask, if these assumptions are dropped?

Admittedly, one could, in addition, be a deflationist about the notion of truth under an interpretation and claim that model theory unduly 'inflates' this notion. Moreover, a deflationist could also say that he is not interested specifically in arithmetic: what he is after is a general account providing *both* a general notion of truth simpliciter *and* a general (deflated) notion of truth under an interpretation, with the last one permitting us to make sense of model theory.

However, in my opinion, for such a deflationist the semantic conservativity condition would be even more problematic.

Firstly, the condition is formulated - at least initially - in model theoretic terms and it involves quantification over *all* interpretations. The last conjunct is important: typical piecemeal strategies of deflating the notion of truth (e.g. characterising "truth under the interpretation *I*" by means of appropriate Tarskian biconditionals) won't work here, unless some way is found to simulate the quantification over all interpretations. This makes the position even more demanding: there are additional problems to overcome!

Secondly, even if this can be done, it would still remain unclear why admissibility of all interpretations should matter. [28] It is exactly at this point where the really troublesome questions appear (closely related to the arguments presented earlier in this section). Will the general notion of truth permit us to make sense of the notion of the *intended interpretation* of our overall theory? If yes, what's the point of insisting on the admissibility of all interpretations, including those which are not intended? And if not – if the notion of the intended interpretation would still not be captured by our truth axioms – then any appeal to such a notion in an argument for semantic conservativity would remain exactly as illegitimate as before.

I have found no good arguments for semantic conservativity demand. It seems to me in fact that the demand (or at least the motivation behind it) is at odds with some deflationary tenets. However, this is not to say that investigating the properties of semantically conservative truth theories is a pointless endeavour. The question of which reasoning involving the notion truth

[28] Interpretations of *what*, one could ask? In the literature the conservativeness demand is often presented as a requirement of being conservative over theory of syntax, which plays the role of the base theory. Here, however, we are discussing a more sweeping picture, with the base theory being identified with all of our knowledge expressed in the language without the general truth predicate.

can be carried out independently of the choice of the model of our theory is interesting for its own sake. More information in this direction can be found in (Fischer, 2009).

6 Syntactic conservativity – what sort of argument?

The aim of this section is to discuss arguments in favour of the syntactic conservativity constraint for deflationary truth theories. As far as I can see, there is really just one serious candidate for the role of such an argument. Roughly, it consists in deriving the syntactic conservativity constraint from instrumentalist claim that the concept of truth is just a tool which in principle can be disposed of in explanations or justifications of non-semantic facts (see Section 2 for some textual basis).[29] Let me stress at the start that these last two concepts should not be conflated and the analysis of arguments for conservativity must be sensitive to differences between them. The proof of a given theorem may play two basic roles: justificatory or explanatory (sometimes both at the same time). In the first role, the proof convinces us of the truth of the theorem. This is particularly crucial if the result is a new, previously unknown discovery. It can also be important if the theorem has been reproved by more modest (and more believable) means than before. On the other hand, sometimes mathematicians look for alternative proofs of known results for different (i.e. not justificatory) reasons. In the words of Jamie Tappenden:

> A proof or proof sketch can give cogent grounds for believing a claim, but it might fail nonetheless to provide the sort of illumination we can hope for in mathematical investigation. It is not unusual, nor is it unreasonable, to be dissatisfied with a proof that doesn't convey understanding and to seek another argument that does. Sometimes one proof may be counted superior to a second even though both proofs are carried out within the same theoretical context (same definitions, primitive concepts, formal or informal axiomatic formulations, etc. In other cases […] the advantages of one argument over another appear to derive partly from the definitions and/or axioms in terms of which they are framed. (Tappenden 2005, p. 152)

The approach we are going to consider sends us back to (Shapiro, 1998). Let's recall the key question, posed in Shapiro's paper: "How thin can the notion of arithmetic truth be, if by invoking it we can learn more about the natural numbers?" The present observation is that "learning more" can mean two things: the notion of truth (that is, adding truth axioms) may give us an ability to *explain* previously unexplained phenomena or it may endow us with the possibility to *justify* new arithmetical theorems.

The argument for conservativity, which takes non-explanatory role of truth as given, could take the following form:

1. Truth is never explanatory.
2. If a theory of truth proves new non-semantic facts, then these new facts are explained by truth-theoretic considerations.

[29] This argument was proposed by Shapiro (1998) and Ketland (1999). In particular, Ketland explicitly linked the conservativeness/deflationism issue to a certain instrumentalist program; namely, to Field's (1980) attempt to show the conservativeness of mathematical axioms over any nominalistic theory of concrete objects. Such an attempt, if successful, would justify the claim that mathematics is a "mere instrument" and – in Ketland's words – it would serve "to 'deflate' the platonist notion that there is a realm of abstract mathematicalia". (See Ketland 1999, p. 71).

3. Therefore, a theory of truth does not prove new non-semantic facts i.e. it is syntactically conservative over its base.

The main stumbling block in regard to assessing this argument is that the concept of explanation in mathematics is at present neither well understood nor sufficiently studied, with the research still in the initial phase.[30] In such a situation, mathematical examples should be taken with a grain of salt. With this word of caution, there are however a couple of issues with such a line of thinking. For an illustration of possible problems, consider the following reasoning (discussed in a different context by Halbach (2011)[31]), which is carried out in a theory of truth consisting of Tarski biconditionals for arithmetical sentences with first order logic as a base theory:

$T(\ulcorner 0=0 \urcorner)$ iff $0 = 0$

$T(\ulcorner 0\neq0 \urcorner)$ iff $0 \neq 0$

Therefore $\ulcorner 0=0 \urcorner \neq \ulcorner 0\neq0 \urcorner$.

In effect, the theory of truth proves the existence of two distinct objects, clearly going beyond the base theory (in this case, beyond logic). The truth axioms permitted us to prove a new non-semantic fact, but did they permit us to *explain* this fact? At least on some accounts of explanation in mathematics, they did not. For example, Mark Steiner offers the following criteria for the proof to count as explanatory:

> an explanatory proof depends on a characterizing property of something mentioned in the theorem: if we 'deform' the proof, substituting the characterizing property of a related entity, we get a related theorem. A characterizing property picks out one from a family ('family' in the essay undefined); an object might be characterized variously if it belongs to distinct families. 'Deformation' is similarly undefined - it implies not just mechanical substitution, but reworking the proof, holding constant the proof-idea. (Steiner 1978, p. 147)

Thus, the criteria for explanatory proofs offered by Steiner, comprise: first, the dependence on a 'characterising property' of an object or a structure mentioned in the theorem; second, the possibility of generalising the result by the procedure of varying this property. As I take it, these conditions are simply not met by the proof just given. A minor reason is that the theorem is existential, which is a case not covered by Steiner. A more important reason is that the only 'characterising property' in this case is that of two objects being different (nothing else is mentioned in the theorem) and I can see no plausible candidates for the role of 'related theorems' to be obtained by the 'deformation' of the proof. I conclude, in effect, that premise 2 is not true by default – at least it is far from obvious that all proofs are explanatory. The connection between instrumentalism and conservativity just cannot be *that* direct.

Ketland (1999) proposed the slogan: "non-substantiality ≡ conservativeness" (p. 79). In the present context this is not satisfactory for another additional reason: even conservative truth theories may permit us to build explanatory truth-theoretic proofs of theorems in the base language. Conservativity means only that another truth-free proof will be available. What it does not rule out is that a proof in the extended language will be more informative, more general, or more explanatory. Consider the following simple proof:

[30] For an overview, see http://plato.stanford.edu/entries/mathematics-explanation/.

[31] See p. 55 and also p. 314. Halbach discusses this argument in order to criticise the demand of conservativity (of the theory of truth) over logic. My aim here is different, the focus being on the notion of explanation.

(P) Fix an arbitrary arithmetical sentence φ. We reason in CT^-, arriving at the weak law of identity for φ, i.e. the formula "$\varphi \rightarrow \varphi$", at the last step of our proof. The reasoning proceeds via compositional truth axioms of CT^-: since for every ψ, $T(\psi) \rightarrow T(\psi)$, compositional principles permit us to obtain a general statement "for every ψ, $T(\psi \rightarrow \psi)$", from which $T(\varphi \rightarrow \varphi)$ trivially follows. Applying disquotation (valid in CT^-) we reach finally the conclusion: $\varphi \rightarrow \varphi$.

Obviously, the conclusion of (P) is trivial: the detour via truth in CT^- is not necessary to obtain the weak law of identity for φ (CT^- is in fact a conservative extension of PA). However, this still leaves intact the question about the explanatory value of this proof. Let us take again Steiner's criteria as our starting point. A natural candidate for the role of the 'characterising property' on which the proof relies is a propositional structure of the theorem. Admittedly, this propositional structure is not *mentioned* in the theorem, as Steiner wants to have it, but I take it as a moot point. The proof (P) consists really in distributing the truth predicate over an arbitrary formula with an indicated propositional structure, observing the validity of the result and concluding (by compositionality) that the whole formula will always be true. We obtain related results by 'deforming' the proof – by substituting "the characterizing property of a related entity", i.e. by choosing a different propositional structure, e.g. "$\varphi \vee \neg\varphi$". After introducing such a deformation, we are able to "rework the proof, holding constant the proof idea". Again, just distribute the truth predicate, observe the validity of the result, and apply compositionality to justify the truth of the whole formula. In effect, I gather that Steiner's criteria for explanatory proof are satisfied in this case.

Investigation of the notion of mathematical explanation is an emerging area of research, where very little consensus has been achieved so far. Since the notion of explanation in mathematical contexts remains obscure, the example given above can be contested. One could introduce different – or perhaps additional – demands for the proof to count as explanatory.[32] As a side comment, let us note one curious trait of (P). Imagine that (P) is given as an explanation of someone's acceptance of the weak law of identity for φ. It is easy to observe that the same law – admittedly, for a formula different than φ and containing the truth predicate – has been in fact *used* in the proof (P) (what we have there as a step in a proof is a generalization "for every ψ, $T(\psi) \rightarrow T(\psi)$").

Here, we appeal in effect to (a form of) the weak law of identity in order to explain our acceptance of (a different form of) the weak law of identity. Is it acceptable in an explanatory proof? Observe that a negative answer to this question would have far reaching consequences: it would give the deflationist nothing short of full access to non-conservative theories of truth (as I take it, it would mean simply saying farewell to the conservativeness condition). The standard way to prove the non-conservativity of a truth theory *Th* proceeds via proving in *Th* the so-called "global reflection principle" (GR) for the base theory *B*. We do this by proving "All theorems of *B* are true", and then by deducing the consistency of *B*, which by Gödel's second theorem, ensures non-conservativity. If the theory *B* in question is schematically axiomatised, the proof of GR typically uses an instance of the axiom schema of *B* in the extended language (with the truth predicate). For example, if *B* is Peano arithmetic, axiomatised by means of the induction schema, a part of the proof of GR consists in showing that all the (arithmetical) axioms of induction are true, which is typically done by *using* induction in the extended language containing the truth predicate. The problem becomes perhaps most visible after

[32] Or one could reject the notion of explanatory proof in mathematics altogether; cf. (Resnik and Kushner, 1987).

presenting our explanation as a series of answers to the "why" questions: (1) why Con_{PA}? Because of (GR); (2) why (GR)? Because all the axioms of PA are true and our rules of inference preserve truth; (3) why in particular are the axioms of induction true? Because we can prove it by (a different version of) induction. Although unlike the case of (P), the final statement (that is, Con_{PA}) is not derived by means of an instance of the same statement, it is still the case that in part (3) our explanation contains a circular argument.[33] In effect, if someone wanted to contest the explanatory value of (P) for reasons of its 'circularity', then he would have to question the familiar consistency proofs for exactly the same reason.[34]

Anyway, the moral is that conservativity per se does not guarantee the non-existence of explanatory truth theoretic proofs; neither non-conservativity implies the existence of such proofs. I therefore conclude that a different argument is needed to validate the conservativeness requirement.[35]

So far we have concentrated on the explanatory role of truth, with negative results (and with the main trouble being perhaps that – as it seems – the prospects for building a good account of mathematical explanation look dim at the moment). The second possible approach takes *justification*, not explanation, as the basic concept. Accordingly, the argument for conservativity could take the following form:

1. Truth is never justificatory.
2. If a theory of truth proves new non-semantic facts, then these new facts are justified by truth-theoretic considerations.
3. Therefore, a theory of truth does not prove new non-semantic facts, i.e. it is syntactically conservative over its base.

However, in this version the argument still remains vulnerable and weak, even if we take for granted the attribution of premise 1 to the deflationists. The point is that premise 2 faces a serious problem; being, broadly, the issue of the justificatory value of truth theoretic arguments. After all, it might well be the case that, from a justificatory point of view, proofs of new non-semantic facts in a non-conservative theory of truth are quite worthless; that is, these facts are not accepted by us *because of* these proofs, nor our degree of belief in these facts increases once we are presented with their truth theoretic proofs. In such a situation, premise 2 would become false, with the whole argument breaking down. Now, how realistic is this scenario?

A typical example of the "new fact" proved by a non-conservative theory of truth is the consistency of the base theory. Thus, a non-conservative theory CT, with full induction, proves the consistency of Peano arithmetic. How compelling is such a proof? The last question – let me stress – is not about formal correctness of the proof of Con_{PA} in CT (the proof *is* formally correct!). It rather concerns its justificatory power: to what degree does the proof justify our belief in consistency of Peano arithmetic? To put the matter in different terms, imagine that someone has serious doubts about the consistency of PA. After seeing and understanding the proof in CT, will he lose these doubts? Or, more importantly, *should* he lose them?

[33] Stages (1) and (2) are admittedly non-circular, but it's a weak consolation. If circularity is unacceptable in explanations, I can see no reason why it should matter in which part of an explanation it occurs.

[34] Nevertheless, some philosophers explicitly accepted such consistency proofs as explanatory, e.g. Shapiro wrote: "On an intuitive level, however, I submit that we do have a good explanation of G [the Gödel sentence], and that this explanation invokes truth in the explanation. The burden is on the deflationist to show what is wrong with this picture." (Shapiro 1998, p. 507)

[35] It's also worth stressing that perhaps the most natural context for truth-theoretic explanations is when the explanation concerns a semantic fact, not an arithmetical one. However, if the deflationist were to reject truth as explanatory notion also in such contexts, conservativity wouldn't cover it.

Let us start with being clear about the means which are used in the proof. When proving $\mathrm{Con}_{\mathrm{PA}}$ in a truth theory like CT, what we in fact employ is some theory of syntax (here $\mathrm{I}\Sigma_1$ would be enough), we also use compositional truth axioms combined with extended induction (as a matter of fact it is rather easy to verify that Π_1 induction for the extended language is quite satisfactory for this purpose). Initially this approach could look appealing, modest, and trustworthy; after all, it is just $\mathrm{I}\Sigma_1$ and partially inductive notion of compositional truth! Let us look, however, at some of the details, starting with the following simple observation:

Observation. *Let Th = $I\Sigma_1$ + compositional truth axioms + induction for Δ_0 formulas of the extended language (with the truth predicate). Then Th proves: "all the axioms of PA are true", with PA taken as the theory axiomatised by means of a parameter free induction schema.*

Proof. All axioms of PA, except the inductive ones, obviously belong to *Th*, so by disquotation (valid in *Th*) they are true. For the truth of inductive axioms, working in *Th* fix an arithmetical formula $\varphi(x)$ with one free variable. It is enough to obtain:

$$(*) \quad \mathrm{T}(\varphi(0)) \wedge \forall x[\mathrm{T}(\varphi(x)) \rightarrow \mathrm{T}(\varphi(x+1))] \rightarrow \forall x \mathrm{T}(\varphi(x)).$$

Then the truth of inductive axiom for $\varphi(x)$ will follow by compositionality. Let us assume the antecedent of (*). For an indirect proof, assume also $\exists x \neg \mathrm{T}(\varphi(x))$ and choose (using Δ_0 induction) the smallest x with this property. By the antecedent of (*) such a smallest x can be neither zero nor a successor number, which generates a contradiction. □

It follows that even *Th*, with a seemingly weak base theory, is at least as strong as full PA. Further extension of *Th* with Π_1 induction for formulas with the truth predicate produces a non-conservative theory.[36]

Where does it leave us in terms of our justificatory purpose? Let us go back to our imagined opponent, to the person who (at least initially) has doubts about the consistency of PA. How should he react to the truth-theoretic consistency proof? The proof under consideration clearly requires some theory of syntax. As we have stressed, this theory of syntax does not have to be full PA (the consistency of which after all is *doubted* by the opponent). The real trouble comes with the truth axioms combined with extended induction. As we saw in the proof of Observation, it is the extended Δ_0 induction that licences a move from $\exists x \neg \mathrm{T}(\varphi(x))$ to the choice of the smallest x with this property. It is also at this point where the opponent has every right to feel cheated. Accepting the least number principle in such a form is – he could say – nothing short of accepting full arithmetical induction as credible. It does not matter that the extended induction used in the proof is "just" Δ_0: the principle works for an *arbitrary* arithmetical formula $\varphi(x)$, which is turned into a Δ_0 formula by a mere quirk of syntax (i.e. by appending "T"). In effect, for someone who doubts the consistency of PA, a proof which assumes a truth theoretic version of the least number principle (for arbitrary arithmetical formulas) does not add much in terms of justification.[37] It is perfectly possible for a theory of truth to be non-conservative *and* at the same time for the truth predicate to have very little justificatory power. In the end, this version of the argument for conservativity breaks down as well.[38]

[36] Whether *Th* itself is that strong, i.e. whether it is non-conservative over PA, remains an open problem.

[37] Cf. the following remark of Pohlers (2009) about Gentzen's consistency proof: "At this point our opponent will argue that doing so we exhaust full first order number theory and even a bit more. But (s)he doubts full number theory. Therefore (s)he cannot accept the proof. We hardly can advance a mathematical argument against that." (p. 129).

[38] Although we discussed the explanatory and justificatory role of truth separately, much the same can be said about the disjunction of these two properties. In other words, identifying "truth is substantial"

7 Conclusion

Neither semantic nor syntactic conservativeness fares well as an explication of traditional deflationary claims. A commitment to a conservative truth theory is not supported by the views typically attributed to the deflationists. In addition, some arguments for conservativity are at odds with deflationary tenets. The best which can be said about conservativeness is that, in some respects, it is a *convenient* property. If each arithmetical theorem has not only truth-theoretic, but also arithmetical proof, the adherent of a given truth theory has at least a *candidate* for the role of a purely arithmetical explanation/justification of an arithmetical claim. However, we can say no more than that. There is no place for conservativeness as a commitment of traditional deflationary standpoint.

REFERENCES

AYER, A. J. 1935, "The criterion of truth", Analysis 3, pp. 28-32.

CANTINI, A. 1989, "Notes on formal theories of truth", Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 35(2), pp. 97-130.

CIEŚLIŃSKI, C. 2015, "Typed and untyped disquotational truth", in: T. Achourioti, H. Galinon, K. Fujimoto, and J. Martínez-Fernández, eds., *Unifying the philosophy of truth*, Springer, to appear.

CIEŚLIŃSKI, C. 2011, "T-equivalences for positive sentences", Review of Symbolic Logic 4(2), pp. 319-325.

CIEŚLIŃSKI, C. 2010, "Truth, conservativeness, and provability", Mind 119(474), pp. 409-422.

ENAYAT, A. and VISSER, A. 2015, "New constructions of satisfaction classes", in: T. Achourioti, H. Galinon, K. Fujimoto, and J. Martínez-Fernández, eds., *Unifying the philosophy of truth*, Springer, to appear.

FIELD, H. 1999, "Deflating the conservativeness argument", Journal of Philosophy 96(10), pp. 533-540.

FIELD, H. 1994, "Deflationist views of meaning and content", Mind 103(411), pp. 249-285.

FIELD, H. 1980, *Science without numbers*. Princeton: Princeton University Press.

FISCHER, M. 2009, "Minimal truth and interpretability", Review of Symbolic Logic 2(4), pp. 799-815.

GAIFMAN, H. 2003, "Non-standard models in a broader perspective", in: A. Enayat and R. Kossak, eds., *Nonstandard models of arithmetic and set theory*, The Contemporary Mathematics Series, American Mathematical Society, pp. 1-22.

GROVER, D. 1992, *A Prosentential Theory of Truth*, Princeton, NJ: Princeton University Press.

HALBACH, V. 2011, *Axiomatic theories of truth*, Cambridge: Cambridge University Press.

HALBACH, V. 2009, "Reducing compositional to disquotational truth", Review of Symbolic Logic 2(4), pp. 786–798.

HALBACH, V. and Horsten, L. 2005, "Computational structuralism", Philosophia Mathematica 13(2), pp. 174-186.

HORSTEN, L. 2011, *The Tarskian turn. Deflationism and axiomatic truth*, Cambridge, MA: MIT Press.

HORSTEN, L. 1995, "The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth", in: P. Cortois, ed., *The Many Problems of Realism*, vol. 3 of Studies in the General Philosophy of Science, Tilburg: Tilburg University Press, pp. 173-87.

HORWICH, P. 1999, *Truth*, second edition, Oxford: Clarendon Press.

KAYE, R. 1991, *Models of Peano Arithmetic*, Oxford: Clarendon Press.

KETLAND, J. 1999, "Deflationism and Tarski's paradise", Mind 108(429), pp. 69-94.

KETLAND, J. 2010, "Truth, conservativeness, and provability: reply to Cieslinski", Mind 119(474), pp. 423-436.

with "truth plays an explanatory *or* a justificatory role" leads to the same troubles with conservativity: on the one hand, proofs of arithmetical facts in non-conservative theories could be neither explanatory nor justificatory, on the other - proofs of arithmetical facts in conservative theories could be explanatory *or* justificatory.

KOTLARSKI, H., KRAJEWSKI, S., and LACHLAN, A. H. 1981, "Construction of satisfaction classes for nonstandard models", Canadian Mathematical Bulletin 24(3), pp. 283-293.
KRIPKE, S. 1975, "Outline of a theory of truth", Journal of Philosophy 72(19), pp. 690–716.
LEITGEB, H. 2007, "What theories of truth should be like (but cannot be)", Philosophy Compass 2, pp. 276-290.
MCGEE, V. 2006, "In praise of the free lunch: why disquotationalists should embrace compositional semantics", in: T. Bolander, V. F. Hendricks, and S. A. Pedersen eds., *Self-Reference*, Stanford: CSLI Publications, pp. 95-120.
POHLERS, W. 2009, *Proof theory*, Berlin, Heidelberg: Springer.
RESNIK, M., and KUSHNER, D. 1987. "Explanation, independence and realism in mathematics", British Journal for the Philosophy of Science 38(2), pp. 141-158.
SHAPIRO, S. 1998, "Proof and truth: through thick and thin", Journal of Philosophy 95(10), pp. 493-521.
STRAWSON, P. 1949, "Truth", Analysis 9, pp. 83-97.
STROLLO, A. 2013, "Deflationism and the invisible power of truth", Dialectica 67(4), pp. 521-543.
STEINER, M. 1978, "Mathematical explanation", Philosophical Studies 34(2), pp. 135-151.
TAPPENDEN, J. 2005, "Proof style and understanding in mathematics I: visualization, unification and axiom choice", in: P. Manoscu, K. Jørgensen and S. Pedersen, eds., *Visualization, Explanation and Reasoning Styles in Mathematics*, Dordrecht: Springer, pp. 147-214.
TENNANT, N. 2005, "Deflationism and the Gödel phenomena: reply to Ketland", Mind 114(453), pp. 89-96.
WILLIAMS, M. 1999, "Meaning and deflationary truth", Journal of Philosophy 96(11), pp. 545-564.