Deflationism, conservativeness and maximality

Cezary Cieśliński
Institute of Philosophy, the University of Warsaw
Krakowskie Przedmieście 3, 00-047 Warsaw, Poland
e-mail:c.cieslinski@poczta.uw.edu.pl

1 Preliminaries

The aim of this paper is to put together two logical properties, which could be seen as attractive traits of a deflationary theory of truth: conservativeness and maximality. The issue of conservativeness came to the foreground in recent discussions about deflationism.[1] In their account of truth, deflationists quite often invoke the so-called T-schema:

(T) $\ulcorner\varphi\urcorner$ is true if and only if φ

This schema, restricted in some way in order to avoid the paradoxes, is said to capture the whole content of our intuitive notion of truth. In effect truth is called sometimes a "purely logical" device - a predicate devoid of content, which doesn't express any substantial relation between our language and the world.

According to one proposal, conservativeness is a good explication of this "contentlessness" of truth.[2] Let us recall here the basic definition.

Definition 1 *Let S_1 and S_2 be theories formulated appropriately in languages L_1 and L_2, with $L_1 \subseteq L_2$ (i.e. L_2 may be a richer language, e.g. with some additional predicates). Then we say:*

S_2 is conservative over S_1 iff for every sentence φ of L_1 , if $S_2 \vdash \varphi$, then $S_1 \vdash \varphi$.

[1] See the papers of Field, Ketland, Shapiro and Tennant, listed below in the bibliography.

[2] This proposal is due to Shapiro [7] and Ketland [4]. Admittedly, both authors are not deflationists - they belong rather to the critics. Hartry Field is an example of a deflationist who explicitly accepted conservativeness as a requirement imposed on a theory of truth; see [1].

In short: S_2 doesn't prove any new sentences of the language of S_1 .

Consider now a deflationist who wants to extend some (non-semantic) base theory S by new axioms, characterizing the truth predicate (in some versions, these additional axioms are simply the appropriate T-sentences). It has been urged that such an extension should be conservative over S - otherwise how could a deflationist claim that the notion of truth is contentless or "metaphysically thin"?[3] Quite on the contrary: the notion in question would be then powerful enough to give us a proof of some sentence of the base language, which wasn't provable before. It would seem in effect that truth has a lot of non-semantic content.

The second property, which we are going to discuss, is maximality. Imagine again that we want to extend our base theory S by instantiations of the T-schema. Of course some instantiations will lead to paradoxes and they would be unwelcome as our theorems. But what about the other ones? At first sight an attractive option could be: try to include as many of them as possible. It means in effect that we would opt for nothing less than a maximal consistent extension of our base theory by the instantiations of the T-schema.

This possibility has been investigated by Vann McGee (see [6]). His initial result was quite promising: after taking some arithmetic (say PA) as our base theory, we see that it is indeed feasible to extend it in such a way. Arithmetic has a maximal consistent extension. However in the end McGee didn't have good news for a deflationist. Maximality by itself turns out to be a rather poor guide to the theory of truth, and this for two reasons:

1. There are just too many possible ways of extending our arithmetic to a maximal consistent set. One maximal theory will prove some sentence φ, other - the negation of φ. We need some additional principle to differentiate between such theories. Sheer maximality is plainly not good enough.

2. Maximal consistent extensions of the required sort are not axiomatizable. They are in fact complete, and by Gödel's first theorem, completeness cannot be squared with axiomatizability. On the other hand, axiomatizability is a very useful property and the lack of it certainly diminishes the attractiveness of a given theory of truth.[4]

[3] This last phrase was used by Shapiro, see [7].

[4] The importance of axiomatizability considered as a requirement for a deflationary thory of truth was stressed by Gauker; see [2].

The moral is that in any case we need some additional guiding principles, permitting us to choose the T-sentences, which we will incorporate into our theory. A deflationist must be very careful here: if these additional principles turn out to be "substantial" or "contentful", they may compromise his deflationary standpoint. But at this moment one option presents itself: maybe conservativeness will help? After all, even the critics seem to admit that this property is desirable from a deflationary point of view. It is the aim of the present paper to investigate this option is some detail.

The idea is to start with some base theory (in what follows this role will be played by first order Peano arithmetic) and to extend it conservatively by instantiations of the T-schema in such a way as to attain maximality - in effect we will obtain a conservative extension which can't be enlarged any further without losing its conservative character. The intuition would be that such a theory says all there is to say about truth, but without implying too much - i.e. without proving new arithmetical sentences. All "contentless" aspects of truth would be characterized by such theory. As for the other aspects, one could claim that it's not the task of the truth theorist to describe them. They are "metaphysically thick"[5] and therefore we are fully entitled to leave them alone.

One final comment: I do not want to claim that a deflationist is *committed* to some sort of maximality. I'm saying only that maximality would be for him a convenient property. If our theory of truth is not maximal, the obvious question could be asked: how can we understand the application of the truth predicate to a sentence φ, if (as it may happen) we are not able to prove the appropriate T-sentence for φ. Again, I'm not claiming that questions of this sort cannot be answered. My point is simply that with maximality at hand, they just don't arise.

2 Maximal conservative truth theories

As we said, we take PA as our base theory. Let Ar be the language of first order arithmetic and let Ar^+ be the extension of Ar by a new one-place predicate "$Tr(.)$". We want to add some instantiations of the T-schema to our theory; however at this moment it will be convenient to think about it just in terms of arbitrary sentences of Ar^+, regardless of whether they are such instantiations or not. The following fact (due to McGee [6]) explains the rationale behind this change of perspective.

Fact 1 *Let T be any theory containing PA. Then for every sentence $\alpha \in Ar^+$*

[5] Shapiro's phrase again.

there is a sentence $\beta \in Ar^+$ such that $T \vdash \alpha \equiv (Tr(\ulcorner\beta\urcorner) \equiv \beta)$. In other words: every sentence of Ar^+ is provably equivalent to some T-sentence.

Proof. Pick $\alpha \in Ar^+$. By diagonal lemma, fix β such that $T \vdash \beta \equiv (Tr(\ulcorner\beta\urcorner) \equiv \alpha)$. Then by sentential logic $T \vdash \alpha \equiv (Tr(\ulcorner\beta\urcorner) \equiv \beta)$. □

With this fact at hand, we start with the following theorem about the existence of maximal conservative extensions.

Theorem 1 *Let T be a conservative extension of PA in the language Ar^+. Then there is a theory T_1 in the language Ar^+ that comprises T and is a maximal conservative extension of PA.*

Proof. It's a straightforward application of Zorn's lemma. Define the set A as follows:

$$A = \{B : T \subseteq B \wedge B \text{ is conservative over PA}\}$$

Then A is a family of sets partially ordered by inclusion and such that each chain in A has an upper bound in A (if C is a chain in A, then the sum of C comprises T and is conservative over PA, so it belongs to A). By Zorn's lemma, A has maximal elements and any such element will be a maximal conservative extension of PA. □

We have assumed so far that "$Tr(.)$" is just some one-place predicate. However, if it is really to play a role of the truth predicate, it seems reasonable to demand that our theory prove at least all the arithmetical instantiations of the T-schema, i.e. the instantiations by sentences not involving the predicate "*Tr*". Formally, we will expect from our theory T that:

$$\text{For every } \alpha \in Ar, T \vdash Tr(\ulcorner\alpha\urcorner) \equiv \alpha$$

The good news for a deflationist is that the extension of PA by all such instantiations is conservative (see Ketland [4]). We will claim however that the troublesome traits (1) and (2) characterize not only maximal consistent, but also maximal conservative extensions. The next two theorems will show just that.

Theorem 2 *Let $T = PA \cup \{\ulcorner Tr(\ulcorner\varphi\urcorner) \equiv \varphi\urcorner) : \varphi \in Ar\}$. Then T has continuum many maximal conservative extensions.*

Sketch of the proof. Obviously T has at most continuum maximal conservative extensions, so it is enough to show that the number of the relevant extensions is at least that large. We denote by Q_n the set of arithmetical sentences (i.e. sentences of the language Ar) with exactly n quantifiers. Let ψ_n be the following formula of the language Ar^+:

$$\psi_n := \forall \alpha \in Q_n[Tr(\ulcorner \neg\alpha \urcorner) \equiv \neg Tr(\ulcorner \alpha \urcorner)]$$

Consider now a binary tree with T at the root. Starting from T (level 0), on each level $n+1$ we can choose the path to the left and add $\neg\psi_n$ to our theory, or we may go to the right and add ψ_n. We will claim that each branch in our tree corresponds to a conservative extension of PA, which by Theorem 1 can be then extended to a maximal conservative set. Since the number of branches equals continuum, this will end the proof of our theorem. We show that on each level, whichever path we choose, a conservative extension of PA will be obtained. So let S be a theory obtained at a stage $n+1$. Then $S = W + \psi_n$ or $S = W + \neg\psi_n$, where W is a theory obtained at level n. By our inductive assumption, W is a conservative extension of PA. We show now that for each model of PA, there is an elementarily equivalent model satisfying S (which is tantamount to our desired conservativeness result). Let $K \models PA$. Take a nonstandard M such that $M \equiv K$ (i.e. both models satisfy the same arithmetical sentences) and $M \models W$. There is such a model M since W is conservative over PA. Now we consider two cases.

Case 1 : $S = W + \neg\psi_n$. In this case pick a model M_1, which is just like M with only the interpretation of "Tr" changed. For any model X, we denote by "$Tr(X)$" an interpretation of "Tr" in X. Then we define:

$$Tr(M_1) = Tr(M) - \{\alpha, \neg\alpha\}$$

where α is an arbitrary nonstandard sentence from M such that $M \models \alpha \in Q_n$. Then $M_1 \equiv M$ and $M_1 \models S$. The first conjunct is obtained because arithmetically M_1 is really the same as M - only the interpretation of "Tr" was changed. As for the second, it is obviously a model of $\neg\psi_n$; and it satisfies also W because what happens at any lower level k depends only on the behaviour of Q_{k-1} with respect to the truth predicate, and we leaved this unchanged as well.

Case 2 : $S = W + \psi_n$. Here we also define an appropriate model M_1 differing from M only in the interpretation of "Tr" . It can be easily done by using an arithmetical formula "$Tr_{Q_n}(x)$", being a partial truth predicate for formulas with n quantifiers.[6] With such a formula at hand, we define:

[6] For partial truth predicates and their properties, see Kaye [3], p. 119-129. The key point here is that it's possible to choose a formula "$Tr_{Q_n}(x)$" in such a way as to obtain: $PA \vdash \forall\alpha \in Q_n[Tr_{Q_n}(\ulcorner\neg\alpha\urcorner) \equiv \neg Tr_{Q_n}(\ulcorner\alpha\urcorner)$.

$$Tr(M_1) = (Tr(M) - Q_n(M)) \cup Tr_{Q_n}(M)$$

The expression "$Tr_{Q_n}(M)$" denotes here the set of all a belonging to M such that $M \models Tr_{Q_n}(a)$ and $Q_n(M)$ is the set of all a from M such that $M \models Q_n(a)$. Then again $M_1 \equiv M$ and $M_1 \models S$. Finally let us note that if the predicate "Tr" is inductive in M, it will be also in both cases inductive in M_1. The reason is that in both cases the set $Tr(M_1)$ is definable with parameters in M. So our result holds even if in T we permit the substitutions of formulas with the predicate "Tr" in the induction scheme. □

Now let's deal with the axiomatizability issue. We already know that maximal conservative extensions of PA do exist. Are any of them axiomatizable? When we consider consistent extensions, which are simply maximal, their completeness guarantees a negative reply to our question. But obviously no conservative extension of PA will be complete, so our considerations here must be altogether different. To clear the ground, let's put the trivial cases aside. If the conditions initially imposed on the predicate "Tr" are very weak, then indeed axiomatization may be possible. This will be the case of a theory T (axiomatizable and conservative over PA) such that for some arithmetical formula $\alpha(x)$, T can be extended to a theory T_1 (still conservative over PA) by adding a sentence "$\forall x[\alpha(x) \equiv Tr(x)]$". That is, T could be so weak as to admit (conservatively) the possibility that the set of objects satisfying "$Tr(.)$" is definable by some arithmetical formula. In this case T_1 - the relevant extension of T - would be axiomatizable, conservative and maximal, but for a rather silly reason: every sentence of Ar^+ would be then equivalent (provably in T_1) to some arithmetical sentence. So again in our present context we will consider only theories, which for every arithmetical sentence α prove the equivalence "$Tr(\ulcorner\alpha\urcorner) \equiv \alpha$". Then by Tarski's theorem, the objects satisfying "$Tr(.)$" can't be defined by any arithmetical formula and no consistent extension of our theory can prove anything of this sort. Do such theories have axiomatizable, maximal conservative extensions? They do not.

Theorem 3 *Let T be a conservative extension of PA in the language Ar^+ such that:*

1. *for every $\varphi \in Ar$, $T \vdash Tr(\ulcorner\varphi\urcorner) \equiv \varphi$*
2. *T is axiomatizable*

Then there is a sentence ψ of the language Ar^+ such that $T \nvdash \psi$ and $T+\psi$ is a conservative extension of PA. In other words, T is not a maximal conservative extension of PA.

Before presenting the proof of theorem 3, we fix our terminology and specify some assumptions. Let "$Prov_T(x,y)$" be an arithmetical formula, which represents in T the relation "x is a proof of y in T". Let "$Pr_T(y)$" be the formula "$\exists x Prov_T(x,y)$". We denote by N the standard model of arithmetic. For a theory T in the language Ar^+, by "$N \models T$" we mean: "all arithmetical theorems of T are true in N". As before, for two models A and B we write "$A \equiv B$" in case when A and B satisfy exactly the same arithmetical sentences. We assume that $N \models PA$, so obviously we will also have: $N \models T$, for any T being a conservative extension of PA. In the proof we will make use of Löb's theorem, which may be formulated as follows: for every axiomatizable, consistent theory T containing PA, we have: for every formula β of the language of T, $T \vdash Pr_T(\ulcorner \beta \urcorner) \Rightarrow \beta$ iff $T \vdash \beta$.

Proof of theorem 3. Using diagonalization, fix ψ such that:

$$T \vdash \psi \equiv \forall d[Prov_T(d, \ulcorner \psi \urcorner) \Rightarrow \exists \alpha \in Ar(\alpha < d \land Pr_T(\ulcorner \alpha \urcorner) \land \neg Tr(\ulcorner \alpha \urcorner))]$$

We claim that:

1. $T \nvdash \psi$

2. $T + \psi$ is a conservative extension of PA

Proof of (1). Assume that $T \vdash \psi$. Let d be a proof of ψ in T. Then $T \vdash Prov_T(d, \ulcorner \psi \urcorner)$. Let φ_0 ... φ_k be all arithmetical sentences with gödel numbers smaller than d. Then we have:

$$T \vdash (Pr_T(\ulcorner \varphi_0 \urcorner) \land \neg Tr(\varphi_0)) \lor ... \lor (Pr_T(\ulcorner \varphi_k \urcorner) \land \neg Tr(\varphi_k))$$

and therefore:

$$T \vdash (Pr_T(\ulcorner \varphi_0 \urcorner) \land \neg \varphi_0) \lor ... \lor (Pr_T(\ulcorner \varphi_k \urcorner) \land \neg \varphi_k)$$

But $N \models T$, so for some $s \leq k$, $T \vdash \varphi_s$ and $N \models \neg \varphi_s$. Therefore $N \nvDash T$, which is a contradiction ending the proof.
$Proof of$(2). We show that:

$$\forall K \models PA \; \exists S[S \models (T + \psi) \land S \equiv K]$$

It will mean in effect that $T + \psi$ is conservative over PA.

Fix $K \models PA$. Let $M \equiv K$ such that $M \models T$ (there is such a model M because T is conservative over PA). Let $Th^{Ar}(M) = \{\beta \in Ar : M \models \beta\}$. We will claim that $T + Th^{Ar}(M) + \psi$ is consistent, which will end the proof, giving us via completeness theorem the desired S satisfying $T + \psi$ and elementarily equivalent to K. Assuming the contrary, we have: $T + Th^{Ar}(M) \vdash \neg \psi$.

By compactness, only a finite fragment of $Th^{Ar}(M)$ is needed to establish this result, so let β be an arithmetical sentence belonging to $Th^{Ar}(M)$ such that $T + \beta \vdash \neg\psi$. We show that in this case, for every α belonging to Ar, $T + \beta \vdash Pr_T(\ulcorner\alpha\urcorner) \Rightarrow \alpha$.

Let $\alpha \in Ar$. Since α is a standard formula and $T \nvdash \psi$, we have: $T \vdash \forall d[Prov_T(d, \ulcorner\psi\urcorner) \Rightarrow \alpha < d]$. Now working in $T + \beta$, assume that $Pr_T(\ulcorner\alpha\urcorner)$. Let d be the smallest proof of ψ in T (the existence of such a proof follows from $\neg\psi$). We have: $\alpha \in Ar$ and $\alpha < d$. Then by $\neg\psi$, we obtain $Tr(\ulcorner\alpha\urcorner)$. Therefore α.

So we have shown that $T + \beta \vdash Pr_T(\ulcorner\alpha\urcorner) \Rightarrow \alpha$ for an arbitrary α belonging to Ar. However, β itself (and so obviously its negation) belongs to Ar, therefore via deduction theorem we obtain: $T \vdash \beta \Rightarrow (Pr_T(\ulcorner\neg\beta\urcorner) \Rightarrow \neg\beta)$. Then by sentential logic $T \vdash (Pr_T(\ulcorner\neg\beta\urcorner) \Rightarrow \neg\beta)$. In effect by Löb's theorem $T \vdash \neg\beta$, which means that $T + \beta$ is inconsistent. But M is a model of $T + \beta$, and in this way our proof is finished. □

Where does it leave the deflationist? It seems to us that there are two challenges he must face. First, since by theorem 3 his theory of truth can't be maximal (at least if it's axiomatizable), he owes us some explanation of how (if at all) we can understand our truth predicate in applications, which transcend the bounds of his theory. Theorem 3 guarantees that there are indeed such applications. Moreover, in a certain sense they concern 'truth only' - i.e. their addition to our theory will preserve conservativeness.

The second challenge concerns the choice of a deflationary theory. As we saw, conservativeness requirement leaves us with continuum many options. A deflationist must explain and motivate his decision in such a way as not to lose his deflationary credentials: as it seems, in his explanations he is not allowed to use the notion of truth stronger than that employed within his chosen theory. E.g. if for some conservative T, he chose $T + \psi$ as his theory of truth (where ψ is a sentence constructed in our proof of Theorem 3), then we could ask about his grounds for including ψ. Imagine that we hear the following answer: "I can accept ψ as an additional axiom, because if $T \nvdash \psi$, then ψ holds. And note that indeed T doesn't prove ψ, because otherwise the arithmetical theorems of T couldn't be true, as in fact they are". Now this sort of answer seems hardly accessible to a deflationist, unless he can explain how a deflationary theory of truth can licence the conclusion that all arithmetical theorems of T are true. And it is by no means clear that such an explanation can be given (see Shapiro [7] and Ketland [4]).

In the end, however, let me stress once again: it's not my intention to claim that these challenges cannot be met. My aim was only to provide some logical background for future discussions on these subjects.

References

[1] H. Field "Deflating the conservativeness argument", *Journal of Philosophy* 96 (1999), 533-540.

[2] C. Gauker "T-schema deflationism versus Gödel's first incompleteness theorem", *Analysis* 61 (2001), 129-135.

[3] R. Kaye *Models of Peano Arithmetic*, Oxford: Clarendon Press 1991.

[4] J. Ketland "Deflationism and Tarski's paradise", *Mind* 108 (1999), 69-94.

[5] J. Ketland "Deflationism and the Gödel phenomena: reply to Tennant", *Mind* 114 (2005), 75-88.

[6] V. McGee "Maximal consistent sets of instances of Tarski's schema (T)", *Journal of Philosophical Logic* 21 (1992), 235-41.

[7] S. Shapiro "Proof and truth - through thick and thin", *Journal of Philosophy* 95 (1998), 493-522.

[8] N. Tennant "Deflationism and the Gödel phenomena", *Mind* 111 (2002), 551-582.