Truth, Conservativeness, and Provability

CEZARY CIEŚLIŃSKI

Conservativeness has been proposed as an important requirement for deflationary truth theories. This in turn gave rise to the so called 'conservativeness argument' against deflationism: a theory of truth which is conservative over its base theory *S* cannot be adequate, because it cannot prove that all theorems of *S* are true. In this paper we show that the problems confronting the deflationist are in fact more basic: even the observation that logic is true is beyond his reach. This seems to conflict with the deflationary characterization of the role of the truth predicate in proving generalizations. However, in the final section we propose a way out for the deflationist—a solution that permits him to accept a strong theory, having important truth-theoretical generalizations as its theorems.

1. Introduction

Jeffrey Ketland (1999) and Stewart Shapiro (1998) have proposed conservativeness as an important requirement for deflationary truth theories. It is the deflationist's intuition that truth is in some sense 'innocent' or 'metaphysically thin'.[1] The truth predicate is just a 'logical device' permitting us to formulate useful generalizations (moreover, some of these generalizations will indeed acquire the status of theorems of our theory of truth), but it does not by itself add any new content to our non-semantic base theory. In Shapiro's and Ketland's opinion, conservativeness comes as a handy explication of these intuitions: the deflationist should adopt a theory of truth which is conservative over its base theory. It means in effect that all the sentences of the base language provable in our theory of truth will be provable already in the base theory itself. In Ketland's words:

> Suppose you have some non-semantical theory *S* […] in a language *L* and you extend it to a theory [of truth] S^+. […] Suppose you deduce in this 'semanticized' theory S^+ some non-semantical sentence φ […], perhaps using the concept of truth (in-*L*) in the deduction. Then the conservativeness theorem […] tells you that you can already deduce φ in *S*, without invoking the concept of truth. Hence we have an important sense in which the [deflationary] truth predicate is dispensable. Any non-semantical fact explained with [deflationary] truth can be explained without it. (Ketland 2000, p. 320)

With the above explication at hand, the critic's charge against deflationism is that it cannot explain various 'epistemic obligations', which we should accept once we adopt some base theory *S*. In particular, anyone who accepts a mathematical base theory *S* and has a notion of truth, should accept the so called global reflection principle:

> **(GR)** All theorems of *S* are true.

But if our base theory includes Peano arithmetic (*PA*), then no deflationary truth theory can prove (GR) on pain of losing its conservative character: using (GR) we can easily prove the consistency of *S*, and by Gödel's second incompleteness theorem *S* by itself doesn't prove that. What can the deflationist do? In Ketland's opinion there is only one strategy available to him: he should deny that (GR) should follow from his theory of truth and at the same time offer some non-truth-theoretic analysis of our epistemic obligations (cf. Ketland 2005).

[1] This last phrase was used by Shapiro in the quoted paper.

This challenge has been taken up by Neil Tennant (see Tennant 2002 and also Tennant 2005). Tennant's question is: why should the deflationist be saddled with a commitment to the soundness claim as expressed in the form of the global reflection principle? In his own words:

> We are being asked to believe that the […] claim:
>
> All *S*-theorems are true
>
> is the only—or, if not the only, then at least the most desirable, or an obligatory—way to express our reflective conviction as to 'the soundness of *S*'. But is that the only way to express this conviction? (Tennant 2002, p. 569)

To this last question, Tennant gives a negative answer. In his opinion this 'reflective conviction' is indeed important, but the deflationist has at his disposal a philosophically modest way of 'displaying' (rather than stating) it. As he says:

> If Shapiro demands that the deflationist do justice to the reflective intuition that *all S*-theorems are sound […] then we see no reason why we should not simply add […] the principle
>
> $$\mathrm{Pr}_S(\varphi) \rightarrow \varphi$$
>
> which produces the soundness extension. (Tennant 2000, p. 547)

In the quoted fragment, '$\mathrm{Pr}_S(\varphi)$' is an arithmetical formula which under natural interpretation states 'φ has a proof in *S*'.[2] In effect Tennant's strategy permits us to obtain the theory *S** (it is namely *S* extended by all the arithmetical instantiations of the reflection schema '$\mathrm{Pr}_S(\varphi) \rightarrow \varphi$'), which is obviously stronger then *S* itself. On Tennant's view this however should not be treated as a shortcoming of the proposed solution. On the contrary—the aim here is to present a realistic description of how the deflationist could arrive at stronger theories without burdening himself with any substantial notion of truth. The idea is as follows: start with the theory *S* you are currently using; then reflect about your axiomatic and deductive commitments and try to express them in the form of an appropriate reflection principle. In the process of reflection you note that you are ready to accept any sentence φ for which you can produce a proof in *S*. This gives you a reason to accept any sentence φ for which you can show that it's possible to produce its proof in *S*. In this way you arrive at a theory *S**, which reasonably approximates the statement 'All theorems of S are true'; however, both in *S** itself and in the process of arriving at it you can eschew the notion of truth altogether—the truth predicate does not appear in your reflection. And the whole story does not terminate at this point: in the next stage you can take *S** as your starting point and repeat the whole procedure; in this way *S* gives rise to a sequence of stronger and stronger theories, generated by the process of reflection.

How is one to justify the new reflection axioms, added to our theory? To this question Tennant has a short answer:

> No further justification is needed for the new commitment made by expressing one's earlier commitments. As soon as one appreciates the process of reflection, and how its outcome is expressed by the reflection principle, one already has an explanation of why someone who accepts *S* should also accept all instances of the reflection principle. (Tennant 2005, p. 92)

[2] Clearly some assumptions about *S* (like axiomatizability) are needed if we want to have a formula with good properties of the required sort.

The aim of this paper is to examine this rescue strategy. I will claim that: (1) Tennant's solution in its original form does not help the deflationist; (2) in a modified version it produces a theory, which is sufficiently strong to meet the demands of the critics. However, before turning to philosophical questions, I am going to state some formal results, which (up to my knowledge) have not been considered so far in the debate, and which in my opinion shed light on the issues involved.

2. Formal results

In what follows I am going to take Peano arithmetic as the base theory.[3] After extending the language of arithmetic with a new predicate 'Tr', let us denote as $PA(S)^-$ the theory obtained from *PA* by adding the usual Tarski clauses as new axioms. The stipulation is however, that in $PA(S)^-$ the induction schema will be restricted to arithmetical instantiations only—we are not allowed to use induction for formulas of the extended language, with the truth predicate (that is what 'minus' means here). It is a well known fact that the theory $PA(S)^-$ obtained in this way is a conservative extension of *PA*.[4] But Shapiro's and Ketland's worry is that $PA(S)^-$ fails as a theory of truth for the language of arithmetic exactly because it is conservative—it does not prove the global reflection principle for *PA*. Here however I will start with the claim that the problem is more basic: the theory in question does not even prove that logic is true—in fact $PA(S)^-$ + 'logic is true' is not a conservative extension of *PA*. In order to establish this result, the following theorem will be proved.

Theorem 1. $PA(S)^- + \forall\psi[\mathrm{Pr}_{\varnothing}(\psi) \rightarrow \mathrm{Tr}(\psi)] \vdash \forall\psi[\mathrm{Pr}_{PA}(\psi) \rightarrow \mathrm{Tr}(\psi)]$.

The expression '$\mathrm{Pr}_{\varnothing}(\psi)$' is an arithmetical formula with the intended reading 'ψ is provable from empty set of premises', that is: 'ψ is provable in logic' (in what follows I will use also a notation '$\varnothing \vdash \psi$' in this sense). In effect Theorem 1 reads: if we add to $PA(S)^-$ an additional assumption stating that logic is true, we will be able to prove that all theorems of *PA* are true. Let's turn now to the proof.

Proof. Working in the theory $PA(S)^- + \forall\psi[\mathrm{Pr}_{\varnothing}(\psi) \rightarrow \mathrm{Tr}(\psi)]$, fix ψ such that $\mathrm{Pr}_{PA}(\psi)$; we are going to show that $\mathrm{Tr}(\psi)$. Pick a proof *d* of ψ in *PA*; let $(W, \alpha_0, \ldots, \alpha_s)$ be a sequence of all the axioms of *PA* used in *d*, with $\alpha_0, \ldots, \alpha_s$ being all the induction axioms (*W* is the conjunction of the rest – a single, standard sentence, which could be written down explicitly). Then:

$$\varnothing \vdash (W \,\&\, \alpha_0 \,\&\, \ldots \,\&\, \alpha_s) \rightarrow \psi.$$

Therefore, since logic is true:

$$\text{If } \mathrm{Tr}(W \,\&\, \alpha_0 \,\&\, \ldots \,\&\, \alpha_s), \text{ then } \mathrm{Tr}(\psi).$$

We claim that $\mathrm{Tr}(W \,\&\, \alpha_0 \,\&\, \ldots \,\&\, \alpha_s)$, which will obviously finish our proof. Since $\mathrm{Tr}(W)$ (remember that *W* is a single, standard sentence), it is enough to show that $\mathrm{Tr}(\alpha_0 \,\&\, \ldots \,\&\, \alpha_s)$. For an indirect proof, assume that $\mathrm{Tr}(\neg(\alpha_0 \,\&\, \ldots \,\&\, \alpha_s))$. We have:

$$\varnothing \vdash \neg(\alpha_0 \,\&\, \ldots \,\&\, \alpha_s) \rightarrow (\neg\alpha_0 \vee \ldots \vee \neg\alpha_s).$$

Therefore by the assumption that logic is true:

$$\mathrm{Tr}(\neg\alpha_0 \vee \ldots \vee \neg\alpha_s).$$

[3] About the choice of our base theory, see Halbach 2001a. It seems that what is needed here is some theory adequate for the purposes of coding and arithmetization of syntax; in this respect *PA* looks like a natural choice.

[4] See Kotlarski, Krajewski, and Lachlan 1981. In addition, Kotlarski 1991 is a nice survey of the results obtained in the theory of full satisfaction classes. As for conservativeness, the situation would be completely different if we allowed substituting sentences of the extended language in the induction schema. Such a theory would prove 'All theorems of *PA* are true', and therefore it would not be a conservative extension of *PA*.

We assume that for $r \leq s$, α_r is of the form:

$$[\beta_r(0) \,\&\, \forall x(\beta_r(x) \rightarrow \beta_r(x+1))] \rightarrow \forall x \beta_r(x)$$

which means in effect that α_r is an induction axiom for a formula β_r. Now, denote by $\gamma(x)$ the following formula:

$$[\beta_0(0) \,\&\, \forall y(\beta_0(y) \rightarrow \beta_0(y+1)) \,\&\, \neg\beta_0(x)] \vee \ldots \vee [\beta_s(0) \,\&\, \forall y(\beta_s(y) \rightarrow \beta_s(y+1)) \,\&\, \neg\beta_s(x)]$$

Then we have:

$$\varnothing \vdash (\neg\alpha_0 \vee \ldots \vee \neg\alpha_s) \rightarrow \exists x \gamma(x).$$

So the truth of logic guarantees that $\mathrm{Tr}(\exists x \gamma(x))$; and by the properties of the truth predicate we obtain also: $\exists a \mathrm{Tr}(\gamma(a))$.

We note however that $\forall a \varnothing \vdash \neg\gamma(a)$. To see this it is enough to show that:

$$\forall r \leq s \forall a \varnothing \vdash \{\beta_r(0) \,\&\, \forall y[\beta_r(y) \rightarrow \beta_r(y+1)]\} \rightarrow \beta_r(a).$$

This can be easily proved by induction (the above formula belongs to the language of arithmetic—it does not contain the truth predicate—so induction can be used freely). Then we observe that for every a, logic proves the equivalence of $\neg\gamma(a)$ with the following formula:

$$\{(\beta_0(0) \,\&\, \forall y[\beta_0(y) \rightarrow \beta_0(y+1)]) \rightarrow \beta_0(a)\} \,\&\, \ldots \,\&\, \{(\beta_s(0) \,\&\, \forall y[\beta_s(y) \rightarrow \beta_s(y+1)]) \rightarrow \beta_s(a)\}.$$

So $\neg\gamma(a)$ is logically equivalent to the above conjunction; and moreover, each member of this conjunction is provable in logic. Therefore the conjunction itself is provable in logic (induction again—no truth predicate here!), so $\forall a \varnothing \vdash \neg\gamma(a)$.

Since logic is true, we obtain the conclusion: $\forall a \mathrm{Tr}(\neg\gamma(a))$, which ends the proof, producing the desired contradiction.

The next result is due to Kotlarski (1986), and it answers the question: how strong a theory is obtained by supplementing $PA(S)^-$ with an additional axiom, stating that all theorems of PA are true. It turns out that by adding this axiom we obtain no more and no less than the theory Δ_0-PA(S), which is simply $PA(S)^-$ with one modification: now we are allowed to substitute in the induction schema all the formulas of the extended language (with the truth predicate) which belong to the class denoted usually by Δ_0. This is a class of formulas with bounded quantifiers only, that is: all the quantifiers in formulas belonging to Δ_0 are of the form '$\exists x < y$' or '$\forall x < y$'. Kotlarski's result is formulated below.

Theorem 2. (Kotlarski 1986) $PA(S)^- + \forall\psi[\mathrm{Pr}_{PA}(\psi) \rightarrow \mathrm{Tr}(\psi)] = \Delta_0$-PA(S).

From Theorem 1 and Theorem 2 the following corollary can be easily obtained.

Corollary 1. The following theories are equivalent to Δ_0-PA(S):

T_1 $PA(S)^- + \forall\psi[\mathrm{Pr}_{PA}(\psi) \rightarrow \mathrm{Tr}(\psi)]$,
T_2 $PA(S)^- + \forall\psi[\mathrm{Pr}_{\varnothing}(\psi) \rightarrow \mathrm{Tr}(\psi)]$,
T_3 $PA(S)^- + \forall\psi[\mathrm{Pr}_{\mathrm{Tr}}(\psi) \rightarrow \mathrm{Tr}(\psi)]$,
T_4 $PA(S)^- + Con_{Tr}$,
T_5 $PA(S)^- + \{\mathrm{Pr}_{\mathrm{Tr}}(\psi) \rightarrow \psi : \psi \in L(PA)\}$.

The expression '$\mathrm{Pr}_{\mathrm{Tr}}(x)$' denotes here the formula of the extended language (with the truth predicate) whose natural reading is 'there is a proof of x from the set of true assumptions'. By 'Con_{Tr}' I denote the sentence with a natural reading 'The set of true sentences is consistent'. $L(PA)$ is the language of Peano arithmetic.

Proof.

(1) $T_1 \subseteq T_2$.
This was the content of Theorem 1.

(2) $T_2 \subseteq T_3$.

Obvious

(3) $T_3 \subseteq T_4$.
Assuming Con_{Tr}, fix ψ such that $Pr_{Tr}(\psi)$ and $\neg Tr(\psi)$. Then $Tr(\neg\psi)$, and so Tr is inconsistent.

(4) $T_4 \subseteq T_5$.
In T_5 we have: $Pr_{Tr}(\ulcorner 0 \neq 0 \urcorner) \rightarrow 0 \neq 0$. Since $0 = 0$, we obtain by contraposition $\neg Pr_{Tr}(\ulcorner 0 \neq 0 \urcorner)$, in effect: Con_{Tr}.

(5) $T_5 \subseteq T_1$.
We know from Theorem 2 that $T_1 = \Delta_0$-PA(S), so it is enough to observe that for every ψ, Δ_0-PA(S) $\vdash Pr_{Tr}(\psi) \rightarrow \psi$. Indeed, working in Δ_0-PA(S) assume that $Pr_{Tr}(\psi)$ and let $(\alpha_0 \ldots \alpha_s)$ be a proof of ψ from true premises. Then it is possible to show by induction that $\forall r \leq s Tr(\alpha_r)$ (the last formula belongs to the class Δ_0, so we may use induction freely). Therefore $Tr(\psi)$, and since ψ is standard, we obtain: ψ.

Recall now our basic predicament. Shapiro's and Ketland's worry was that a deflationary (i.e. conservative) theory of arithmetical truth built over *PA* cannot prove the strong form of reflection: 'All theorems of *PA* are true'. What Theorem 1 shows is that in fact a deflationary of truth for *PA* can prove 'All theorems of *T* are true' for no theory *T* at all, no matter how *T* is characterized in the arithmetical language. As we can see, the problem starts already with the empty theory—with logic.[5]

Theorem 2 and Corollary 1 give us information about how strong the reflective theory is, providing various alternative axiomatizations. Before I venture further, let me make here one quick comment. When we consider a sentence like 'All theorems of *PA* are true', we may be ready to think about it as expressing an important fact concerning our arithmetical theory; however, it may not be obvious to us from the start that this sentence expresses also an essential property of our notion of truth. In this respect the reformulation in terms of T_2 or T_3 may be quite revealing: from a philosophical point of view, it is perhaps not so much the relation between truth and *PA*, but between truth and logic, or truth and provability, which matters. What I want to say in effect is that there is more intuitive plausibility to the claim that something like 'Truth is closed under provability' (the content of T_3) expresses an essential property of our notion of truth than that '*PA* is true' does this; although with *PA* as a base, both theories (T_1 and T_3) turn out to be the same.

3. Philosophical discussion

In this paper I want to make the following philosophical claims:

(1) Even if correct, Tennant's argument in its original form is useless to the deflationist.

(2) Tennant's argument can be reformulated in such a way as to give the deflationist a strong theory of truth, probably sufficient for his aims.

I will start with formulating Tennant's argument in the first sub-section below. In the next two sub-sections claims (1) and (2) will be discussed and defended.

3.1 The formulation of Tennant's argument

[5] A qualification is needed here: the assumption 'logic is true' used in Theorem 1 should be read as 'logic in the full arithmetical language (with addition and multiplication) is true'. It does not follow from Theorem 1 that for example $PA(S)^-$ + 'Presburger's arithmetic (arithmetic of addition) is true' is not conservative over *PA*.

Tennant's proposal contains two elements: a descriptive and normative one. On the descriptive side, the process begins with reflecting on my deductive commitments as a user of *PA*: I am ready to accept each sentence φ for which I can furnish a proof in *PA*. Let me formulate it explicitly. In the first step of the process of reflection I accept the following statement:

(D) For any sentence φ, if φ has a proof in *PA*, then I am ready to accept φ.

As I take it, the status of (D) is descriptive. It is a factual statement, concerning the way I use the axioms of *PA* and its proof machinery. I may arrive at (D) by introspection or by some sort of empirical generalization—it does not matter. In what follows I will just assume that I can indeed arrive at (D) without using any concept of truth (just the pragmatic concept of 'accepting' or 'asserting' a given sentence). One could say in effect that (D) expresses simply my trust in *PA* and its proof machinery.

In the next part of the process comes the formalization: I realize that (some of) the content of (D) can be expressed by the infinite set of arithmetical sentences of the form '$\mathrm{Pr}_{PA}(\varphi) \rightarrow \varphi$'—call it the set of reflective axioms. The formalization claim is:

(F) The set of reflective axioms expresses (part of) the content of (D).

And now comes the normative thesis:

(P) Anyone who accepts *PA* should also accept all instances of the reflection schema.

The argument for (P) is as follows: we note that any person accepting *PA* should also accept (D). The reason is that (D) expresses simply the fact that the person in question accepts *PA*; and the claim would be that the data on which (D) is based, whether introspective or empirical, are in principle easily accessible to any rational human being, so it would be a grave mistake to ignore them. In effect, since I have a reason to accept (D), then by (F) I have also a reason to accept all the reflective axioms.

In assessing the above argument, the crucial question is: what is meant here by 'accepting *PA*'? The natural interpretation goes as follows: to accept *PA* means to be ready to accept every sentence for which a proof from the axioms of *PA* can be furnished. On this approach, it is (D) that gives us the meaning of 'I accept *PA*'. And with this interpretation adopted, I find Tennant's argument convincing.[6]

3.2 A criticism of Tennant's argument

The deflationists do not claim that truth is redundant. Quite on the contrary: they stress repeatedly the usefulness of the truth predicate for expressing generalizations. Let *A* be an infinite set of some arithmetical sentences, which we accept.[7] How can we express our acceptance of all sentences belonging to *A*? Assume for a start that we have an arithmetical formula $\alpha(x)$ which defines *A*. Without the truth predicate, we could express our acceptance of all the elements of *A* by means of an infinite conjunction of the form '$(\alpha(\ulcorner\psi_1\urcorner) \rightarrow \psi_1) \,\&\, (\alpha(\ulcorner\psi_2\urcorner) \rightarrow \psi_2) \,\&\, \ldots$', with ψ_1, ψ_2 … being an enumeration of all the

[6] That is, provided that we take (F) for granted. Indeed, one could still wonder about the exact sense, in which the reflective axioms express some of the content of (D)—what does 'express' mean here? It is an intricate question, which I am not going to discuss in this paper—I will just concentrate on showing what can be achieved if we accept (F) as given.

[7] For example *A* is a set of all the instances of the law of excluded middle.

sentences of our language. Having a truth predicate at our disposal, we are able to express it by a single sentence of our language. We state:

(*) $\forall\psi[\alpha(\psi) \rightarrow \mathrm{Tr}(\psi)]$.

According to the deflationist, that is what truth is for.[8]

In view of that, I formulate now the following requirement.

(R) The deflationist should have at his disposal a theory, which proves the basic, sound instances of (*).

The reason behind (R) is that if it is not satisfied, the truth predicate still seems useless (contrary to what the deflationist claims). What is the point of having generalizations—say of the type (*)—expressible in our language, if we do not have the slightest idea of how to arrive at them and how to use them in proofs? Without (R), deflationism would become after all a sort of redundancy theory of truth, which (as the deflationists themselves claim) it is not. In assessing deflationary theories we are therefore entitled to the following strategy: we may consider examples of most basic, intuitive generalizations and ask how the deflationist can explain our acceptance of them. His inability to do that would undermine his philosophical views—that is the outcome.

What Theorem 1 gives us is exactly an example of a basic generality of the required sort, not provable in any conservative truth theory. Confronted with this, what does Tennant's strategy amount to? As I take it, it amounts to rejecting (R). In Tennant's opinion proving basic instances of (*) is not 'obligatory'—we do not need them in our theory.[9] But according to the deflationist that is what truth is for. And this is my reason for concluding: from a deflationary point of view, Tennant's strategy is useless.

3.3 A way out: reflecting on logic

In what follows I am going to propose a Tennant-style argument with the intention of overcoming the deficiencies of the original reflective reasoning, discussed in sub-section 3.2. Taking (R) for granted, I will try to show that the deflationist has a 'deflationary licit' way of arriving at a strong theory, in which the truth predicate adequately performs its generalizing role. This theory will be *PA*(*S*)—arithmetic with Tarski's 'inductive clauses' and full induction for the extended language, with the truth predicate.

In this context the following observation will be useful.

Observation 1. Let Ref_{Log} be the set of all the instantiations of the reflection schema

$$\forall x[\mathrm{Pr}_{\varnothing}(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow \varphi(x)\ ],$$[10]

[8] See Horwich 1990, pp. 31–34. According to Horwich, the truth predicate is useful because it permits us to formulate such generalizations like 'for every *x*, if *x* is a proposition of the form $<p \rightarrow p>$, then *x* is true' (p. 33). On p. 34 he adds: 'And as for alternative functions that [the truth predicate] might have, there simply aren't any plausible candidates'. Tennant did not specify such 'plausible candidates' either.

[9] Alternatively, Tennant could claim that it is possible to accept (R) but reject the generalization 'Logic is true' as not basic enough—as an undesirable instance of (*). However, such a move would require a justification. It is not enough to say 'we reject it because it produces a nonconservative extension'—I find such an answer unacceptable. The deflationist needs to present arguments in favour of conservative truth theories, not to take conservativeness for granted.

[10] The intuitive reading is: 'for every *x*, if logic proves a sentence obtained from φ(.) by substituting a numeral denoting *x* for a free variable in φ(.), then φ(*x*)'. A numeral denoting *x* is a term of the form '*S*...*S*(0)', with the successor symbol *S* repeated *x* times. Some reflection principles of this sort were discussed by Halbach (2001b).

where $\varphi(x)$ is a formula of the extended language. Let T be $PA(S)^{-} + Ref_{Log}$. Then T proves all instances of induction for the extended language.

Proof. Take a formula $\varphi(x)$ of the extended language and assume in T that $\varphi(0) \,\&\, \forall x[\varphi(x) \rightarrow \varphi(x+1)]$. Fix an object a; our aim is to show that $\varphi(a)$. We observe:

$$\mathrm{Pr}_{\varnothing}(\ulcorner[\varphi(0) \,\&\, \forall x(\varphi(x) \rightarrow \varphi(x+1))] \rightarrow \varphi(a)\urcorner).$$

Therefore by reflection for logic:

$$[\varphi(0) \,\&\, \forall x(\varphi(x) \rightarrow \varphi(x+1))] \rightarrow \varphi(a).$$

But then $\varphi(a)$.

As we see, a quite weak reflection schema added to $PA(S)^{-}$—just reflection for logic in the extended language—is enough to give us all benefits of full $PA(S)$.

The reflective reasoning goes now as follows. Imagine a deflationist who at a starting point accepts $PA(S)^{-}$. (Since this theory is conservative over PA, he is entitled to adopt it.) The deflationist may claim that the axioms of $PA(S)^{-}$ fully characterize his notion of truth. At the next stage, our deflationist engages in the process of reflection. He argues as follows:

(D') For any sentence φ, for every a, if $\varphi(a)$ has a proof in pure logic, then I am ready to accept $\varphi(a)$.

(F') The set Ref_{Log} expresses (part of) the content of (D').

And now comes the normative thesis:

(P') Anyone who accepts logic (for the extended language) should also accept all instances of reflection for logic.

The reasoning leading to (P') mimics the argument for (P) given earlier in this paper. I will not repeat the details; I just conclude that in effect the deflationist has a right to extend his initial theory $PA(S)^{-}$ with new reflection axioms. In this way he obtains the theory T from Observation 1. But T is a strong theory—it proves global reflection for Peano arithmetic. In effect Ketland's demand is satisfied, which leaves the deflationist in a quite comfortable position.

I would like to conclude the paper with two short comments.

Comment 1. In the new axioms belonging to the set Ref_{Log} the truth predicate is used—it is after all a reflection for the extended language which is needed to carry out the argument in the proof of Observation 1. Is it possible to use this fact against the deflationist? The worry could be: '$PA(S)^{-}$ does not give us a complete characterization of the meaning of the truth predicate. It is rather $PA(S)^{-} + Ref_{Log}$ that does it; and this theory is not conservative over PA'. I do not find this objection convincing. The crucial question is how we arrive at the axioms of our theory. For a start, consider an axiom:

(Tr-Neg) $\forall\varphi[\mathrm{Tr}(\ulcorner\neg\varphi\urcorner) \leftrightarrow \neg\mathrm{Tr}(\varphi)]$.

When we ask the deflationist 'How do you justify (Tr-Neg)?', he would possibly answer: 'That is how I understand truth and negation—my axiom just formalizes the way I use these notions'. On this approach, (Tr-Neg) gives us a partial analysis of the concept of truth. But how does the deflationist justify his reflection axioms? The key consideration is that no appeal to the concept of truth is needed in this context, just reflection on the fact that the

person in question accepts logic in the extended language. No analysis of the notion of truth is involved here—that would be the deflationist's answer.

Comment 2. The deflationist might wish to apply the above reflective strategy in order to obtain a still stronger theory of self-referential truth.[11] And indeed, this option is available. Imagine that he starts with KF^-—that is, with Kripke-Feferman theory with induction for arithmetical formulas only. It has been shown that KF^- is conservative over *PA* (see Cantini 1989, Corollary 5.9). By the same reasoning as before, extending KF^- with reflection for logic produces full *KF*; in effect the deflationist may use a Tennant-style argument to explain his acceptance of *KF*. Again, this seems a desirable consequence of the present approach.

This is not to say that I advocate *KF* as the proper theory of truth for the deflationist. What seems problematic is again the deflationist's account of the generalizing role of the truth predicate. In *KF* we are not able to express in a general form our acceptance of logic—we are not able to derive a generalization '$\forall\psi[\mathrm{Pr}_\varnothing(\psi) \rightarrow \mathrm{Tr}(\psi)]$', with all the sentences of the extended language falling within the scope of the general quantifier.[12] All I want to say here is that on the present approach a move towards theories of self-referential truth presents itself as a viable option.[13]

University of Warsaw CEZARY CIEŚLIŃSKI
Institute of Philosophy
Krakowskie Przedmieście 3
00-927 Warsaw
Poland
c.cieslinski@uw.edu.pl

References

Cantini, Andrea 1989: 'Notes on Formal Theories of Truth'. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 35, pp. 97–130.
Field, Hartry 2006: 'Truth and the Unprovability of Consistency'. *Mind*, 115, pp. 567–606.
Halbach, Volker 2001a: 'How Innocent is Deflationism?'. *Synthese*, 126, pp. 167–94.
——2001b: 'Disquotational Truth and Analyticity'. *Journal of Symbolic Logic*, 66, pp. 1959–73.
Horwich, Paul 1990: *Truth*. Oxford: Basil Blackwell.
Ketland, Jeffrey 1999: 'Deflationism and Tarski's Paradise'. *Mind*, 108, pp. 69–94.
——2000: 'Conservativeness and Translation-dependent T-schemes'. *Analysis*, 60, pp. 319–27.
——2005: 'Deflationism and the Gödel Phenomena: Reply to Tennant'. *Mind*, 114, pp. 75–88.
Kotlarski, Henryk, Krajewski, Stanisław, and Lachlan, H. Alistair 1981: 'Construction of Satisfaction Classes for Nonstandard Models'. *Canadian Mathematical Bulletin*, 24, pp. 283–93.

[11] About theories of self-referential truth, see Sheard 1994.

[12] Let ψ be the liar sentence. We will be able to prove in *KF* that $Pr_\varnothing(\ulcorner\psi \vee \neg\psi\urcorner)$, but *KF* does not prove that $Tr(\ulcorner\psi \vee \neg\psi\urcorner)$, unless it is inconsistent. See Field 2006, pp. 572–574.

[13] This work has been financed by 2008–2009 scientific grant for research projects, grant number: NN101034235. I would like to thank Konrad Zdanowski and the anonymous referees from the *Mind* journal for their valuable comments.

Kotlarski, Henryk 1986: ‘Bounded Induction and Satisfaction Classes’. *Zeitschrift für Mathematische Logik*, 32, pp. 531–44.
——1991: ‘Full Satisfaction Classes: a Survey’. *Notre Dame Journal of Formal Logic*, 32, pp. 573–9.
Shapiro, Stewart 1998: ‘Proof and Truth—through Thick and Thin’. *Journal of Philosophy*, 95, pp. 493–522.
Sheard, Michael 1994: ‘A Guide to Truth Predicates in the Modern Era’. *Journal of Symbolic Logic*, 59, pp. 1032–54.
Tennant, Neil 2002: ‘Deflationism and the Gödel Phenomena’. *Mind*, 111, pp. 551–82.
——2005: ‘Deflationism and the Gödel Phenomena: Reply to Ketland’. *Mind*, 114, pp. 89–96.