Deflationary truth and pathologies

Cezary Cieśliński
Institute of Philosophy, the University of Warsaw,
Poland

Abstract

By a classical result of Kotlarski, Krajewski and Lachlan, pathological satisfaction classes can be constructed for countable, recursively saturated models of Peano arithmetic. In this paper we consider the question of whether the pathology can be eliminated; we ask in effect what generalities involving the notion of truth can be obtained in a deflationary truth theory (a theory of truth which is conservative over its base). It is shown that the answer depends on the notion of pathology we adopt. It turns out in particular that a certain natural closure condition imposed on a satisfaction class - namely, closure of truth under sentential proofs - generates a nonconservative extension of a syntactic base theory (Peano arithmetic).

1 Preliminaries

1.1 Basic concepts

The notion of a satisfaction class was introduced in order to characterize semantics for nonstandard formulas.[1] Given any nonstandard model M of Peano arithmetic (PA), an arithmetical predicate $Sent(x)$ with an intuitive reading "x is a sentence of the language of PA" will define in M a set containing nonstandard numbers. These numbers can be treated as nonstandard arithmetical sentences (sentences in the sense of M).

In an attempt to describe semantics for such nonstandard languages one defines a satisfaction class: a subset of the universe of the model in question, which - when treated as an interpretation of the truth predicate - will indeed behave in a "truthlike" manner, i.e. it will satisfy Tarski's compositional truth axioms. A membership in such a class serves us then as an explication of the notion of truth for nonstandard sentences. Formally, a satisfaction class will be defined in terms of the theory $PA(S)^{-}$, which is Peano arithmetic with Tarski's "inductive axioms", characterizing the notion of truth. A precise definition of this theory is given below. In what follows it will be assumed that the language of arithmetic contains function symbols for the successor operation, addition and multiplication.[2] We denote as "$Tm^c(x)$" an arithmetical formula with the intuitive reading "x is a closed term of the language of arithmetic"; in an analogous manner, $Sent(x)$, and $Var(x)$ read "x is a sentence" and "x is a variable".

Definition 1 *Let $PA(S)^{-}$ be Peano arithmetic with the following additional axioms:*

1. $\forall t, s, [Tm^c(t) \wedge Tm^c(s) \Rightarrow (Tr(\ulcorner t = s \urcorner) \equiv val(t) = val(s))]$
2. $\forall \psi[Sent(\psi) \Rightarrow (Tr(\ulcorner \neg\psi \urcorner) \equiv \neg Tr(\psi))]$
3. $\forall \varphi, \psi[Sent(\psi) \wedge Sent(\varphi) \Rightarrow (Tr(\ulcorner \varphi \wedge \psi \urcorner) \equiv Tr(\varphi) \wedge Tr(\psi))]$
4. $\forall \varphi(x)\forall a[Var(a) \Rightarrow (Tr(\ulcorner \forall a\varphi(a) \urcorner) \equiv \forall v Tr(\ulcorner \varphi(\dot{v}) \urcorner))]$

The expression "$\forall\varphi(x)$" is taken to express a quantification over formulae with at most one free variable. The dot notation is used here in its usual sense; the expression "$\forall v Tr(\ulcorner \varphi(\dot{v}) \urcorner)$" means that every result of substituting a numeral for a free variable in φ is true.[3] Observe that in $PA(S)^{-}$ we do not have axioms of induction for formulas of the extended language (with the truth predicate). The theory obtained by supplementing also these additional induction axioms will be denoted as $PA(S)$.

The notion of a satisfaction class is then defined in the following way.

Definition 2 *Let M be a model of PA and let S be a subset of M. We say that S is a satisfaction class for M iff $(M, S) \models PA(S)^{-}$.*

In effect a satisfaction class is an interpretation of the predicate "Tr" which makes the truth axioms true. If in addition $(M, S) \models PA(S)$, then we say that S is an inductive satisfaction class for M.

Which models of PA have a satisfaction class? An answer to this question was provided by Kotlarski, Krajewski and Lachlan in [11] (from now on the abbreviation "KKL" will be used): it turns out that such a class can be constructed for every countable, recursively saturated model of PA. There is however a disconcerting twist: many satisfaction classes will contain pathological sentences. A precise formulation of this result is given below.

Theorem 1 (KKL) *Let M be a countable, recursively saturated model of PA. Let φ be an element of M such that for a given nonstandard a in M:*

$$M \models \text{“}\varphi = \ulcorner \underbrace{0 \neq 0 \vee ... \vee 0 \neq 0}_{a\ times} \urcorner\text{”}.$$

Then M has a satisfaction class containing φ.[4]

More precisely, a sentence φ mentioned in Theorem 1 can be specified in the following way. Define ψ_0 as $\ulcorner(0 \neq 0)\urcorner$, define ψ_{k+1} as $\psi_k \vee \psi_k$. We can characterize then our φ from Theorem 1 as ψ_a for a nonstandard $a \in M$.

Intuitively, a satisfaction class containing φ is pathological: the sentence φ is clearly false (it is obtained by iterating an obviously false disjunct); but the model thinks it is true - i.e. it belongs to the satisfaction class for this model.

From Theorem 1 the following corollary can be obtained:

Corollary 1 *$PA(S)^-$ is a conservative extension of PA.*

This is due to the well known fact that for each infinite model M we can find an elementarily equivalent, recursively saturated structure K of the same cardinality as M (see e.g. [9], Proposition 11.4, p. 14). Observe however that Corollary 1 has nothing to do with the pathology mentioned in Theorem 1 - it is just the possibility of constructing a satisfaction class for an arbitrary countable, recursively saturated model of PA which makes it true.

On the other hand, $PA(S)$ with full induction for the extended language is not conservative over PA. In $PA(S)$ one can prove e.g. that all theorems of PA are true, from which consistency of PA follows - and the last statement is clearly not derivable in PA itself.

In this paper I am going to discuss the question of whether the pathology mentioned in Theorem 1 can be eliminated. The next subsection presents philosophical motivations for my investigations.

1.2 Philosophical motivations

The main motivation for this research stems from a recent philosophical debate on deflationism and conservativeness.[5] The following two claims seem to be central to deflationary standpoint:

(1) Truth is insubstantial.

(2) The truth predicate is a purely logical device - its only role consists in permitting us to produce generalizations of the sort "All substitutions of the law of excluded middle are true" - without the truth predicate we wouldn't have the linguistic means for expressing our simultaneous acceptance of all the sentences belonging to an infinite set.[6]

Conservativeness condition was proposed as an explication of (1): a theory of truth built over some syntactic base theory B should not permit us to establish any new facts in the language of B - facts which cannot be proved already in B. Otherwise it would seem that our notion of truth has indeed a lot of nonsemantic content and can't be called "insubstantial" in any decent sense of this word. It should be stressed at this point that this explication enables the deflationist to propose truth theories which are quite rich indeed, at least in comparison with austere proposals like Horwich's minimal theory, having just T-sentences as axioms. Considering some richer theories, involving not just T-sentences but also compositional principles, McGee wrote:

> The conservativeness theorems show that the disquotationalist is permitted to avail herself of a theory of truth a lot richer than what we get merely by taking the T-sentences as axioms. [...] Taking the richer theory as axiomatic is something the disquotationalist is permitted to do without endangering her standing as a disquotationalist. ([15], p. 108.)

This approach will be adopted also in the present paper. On this proposal the deflationist is construed as an adherent of conservative truth theories - his claim would be that such theories are adequate for all the purposes we might have in introducing the truth predicate into our language. These purposes are characterized by thesis (2): it is the generalizing role of truth that matters. In this context the crucial issue is which generalizations involving the notion of truth should be decided by an adequate truth theory. (Some of them, like "All theorems of B are true", are clearly beyond the scope of any conservative truth theory built over B.) This issue was discussed by Halbach in [7]. He wrote:

> As far as I can see, only one really sensible answer has emerged. A natural strengthening of the T-sentences is achieved by picking the "inductive clauses" for truth. They allow the deflationist to prove many interesting generalizations in a natural way. ([7], p. 184.)

And in a footnote he added:

> Now, after nearly seven decades of addiction to them, the "inductive" clauses have proven to be natural axioms and all generalizations not provable from them seem to be better left undecided by a good theory of truth. ([7], p. 192, footnote 23.)

In what follows we will concentrate exclusively on the notion of arithmetical truth; Peano arithmetic will play the role of a base theory B. Accordingly, the relevant truth axioms will characterize the notion of truth just for the language of arithmetic. The scope of our investigation is thus rather limited - obviously one could consider both weaker and stronger base theories in place of PA.[7] However, it should be kept in mind that building a truth theory over PA is treated here not as an aim in itself, but as a convenient paradigm case - in my opinion philosophical claims endorsed by the deflationist would be worthless if they fail a model test, provided by the task of constructing a suitable notion of arithmetical truth.

In effect the "inductive" clauses for truth mentioned by Halbach form the set of axioms of $PA(S)^{-}$. Should we rest happy with $PA(S)^{-}$ then? That would be a rather hasty conclusion.[8] There is no need for us to disagree with Halbach: the clauses in question seem to enjoy a privileged status in (nearly) all discussions on the subject. But their special character stems from the fact that these natural truth axioms combine nicely with induction for formulas of the extended language (with the truth predicate), producing a strong theory which we denoted as $PA(S)$. In other words: if we have all the induction there is to have, we do not need any additional truth axioms to derive strong and natural truth-theoretical conclusions. However, this sort of consideration hardly matters in our context. We do not have this induction - accepting it would mean that we are opting quite outrightly for a nonconservative truth theory, prejudging the issue against the deflationist. And it would be interesting to know how far we can go (conservatively) beyond $PA(S)^{-}$ in pursuing the aim of proving truth-theoretical generalizations. This is exactly the point where an attempt to eliminate pathology from a satisfaction class can provide us with some information. Theorem 1 shows that the following generalization is not provable in $PA(S)^{-}$: take a false sentence α, produce a disjunction of an arbitrary length with α as the only disjunct, and the result of your operation will also be false. This raises the question of which generalizations can be salvaged? It is quite possible that various interesting generalizations involving the notion of truth not derivable in $PA(S)^{-}$ can nevertheless be obtained in some natural truth theory which contains $PA(S)^{-}$, but is still conservative over PA. In my opinion it is a serious issue, which should be considered by the deflationist insisting on the generalizing role of truth as the main reason for its usefulness.

2 Some known results

In this section I will formulate some results established elsewhere in the literature. However, in order to discuss in a precise manner the issue of eliminability of pathologies, we must first answer two questions: (a) what does it mean to "eliminate" pathology? (b) what is pathology?

As to (a): imagine that a set P of pathological instances is somehow characterized. An elimination of pathology could be achieved by constructing, for an arbitrary countable, recursively saturated model of PA, a satisfaction class which does not contain any member of P. Another way would consist in proving a conservativeness result: our aim would be to show that if we extend $PA(S)^-$ with an axiom of the form "No element of P is true", we obtain a conservative extension of PA.[9]

As to (b), there are various possible approaches one could take. In order to suggest some directions (and at the same time to stress the specific deficiencies of $PA(S)^-$ as a theory of truth) let us ask why exactly we are inclined to think about KKL's case as pathological. I discern below three reasons for that.

(1) The pathological sentence φ presented by KKL belongs to the class Δ_0 (indeed, it is even quantifier free!). It is a well known fact that we have an arithmetical truth predicate $Tr_{\Delta_0}(.)$ available for Δ_0 sentences of the language of arithmetic (see e.g. [5], p. 56, Theorem 1.70). And it is not difficult to prove that in our model φ will not belong to the extension of Tr_{Δ_0} (since $Tr_{\Delta_0}(.)$ is arithmetical, we are free to use induction for it). In effect in $PA(S)^-$ we have no guarantee that our general notion of truth coincides with the partial ones. This, I think, presents a problem. A truth theorist opting for $PA(S)^-$ would have to handle somehow the question about the relation between arithmetically expressible notions of truth and the general, arithmetically inexpressible concept. What is the link between them? Should we treat the former ones as limited variants of the latter? If so, in what sense? Some explanation is needed here.[10] Anyway, we pinpoint here the first reason of pathologicality of φ and we generalize it in the following manner: a class P of pathological cases will consist of all the sentences ψ (in the sense of the model M) such that for some natural number n, $M \models Tr_n(\ulcorner \neg\psi \urcorner)$, with "$Tr_n(.)$" being an appropriate partial truth predicate.

An example of a pathology with a standard, but non-zero quantifier rank can be obtained by adding to "$0 \neq 0$" a quantifier prefix of a standard complexity, but containing nonstandardly long sequences of similar quantifiers (e.g. for rank 1 we could consider a formula "$\exists x_0 ... x_a 0 \neq 0$" with a nonstandard).

(2) The second reason of the pathological character of KKL's formula φ is

that its negation is provable in pure logic (and the model thinks it is "true"!). Indeed, it is easy to prove by induction in M that

$$\forall a \; Pr_{\emptyset}(\neg(\underbrace{0 \neq 0 \vee ... \vee 0 \neq 0}_{a \; times})).$$

That is: no matter how long the disjunction is, its negation is provable in logic. On this approach, the set of pathologies would be simply the set of all sentences disprovable in first order logic. (As an example, we could consider expressions of the form "$\exists x_0 ... x_a \varphi \equiv \neg\varphi$" for a nonstandard number a. The quantifier ranks of such sentences can be arbitrarily large, depending on the choice of φ. Their negations are provable in pure logic, but they are made true by some satisfaction classes.) In an attempt to eliminate the pathology one could try to construct a satisfaction class which makes logic true.

(3) The third option to consider is based on the following fact:

Fact 1 *Let S be a satisfaction class for M containing φ. Then there is a sentence ψ provable in sentential logic such that $\ulcorner \neg\psi \urcorner \in S$.*

Proof.

Define ψ as the sentence: "$\varphi \Rightarrow 0 \neq 0$". Obviously ψ is provable in sentential logic, but it cannot belong to S because otherwise "$0 \neq 0$" would belong to S, which is impossible. Therefore $\ulcorner \neg\psi \urcorner \in S$.

□

In view of the above fact we could demand that a satisfaction class does not contain any sentence disprovable in sentential logic - it is the class of such sentences that would be considered pathological on this approach.

In what follows I will comment briefly on (1) and (2), presenting results on (in)eliminability of pathologies obtained elsewhere in the literature.

2.1 Partial truth predicates

The pathology understood in the sense of (1) turns out to be eliminable. In fact, a theory obtained from $PA(S)^{-}$ by adding axioms which stipulate that partial truth predicates coincide with the general one for appropriate classes of formulae, is a conservative extension of PA. This result is an immediate corollary from the following theorem obtained by Engström:

Theorem 2 *Let M be a countable, recursively saturated model of PA and let n be a natural number. Then M has a satisfaction class S such that:*

$$(M,S) \models \forall\psi \in \Sigma_n\ [Tr_{\Sigma_n}(\psi) \equiv Tr(\psi)].$$

For the proof, see [3], p. 56-57. By an easy argument from compactness, it follows that a theory obtained from $PA(S)^{-}$ by adding as new axioms all sentences of the form "$\forall\psi \in \Sigma_n\ [Tr_{\Sigma_n}(\psi) \equiv Tr(\psi)]$" for all natural numbers n, is a conservative extension of PA.

2.2 First order logic

The theorem formulated below was obtained by Cieśliński in [2]:

Theorem 3 *The following theories are equivalent:*

T_1 $\quad PA(S)^{-} + \forall\psi\ [Pr_{\emptyset}(\psi) \Rightarrow Tr(\psi)]$

T_2 $\quad PA(S)^{-} + \forall\psi\ [Pr_{Tr}(\psi) \Rightarrow Tr(\psi)]$

T_3 $\quad \Delta_0\text{-}PA(S)$

The expression "$Pr_{\emptyset}(\psi)$" denotes an arithmetical formula with a natural reading "ψ is provable in logic"; "$Pr_{Tr}(\psi)$" is a formula of the extended language which reads "ψ is provable from true premises". In effect T_1 is a theory obtained by supplementing $PA(S)^{-}$ with the membership condition for first order logic: "All first order tautologies are true". T_2 contains a closure condition: "Truth is closed under first order provability". Both T_1 and T_2 turn out to be equivalent with the theory denoted here as Δ_0-$PA(S)$. It is $PA(S)^{-}$ supplemented with all the induction axioms for just those formulas of the extended language, which belong to the class Δ_0. However, Δ_0-$PA(S)$ is not a conservative extension of PA - Δ_0 induction for the extended language permits us to prove consistency of Peano arithmetic. It turns out that the pathology in the sense (2) is ineliminable - a satisfaction class which makes logic true cannot be constructed for an arbitrary recursively saturated model of PA. In general, a conservative truth theory is too weak to prove the truth of first order logic.

The degree of damage which this result inflicts on the deflationist is a debatable issue. Some authors claim that various generalities involving the notion of truth (notably "Peano arithmetic is true") do not have to be provable in a truth theory as such - on the contrary, additional non-truth-theoretic principles may be used in order to derive them, and the deflationist is committed only to conservativeness of his specific set of "truth-theoretic" axioms over the base theory.[11] But even so, the question still remains about the possible room for manoeuvre for the deflationist, compiling his list of "specifically truth-theoretic" axioms. Which of them are admissible if we take conservativeness as our guiding constraint? In the sections to follow some further candidates will be considered.

3 Truth and propositional logic

The results described in the last section still leave us with the question: which truth-theoretic generalizations can be salvaged? Which of them are accessible to an adherent of a conservative truth theory? A natural starting point for further investigation is propositional logic. We saw that truth of propositional logic is not derivable in $PA(S)^{-}$ (Fact 1); but what about possible (conservative over PA) extensions of $PA(S)^{-}$? A still more general question would concern possible closure conditions, which could be imposed (conservatively) on a set of true sentences. We would like to have a satisfaction class which is closed under logic in some nontrivial and not too weak sense of the word. How far can we go in this direction without compromising conservativeness?

In this section I want to show a (partial) answer to this question - a result concerning the closure condition for propositional logic. It turns out that closure of truth under propositional logic produces a nonconservative truth theory (in fact this extension is again Δ_0-$PA(S)$). This is the content of the theorem formulated below.

Theorem 4 *Denote by T a theory: $PA(S)^{-} + \forall\psi[Pr^{Sent}_{Tr}(\psi) \Rightarrow Tr(\psi)]$. Then $T = \Delta_0$-$PA(S)$.*

In the above formulation, $Pr^{Sent}_{Tr}(x)$ is a one place predicate of the extended language (with the truth predicate) which reads: "x is provable from true premises in sentential logic" (no special rules or axioms for handling quantifiers are allowed). In effect apart from $PA(S)^{-}$, T contains an additional axiom stating that truth is closed under propositional logic.

It is very easy to show in Δ_0-$PA(S)$ that truth is closed under sentential logic, so I will take as obvious that $T \subseteq \Delta_0$-$PA(S)$. Our task is to prove the opposite inclusion.

Proof. Let M be a model of T; we are going to show that M satisfies induction for Δ_0 formulas of the extended language. We start with defining, for an arbitrary Δ_0 formula φ of the extended language, a translation function $F_{\varphi}(.)$.[12] This one place function takes as arguments (codes of) finite valuations (variable assignments) in M and produces as values formulas (possibly nonstandard) of the language of PA in such a way that the following condition is satisfied:

(*) $\quad (M, Tr) \models \varphi[m]$ iff $(M, Tr) \models Tr(F_\varphi(m))$.

The idea of constructing F_φ is as follows: just substitute numerals for free variables occurring in φ in a way required by the valuation m (so e.g. if φ is a formula $\ulcorner v_5 + v_3 = v_8 \urcorner$, the translation function for a valuation m will produce a formula $\ulcorner m_5 + m_3 = m_8 \urcorner$, with numerals for the appropriate objects belonging to the sequence m). A special treatment will be needed though for a case of a bounded quantifier - we translate a formula with such a quantifier into a conjunction, whose length in a given model may be nonstandard. The inductive definition of F_φ proceeds as follows:

- $F_{t_1=t_2}(m) = \ulcorner sub(t_1, m) = sub(t_2, m) \urcorner$
- $F_{Tr(t)} = \begin{cases} val(t, m) & \text{if } val(t,m) \text{ is an arithmetical sentence} \\ \ulcorner 0 \neq 0 \urcorner & \text{otherwise} \end{cases}$
- $F_{\neg\varphi}(m) = \ulcorner \neg F_\varphi(m) \urcorner$
- $F_{\varphi \wedge \psi}(m) = \ulcorner F_\varphi(m) \wedge F_\psi(m) \urcorner$
- $F_{\forall v_i < v_j \varphi}(m) = \bigwedge_{a < m_j} F_\varphi(m\frac{a}{m_i})$

We check (*) for the bounded quantifier case, leaving other cases for a reader to verify. We claim that the following conditions are equivalent:

1. $(M, Tr) \models \forall v_i < v_j \varphi[m]$,
2. $\forall a <_M m_j (M, Tr) \models \varphi[m\frac{a}{m_i}]$,
3. $\forall a <_M m_j (M, Tr) \models Tr(F_\varphi(m\frac{a}{m_i}))$,
4. $(M, Tr) \models Tr(\bigwedge_{a<m_j} F_\varphi(m\frac{a}{m_i}))$,
5. $(M, Tr) \models Tr(F_{\forall v_i < v_j \varphi}(m))$.

The first equivalence is obvious, the second holds by the inductive assumption, the last one by the definition of F for the case of a bounded quantifier. The crucial step comes with the third of these equivalences: in order to obtain it we use the assumption that propositional logic preserves truth. It is this assumption which permits us to move freely between "all members of a given conjunction are true" and "the conjunction itself is true", where the conjunction in question is of arbitrary length.

With (*) at hand, we are ready to show that our model (M, Tr) satisfies Δ_0 induction. Let $\varphi(x)$ be a Δ_0 formula of the extended language and let's assume that $(M, Tr) \models \exists x \varphi(x)$. We will argue that in such a case there is the smallest object in (M, Tr) satisfying $\varphi(x)$ - this amounts to proving the least number principle, equivalent with induction. So fix a number a such that $(M, Tr) \models \varphi(a)$. By (*) we obtain: $(M, Tr) \models Tr(F_\varphi(a))$. Our next observation is that in such a case:

$$(M, Tr) \models Tr(\bigvee_{b \leq a}(F_\varphi(b) \wedge \bigwedge_{c<b} \neg F_\varphi(c))).$$

I will explain the idea behind the above step. What we use here is the principle which could be dubbed "propositional minimalization": take any finite sequence of sentences $(p_1 ... p_n)$, then if some sentence in this sequence is true, then there is the first sentence in this sequence, which is true. What is crucial here is that this principle can be "translated" into propositional tautologies. Assume for example that p_n holds. Then by propositional logic we obtain the consequence:

$$p_1 \vee (p_2 \wedge \neg p_1) \vee (p_3 \wedge \neg p_2 \wedge \neg p_1) \vee ... \vee (p_n \wedge \neg p_{n-1} \wedge ... \wedge \neg p_1)$$

For $k \leq n$, the kth disjunct of the above formula can be read as stating: "p_k is the first true sentence in the relevant sequence". By applying this to our case, we find out that the implication:

$$F_\varphi(a) \Rightarrow \bigvee_{b \leq a}(F_\varphi(b) \wedge \bigwedge_{c<b} \neg F_\varphi(c)))$$

is a propositional tautology, so it is true (by our initial assumption), therefore since its antecedent is true, its subsequent is also true.

Now: since the above disjunction is true, there must be one particular disjunct which is true. This remark may sound obvious, but we ask the reader not to treat it too lightly. We have no right to it when working in $PA(S)^-$ without any additional assumptions (remember KKL's pathological example!). In our context however it is fully justified: closure of truth under propositional logic guarantees it (if the negation of every disjunct was true, the negation of the whole disjunction would follow from the set of true sentences in propositional logic). So pick a b such that $(M, Tr) \models F_\varphi(b) \wedge \bigwedge_{c<b} \neg F_\varphi(c))$. And now it is enough to translate back, using (*) again. We obtain: $(M, Tr) \models \varphi(b)$ and $(M, Tr) \models \forall v < b \neg \varphi(v)$. This ends the proof since it means that b is the smallest number satisfying φ.[13]

□

4 Perspectives for further work

As we saw, the outcome is that closure of truth under propositional logic is a strong condition, producing a nonconservative extension of PA. This however still leaves open some interesting possibilities. When discussing first order logic (with quantifiers), we observed that the following two conditions are equivalent:

- **closure condition:** Truth is closed under first order provability,
- **membership condition:** All first order tautologies are true.

But for propositional logic, all we have at the moment is the information about the closure condition: we know that it is strong indeed. How about the membership condition of the form "All propositional tautologies are true"? We saw that this generalization is not provable in $PA(S)^{-}$, but it might - just might - be that when added to $PA(S)^{-}$, it does not produce new arithmetical theorems.[14]

A still more general approach would involve also (some of) the logic of quantifiers. The guiding question is what natural and nontrivial closure conditions can be conservatively imposed on the set of true sentences. It seems to me that a promising candidate to consider is closure of truth under proofs of finite length. Take some complete, axiomatic logical system S and consider a theory obtained by supplementing $PA(S)^{-}$ with a set of axioms, containing for each natural number n a sentence of the form "For every ψ, if ψ is S-provable from true premises in n steps, then ψ is true". The theory obtained in this way will prove that applying finitely many logical manipulations to true sentences will produce true results. And the question would be: for which logical manipulations, i.e. for which logical systems, is the resulting theory a conservative extension of PA. (If e.g. S is a system containing all propositional tautologies as axioms, membership condition for propositional logic would follow trivially from a resulting theory.) From a philosophical point of view, it seems that such a theory could offer the deflationist a decent compromise between conservativeness and generalization, conceived of as a rationale for introducing the truth predicate into our language. The matter seems to well deserve some further investigation.[15]

Notes

[1] [16] and [14] are classical texts in this area of research. For a review of results on satisfaction classes, see [13].

[2] The classical paper [11] describes a satisfaction class for a relational structure - i.e. for a model of arithmetic formulated in relational language. A construction of a satisfaction class for a model of arithmetic with function symbols is presented in [9].

[3] In other, perhaps more conspicuous notation, the formula in question could be written as: "$\forall v Tr(sub(\varphi(x), name(v)))$". In this version the expressions "sub" and "$name$" are used as function symbols; they correspond however to formulas (provably functional in PA) "$z = sub(x, y)$" and "$y = name(x)$". The first of them represents in PA a recursive function of substitution, which for a term y and a formula x with one free variable, produces the result of substituting y for a free variable in x. The second represents a recursive function giving as value, for a number x, a numeral denoting x.

[4] See [11]. Cf. also [3] for a discussion of various types of pathologies in satisfaction classes.

[5] Horwich's book [8] is an exposition of deflationism as a philosophical standpoint. For the conservativeness debate, see [10] and [17].

[6] On truth as a device for expressing infinite conjunctions, see [6].

[7] An interesting possible choice of a weak base theory would be $I\Delta_0$ - arithmetic with induction for Δ_0 formulas only. Analyzing truth theories built over $I\Delta_0$ could possibly shed some light on familiar (and notorious) open problems concerning $I\Delta_0$ itself. On the other hand, ZFC could be a natural choice of a strong base theory; the aim would be then to characterize the notion of truth for the language of set theory.

[8] Just to make it clear: I am not attributing such a conclusion to Halbach. This would be in fact a grave misinterpretation, out of tune with the main points he makes in the quoted paper.

[9] Obviously this makes sense only if we are able to characterize the class P with the linguistic means we have at our disposal within our theory.

[10] The problem does not arise if move to full $PA(S)$, because there we are able to prove that partial notions of truth indeed do coincide with the general one for appropriate classes of formulae.

[11] For example Field in [4] argues, that induction for formulas with arbitrary new predicates (including truth predicate) is arithmetical and not truth-theoretic; accordingly the deflationist may use it freely.

[12] Such translation functions were introduced by Kotlarski in [12], although the characterization I give differs in some respect from his definition (the main difference concerns the last clause - the quantifier case).

[13] In the above proof we reasoned with our function F_φ in a rather informal manner. In a more rigorous treatment one would have to observe that the function in question is recursive and work with an appropriate formula representing this function in PA.

[14] In my proof of Theorem 4 I used a stronger closure assumption when I moved from "a disjunction is true" to "some member of this disjunction is true". I do not know whether it is possible to justify such a move with merely a membership condition at hand.

[15] The relation of n-step provability received a lot of attention in logical literature. The main motivation for this research was the so called Kreisel's Conjecture. Given the set Ax of the usual axioms of PA and a logical system S, assume that there is a uniform upper bound on the lengths of proofs of $\varphi(n)$ from Ax in S (i.e. for every natural number n, there is a S-proof of $\varphi(n)$ from Ax in k steps, with k being fixed). Is it true then that the general statement "$\forall x \varphi(x)$" can be obtained in PA? The answer seems to depend on the choice of our proof system and the question is whether in our case we also encounter a similar phenomenon of system dependency. Fore more information on n-step provability, see [1].

References

[1] Buss, S. (1994). “On Gödel’s theorems on lengths of proofs I: number of lines and speedup for arithmetics”, *Journal of Symbolic Logic* 39, pp. 737–756.

[2] Cieśliński, C. (2010) “Truth, conservativeness, and provability”, *Mind* 119, pp. 409-422.

[3] Engström, F. (2002). *Satisfaction classes in nonstandard models of first order arithmetic*, Chalmers University of Technology and Göteborg University.

[4] Field, H. (1999). “Deflating the conservativeness argument”, *Journal of Philosophy* 96, pp. 533–540.

[5] Hajek, P. & Pudlak, P. (1993). *Metamathematics of first order arithmetic*, Springer Verlag.

[6] Halbach, V. (1999). “Disquotationalism and infinite conjunctions”, *Mind* 108, pp. 1–22.

[7] Halbach, V. (2001). “How innocent is deflationism?”, *Synthese* 126, pp. 167–194.

[8] Horwich, P. (1990). *Truth*, Basil Blackwell.

[9] Kaye, R. (1991). *Models of Peano arithmetic*, Oxford: Clarendon Press.

[10] Ketland, J. (1999). “‘Deflationism and Tarski’s paradise”, *Mind* 108, pp. 69–94.

[11] Kotlarski, H. Krajewski, S. & Lachlan, A. H. (1981). “Construction of satisfaction classes for nonstandard models”, *Canadian Mathematical Bulletin* 24, pp. 283–293.

[12] Kotlarski, H. (1986). “Bounded induction and satisfaction classes”, *Zeitschrift für Mathematische Logik* 32, pp. 531–544.

[13] Kotlarski, H. (1991). “Full satisfaction classes: a survey”, *Notre Dame Journal of Formal Logic* 32, pp. 573–579.

[14] Krajewski, S. (1976). “Non-standard satisfaction classes”, in *Set theory and hierarchy theory (Proc. Second Conf. Bierutowice, 1975)*, Lecture Notes in Mathematics, vol. 537, pp. 121–144.

[15] McGee, V. (2006). "In praise of the free lunch: Why disquotationalists should embrace compositional semantics", in V. Hendricks, S. Pedersen and T. Bollander (eds.) *Self-Reference*, CSLI, Stanford, pp. 95–120.

[16] Robinson, A. (1963). "On languages which are based on non-standard arithmetic", *Nagoya Mathematical Journal* 22, pp. 83–117.

[17] Shapiro, S. (1998). "Proof and truth - through thick and thin", *Journal of Philosophy* 95, pp. 493–552.