CHAPTER 11

Problems of Deflationism

PANU RAATIKAINEN

1. Introduction

The larger part of the more recent philosophical discussion on truth has focused on the conflict between 'deflationist' and substantial views of truth. However, it is not always altogether clear exactly what the essence of deflationism is—what theses are constitutive for it. Accordingly, one may distinguish at least four different issues that have been presented, together or separately, as the defining characteristics of deflationism (cf. Halbach 2001):

(1) Truth is not a property, or, at least, not a genuine or substantial property.
(2) T-sentences govern the meaning of the truth predicate and thus they are analytic or necessary.
(3) Truth is only a device of disquotation. All there is to say about truth follows from T-sentences; all facts involving truth can be explained on the basis of T-sentences alone.
(4) The notion of truth is only needed for certain logical purposes; truth is only a device of generalization, or semantic ascent, or it serves the purpose of expressing infinite conjunctions and disjunctions.

By 'T-sentences' one obviously means the equivalences, made famous by Tarski, of the form:

(T) 'p' is true if and only if p.

The familiar example is " 'Snow is white' is true if and only if snow is white".

Each of the above theses has been proposed as the essence of deflationism. Now, it is possible to hold the claims (1)–(4) separately or in different combinations. Different combinations and different interpretations may lead to a different understanding of what deflationism is. No wonder there is so much confusion about the true content of deflationism. Let us first briefly elaborate each of these theses one at the time.

'Truth is not a genuine property'. The claim (1) is the common starting point of various different deflationist views. However, although familiar

in the literature, it is in itself too vague to serve as the subject of interesting philosophical argumentation. I must say that I, for one, find it difficult to see exactly what is claimed. So does, for example, Hartry Field, a leading deflationist: "I am not clear enough as to what that is supposed to mean" (Field 1994, fn19). Paul Horwich (1998, p. 2), another influential deflationist, in turn admits that 'is true' is a property, but adds that it is not an ordinary sort of property which can be expected to participate in some deep theory of that to which it refers.

Without some further clarification, it is quite difficult to say anything intelligible for or against (1). However, claims (3) and (4), for example, can be considered, separately or together, as a possible explication of (1). Indeed, one popular way to explicate this vague idea seems to be to submit that all there is to say about truth is provided by T-sentences, which one takes to be true solely by virtue of the meaning of 'true', that is, analytic (as in (2)).

As a development of this idea, one possible, relatively exact way to draw the line is to contrast the full Tarskian theory of truth, with its inductive clauses, with the mere addition of T-sentences to the base theory, and understand the former as a substantial and the latter as a deflationary theory. Field, for example, contrasts deflationism with Tarskian approaches. It must be added, however, that just how substantial the Tarskian theory is, remains controversial. (Some take it as a version of the correspondence theory, whereas others think that it is essentially a deflationist view.)

Another, somewhat different possible explication of the idea that truth is not a genuine or substantial property is to require that additions of the truth predicate and axioms governing it, should not imply any new truths on issues not concerning the notion of truth, or, put more exactly, to require that the extension by the notion of truth should be *conservative* over the base theory. This idea also harmonizes nicely with disquotationalism, or T-schema deflationism (see below), for under certain quite general conditions, the addition of T-sentences to a base theory indeed results in a conservative extension. The full Tarskian theory (with the induction scheme extended to apply to the language containing the truth predicate), in contrast, is non-conservative.

'T-sentences are analytic, or necessary'. It is possible to understand the thesis (2) as a consequence of deflationism (and in particular, it is certainly a consequence of (3)), or as the defining thesis of deflationism. However, it seems that it cannot define deflationism, for several opponents of deflationism have also endorsed this thesis (Gupta, 1978; Gupta & Belnap, 1993; Milne, 1997; Raatikainen, 2003). An opponent should not go as far as to admit that P and "'P' is true" are synonymous. However, it seems to be possible to subscribe to the idea that T-sentences are analytic or necessary without committing oneself to this stronger view.

In any case, the view that T-sentences are analytic or necessary is at all sensible to hold only if one assumes a strict Tarskian distinction between the object language and the metalanguage, that is, if the truth predicate is not allowed to apply to the sentences which contain the truth predicate. Otherwise, not only are T-sentences not all analytically true, they are not all even true. Further, identifying the true ones is quite impossible, for the set of true T-sentences is then not recursive or recursively enumerable, and not even arithmetical (see McGee 1992). We will return to this issue in much more detail below.

'Truth is disquotation'. This slogan too allows different weaker and stronger interpretations. At one extreme one has the redundancy theory, according to which the predicate 'is true' can always be eliminated. As we will see, it has long been known that this view is untenable. However, it is possible, without committing oneself to this radical view, to hold that T-sentences *entail* all there is to say about truth, or that all facts involving truth can be explained solely on the basis of T-sentences. Let us call this latter view T-schema deflationism, or disquotationalism.

'Truth is a logical notion'. Again, this thesis in itself is not totally clear. To what is it opposed? To a substantial theory? But the latter reply shares the unclarity of (1). Often this is interpreted to mean that truth is only a device of generalization, or semantic ascent, or its serves the purpose of expressing infinite conjunctions and disjunctions. Even after this addition, much unclearness remains. For example, is the Tarskian theory of truth a version of the correspondence theory, and more generally, of a substantial theory of truth? Or is it rather a logical device of the sort intented here? Can it (or any other theory) be both? It would be clearer if one combined this claim with disquotationalism, or with the requirement of conservativity.

Indeed, it seems to be usual among deflationists to further assume that T-sentences are all that is needed for these logical purposes (e.g. Horwich 1998; Field 1994). Thus Field (1994) says: "Deflationism is the view that truth is at bottom disquotational"; and according to Horwich, minimalism, that is, the theory that consists of T-sentences (or, rather, in their propositional counterparts) can explain all the facts about truth (see also below). It remains to be seen whether that is really possible. But be that as it may, the combination of (3) and (4) appears to be, at least in practice, the most usual understanding of what deflationism amounts to. In what follows, I shall mostly assume this interpretation of deflationism.

2. From redundancy to deflationism

It is nowadays generally agreed—pace the traditional redundancy theory of truth—that indirect ascriptions of truth and generalizations involving truth cannot be eliminated; that was indeed the stumbling block of traditional redundancy theory.[1] By indirect ascriptions of truth, one means uses

of truth such as 'What Frank said yesterday is true', 'Fermat's last theorem is true', or 'That is true'. By generalizations involving truth, uses such as the following are intended: 'Every proposition that is entailed by a true proposition is true', 'All theorems of the theory T are true', 'There are arithmetical sentences that are true but not provable in T', et cetera. Or, to give a less logical example, 'Everything the Pope says is true'. In general, the sentences to which the truth is ascribed are not explicitly given in such cases.

It was noted early on that in such contexts, the truth predicate just cannot be eliminated with the help of T-sentences. Present-day deflationists grant this. Subsequent discussion has usually focused mainly on logical generalizations, but the indirect ascriptions can be handled, if at all, analogously with them, so this is not a significant omission. The contemporary deflationists add that these and related ('logical') contexts are the only contexts where the notion of truth is needed. It is thus important not to conflate present-day deflationism with the more traditional and more radical redundancy theory of truth, although the former is certainly a descendant of the latter.

Field (1986)[2] suggested that at least most of the uses of 'true' that have proved difficult to paraphrase away are uses where 'true' is serving as a surrogate for infinite conjunction or infinite disjunction. Thus consider, for example: 'There are true sentences which no one will ever have grounds to accept'. This use of 'true' can be, according to Field, naturally understood as a device of infinite disjunction:

p_1, but no one will ever have grounds to accept it; or
p_2, but no one will ever have grounds to accept it; or
p_3, but no one will ever have grounds to accept it; or ...

Instead of 'true', the equivalent idea could also be expressed by using a substitutional quantifier. A substitutional quantifier can also be used to define *a certain notion* of truth: according to it, 'x is true' is defined as the infinite conjunction of the sentences

If x is 'p_1', then p_1;
If x is 'p_2', then p_2; etc. ...

Field (1986), inspired by Quine (1970),[3] called this notion "a notion of disquotational truth". At the time, Field only suggested that even someone who accepts a correspondence notion of truth needs this notion. He then defined "a deflationary notion of truth" as the view which proposes that the latter serves no useful purpose at all, while at the same time preserves, contrary to the radical redundancy view, a use for the word 'true' (the disquotational use).

Few years later, Horwich then published his book-length defence of deflationism (Horwich 1990; cf. 1998). The main difference to the above definition is that he took propositions as the primary bearers of truth. But besides that, it generally agreed with Field's characterization. According to

Horwich, "the truth predicate exists solely for the sake of a certain logical need", namely, the problem of having a single, finite proposition that has the intuitive logical power of an infinite conjunction. The (unproblematic) instances of the schema

The proposition *that* p is true if and only if p

form what Horwich calls "the minimal theory of truth". Horwich adds that all facts involving truth can be explained on the basis of the minimal theory. By 1994, Field himself had become convinced of the correctness of deflationism as he had earlier defined it (see Field 1994).

Field (1986) reports that he got the term "deflationary conception of truth" from Horwich (1982). Horwich has later (1990) called his own particular brand of deflationism "minimalism". In any case, it is fair to consider these two thinkers as the founding fathers of the contemporary deflationism. I also think that this brief historical sketch justifies my choice of taking (3) and (4) together as a fair definition of contemporary deflationism.

3. Generalizations

Recall that a key motivation of contemporary deflationism was the observation that we need truth to make generalizations; deflationists often talk here about expressing infinitary conjunctions and disjunctions. This talk is in need of explication.

Perhaps it is only assumed that the truth predicate and T-sentences enable one to *express* infinite generalizations. Although such ideas have been stated repeatedly in the deflationist literature, somewhat surprisingly they were never made more exact before Halbach (1999). Halbach showed there that if one focuses on *definable* infinite sets of sentences, generalizations using the truth predicate and the corresponding infinite conjunction have exactly the same consequences. That is, let T^{Tr} be the theory obtained from the base theory T by adding T-sentences. Then, for example,

$$T + \mathrm{Prov}([A]) \to A \text{ for all } A \in L(T)$$

and

$$T^{Tr} + \forall x(\mathrm{Prov}([x]) \to \mathrm{Tr}([x]))$$

prove exactly the same formulae of $L(T)$. So with the help of the truth predicate, we can expess what would otherwise require infinitely many instances of the schema $\mathrm{Prov}([A]) \to A$. And similary for any definable infinite conjunction of the sentences of $L(T)$. So in this exact sense, it is indeed true that the disquotational notion of truth enables one to *express* (definable) infinite conjunctions.

However, the question of how we justify, or come to know, such generalizations involving the notion of truth still remains. For, as was noted above, often the claim is that T-sentences *entail* all there is to say about truth, or that all facts involving truth can be explained solely on the basis of

T-sentences. And it is this question that poses one of the hardest problems for deflationism.

As Tarski (1935) already pointed out, mere T-equivalences are insufficient to entail any general facts about truth, such as

Every sentence of the form '$p \rightarrow p$' is true.

T-sentences only imply every particular instance of such generalisations, but not the generalizations (see also Ketland 1999). Moreover, both Horwich (1998) and Field (1999), for example, admit that one should be able to prove such generalizations.[4] This is a serious problem for deflationism.

Horwich (1998) has proposed a way out. He suggests that we use of a rule of inference which allows the step from the provability of all instances of the generalization to the generalization. However, this amounts to the ω-rule and highly infinitary logic. Field (1994) also seems to suggest something similar. But there are several problems with such a solution.[5]

One problem is that the ω-rule is intimately connected with the substitutional interpretation of quantifiers. Field (1994) also seems to lean more directly on substitutional quantification. The substitutional quantification, however, is usually explained with the help of the notion of truth. Thus there is a serious threat of begging the question here (cf. David 1996). Moreover, logic equipped with the ω-rule is so powerful that it can, not only define, but even decide every arithmetical truth. By Tarski's undefinability theorem, the rule itself is not even arithmetically definable. It is somewhat preposterious to escape to such highly non-axiomatizable logic, which in a sense already has truth built into it, to save the deflationist view of truth. And if one is prepared to use infinitary logic, it would then be only a short step to allow infinite conjunctions and disjunctions directly. But avoiding them was a main starting point of deflationism.

4. Conservativity

One way to interpret deflationism is to require that the addition of the notion of truth should be a conservative extension of the base theory. The idea is that if truth is not substantial, then adding the truth predicate and the axioms essential to truth should not imply any truths not related to truth. Indeed, Shapiro (1998, 2002) and Ketland (1999) have argued that the conservativity requirement is essential for deflationism. (See Shapiro 1998 for a good discussion of the motivations of requiring conservativity.) T-sentences (restricted to the base language) are conservative over the base theory. Disquotationalism thus satisfies the conservativity requirement.

Conservativity has, however, its price. Let T be a theory that satisfies the assumptions of Gödel's incompleteness theorem, and T^{Tr} be an arbitrary theory which contains the truth predicate for $L(T)$. Shapiro (1998) has

pointed out that the following three claims are jointly inconsistent with Gödel's theorem:

(I) All talk, in T^{Tr}, of truth (of sentences of T) conservatively extends T.

(II) T^{Tr} enables one to assert that all theorems of T are true.

(III) T^{Tr} contains T-equivalences of all sentences of T.

You just cannot have them all. (Ketland (1999) made a similar observation.) If your theory of truth is conservative, you cannot have generalizations such as (II); if you are able to prove such generalizations, your theory is not conservative.

Tennant (2002) has proposed, in opposition to Shapiro and Ketland, that it is possible to argue for the truth of the Gödel sentence G without appealing to a substantial notion of truth. He suggests that one can instead use the so-called reflection principles, that is, formulas of the form

$$\mathrm{Prov}([\varphi]) \to \varphi.$$

Now the Gödel sentence certainly follows from this scheme. However, the real issue concerns how the reflection principles (or, 'T is consistent') are to be justified. Why should anyone accept the reflection scheme? Arguably one needs to lean on considerations involving a substantial notion of truth to justify it. Moreover, the addition of reflection principles anyway leads to a non-conservative extension.

5. Paradoxes

Deflationists have had strikingly little to say about paradoxes of truth and their solution. The general attitude seems to be that paradoxes are something that any theory of truth faces, and therefore deflationism need not specifically address them. Simmons (1999) and Glanzberg (2003) have, however, presented arguments concluding that the situation is not the same for everyone, but that deflationism has special difficulties in solving the paradoxes of truth.

One thing is certain: one cannot accept all T-sentences, if we allow that the truth-predicate can be iterated, that is, that p, in the T-sentence $\mathrm{Tr}([p]) \leftrightarrow p$, can itself contain the truth-predicate $\mathrm{Tr}(x)$. For this theory leads to the Liar paradox, and is thus inconsistent.

Tarski's classical solution was to draw a strict distinction between the object language and the metalanguage, and ban the iteration of the truth-predicate; the object language is a fragment of language which does not contain the truth-predicate. Indeed, in the above considerations, I have mostly assumed this setting for simplicity.

Both Field and Horwich, however, think that the Tarskian solution is too restrictive. Horwich suggest that we give up 'problematic' T-sentences. But how are we to recognize them? There is no effective method for recognizing

the 'problematic' instances of T-schema or even generating (i.e., recursively enumerating) them. And even worse, there is no unique set of unproblematic instances: McGee (1992) has showed that there are in fact uncountably many maximally consistent sets of T-sentences. Many of their 'unproblematic' T-sentences entail falsities (see also below). Therefore, under closer scrutiny, it might be wiser for a deflationist to stick to the Tarskian view and not iterate the truth predicate. However, both Field and Horwich seem to suggest that this would not be coherent with the general aims of deflationism.

6. Whither bivalence?

Under certain natural assumptions, T-schema deflationism implies bivalence. Horwich, for example, explicitly commits himself to this conclusion. The argument is simple: Let B be any sentence (not necessarily paradoxical) which is neither true nor false. Then "'B' is true" is false. But it then follows that the equivalence "'B' is true $\leftrightarrow$ B" cannot hold, for the left-hand side is false but the right-hand side is not.

Now there are logically possible ways to avoid this conclusion (see e.g. Holton 2000, Beall 2002).[6] But they are certainly not that attractive. They involve many-valued logics, deviant connectives and several negations. One may wonder whether one is still talking about T-sentences here, or whether one only has something superficially resembling them but in fact quite different from them. After all, deflationism is usually understood as applying primarily to a language we understand. But here we should conclude that the language and its logic were quite different from what we had thought.

Horwich does not accept Tarskian restrictions on T-sentences. But if one is allowed to iterate the truth predicate, it is difficult indeed to see how one could still stick to bivalence without qualifications. Opening that can of worms makes it possible to formulate the paradoxical Liar sentence, and it is then quite impossible to hold that sentence as either true or false.

7. Stranger T-sentences

We noted above that the addition of T-sentences to a base theory leads to a conservative extension. This holds under the assumption that the truth-predicate is not allowed to be iterated. If this restriction is removed, conservativity is lost. Should the iteration of the truth predicate be allowed or not? Both Horwich and Field answer affirmatively. However, it seems that they underestimate the troubles that follow.

Assume now that the truth predicate may be iterated. There is a trick, due to Vann McGee (1992), which provides, for every sentence S, an instance of (T) that is materially equivalent to S. This is achieved as follows: Let L be the usual language of arithmetic, and let L_T be L extended by the truth predicate $Tr(x)$. Now let S be any sentence of L_T. Consider the open formula $Tr(x) \leftrightarrow S$. By applying the diagonalization lemma, we can find a

sentence A so that PA proves $A \leftrightarrow (\mathrm{Tr}([A]) \leftrightarrow S)$. But this is equivalent, in propositional logic, to $S \leftrightarrow (\mathrm{Tr}([A]) \leftrightarrow A)$.

Of course, if such T-sentences are included, the addition of certain T-sentences to the base theory may not lead to a conservative extension any more. Take, for example, PA as the base theory, and let S above be $\mathrm{Cons}(PA)$. The addition of the resulting T-sentence to PA obviously leads to a non-conservative extension. And if conservativity is then agreed to be essential for deflationism, such T-sentences cause trouble. Note also that not all T-sentences are analytically or necessarily true any more—some are clearly not even true: take S above to be, for example, $\neg\mathrm{Cons}(PA)$.

More generally, for any, however strong axiomatizable theory F that extends PA, one can construct a recursively enumerable set T^* of T-sentences so that $PA + T^*$ has the same arithmetical theorems as F. We may take F to be, for example, ZFC. In particular, because we can take S to be a sentence which contains the truth-predicate, we can take it to be, for example

$$(\forall x)(\mathrm{Prov}(x) \rightarrow \mathrm{Tr}(x)).$$

That is, such self-referential T-sentences can also entail all sorts of generalizations involving truth that were unprovable with ordinary T-sentences. The problem with such T-sentence extensions is, of course, that they are completely ad hoc and parasitic to the theory F to which McGee's trick is applied. One could justify anything, also all sorts of falsities, with such self-referential T-sentences. It would be certainly very interesting if an intrinsic natural condition on T-sentences could be found which would lead to a non-conservative, sound extension merely by adding T-sentences. But at the moment, nothing of the sort is known, and the only examples of such non-conservative extensions are highly artificial. And in any case, if one takes conservativity to be essential for deflationism, then such a theory built from self-referential T-sentences just is not a deflationist theory. In sum, exotic self-referential T-sentences do not seem to much help deflationism.

University of Helsinki

Acknowledgements

My understanding of many issues dealt with here has benefited from e-mail correspondence with Volker Halbach and Jeff Ketland.

Notes

1. For the history and development of deflationist views, see Field (1986), Halbach (2001, Chapter 2).
2. At the time Field had not yet been converted to deflationism; nevertheless, Field (1986) is one of the most useful discussions of deflationism.
3. Although it is certainly true that Quine's remarks on truth considerably inspired contemporary deflationism, I think it is problematic to count Quine himself unqualifiedly as a deflationist.
4. Azzouni (1999), on the other hand, submits that a deflationist does not need to be able to achieve such generalizations. Clearly his understanding of deflationism is very different from the usual one.
5. I discuss them in more detail in Raatikainen (2005).
6. Also Field seems to be willing to follow this line.

References

Azzouni, Jody (1999). "Comments on Shapiro", *Journal of Philosophy* 96, 541–544.

Beall, J. C. (2002). "Deflationism and gaps: Untying 'not's in the debate", *Analysis* 62, 347–349.

David, Marian (1996). *Correspondence and Disquotation: An Essay on the Nature of Truth*, New York: Oxford University Press.

Field, Hartry (1986). "The deflationary conception of truth", in G. MacDonald and C. Wright (eds), *Fact, Science and Morality*, Oxford: Blackwell.

Field, Hartry (1994). "Deflationist views of meaning and content", *Mind* 103.

Field, Hartry (1999). "Deflating the conservativeness argument", *Journal of Philosophy* 96, 533–540.

Glanzberg, Michael (2003). "Minimalism and paradoxes", *Synthese* 135, 13–36.

Gupta, Anil (1978). "Modal logic and truth", *Journal of Philosophical Logic* 7, 441–472.

Gupta, Anil (1993). "A critique of deflationism", *Philosophical Topics* 21, 57–81.

Gupta, Anil and Neil Belnap (1993). *The Revision Theory of Truth*, Cambridge: MIT Press.

Halbach, Volker (1999). "Conservative theories of classical truth", *Studia Logica* 62, 353–370.

Halbach, Volker (2001). *Semantics and Deflationism*, unpublished habilitation thesis.

Holton, Richard (2000). "Minimalism and truth-value gaps", *Philosophical Studies* 97, 135–165.

Horwich, Paul (1982). "Three forms of realism", *Synthese* 52, 181–201.

Horwich, Paul (1990). *Truth*, Oxford: Basil Blackwell.

Horwich, Paul (1998). *Truth*, second edition, Oxford: Clarendon Press.

Ketland, Jeffrey (1999). "Deflationism and Tarski's paradise", *Mind* 108, 69–94.

McGee, Vann (1992). "Maximally consistent sets of instances of Tarski's schema (T)", *Journal of Philosophical Logic* 21, 235–41.

Milne, Peter (1997). "Tarski on truth and its definition", in Childers, Kolár and Svoboda (eds), *Logica'96: Proceedings of the 10th International Symposium*, Prague: Filosofia, 189–210.

Quine, W. V. (1970). *Philosophy of Logic*, Englewood Cliffs: Prentice Hall.

Raatikainen, Panu (2003). "More on Putnam and Tarski", *Synthese* 135, 37–47.

Raatikainen, Panu (2005). "On Horwich's way out", *Analysis*, forthcoming.

Shapiro, Stewart (1998). "Proof and truth: Through thick and thin", *Journal of Philosophy* 95, 493–521.

Shapiro, Stewart (2002). "Deflation and conservation", in V. Halbach and L. Horsten (eds), *Principles of Truth*, Frankfurt: Ontos Verlag, 103–128.

Simmons, Keith (1999). "Deflationary truth and the liar", *Journal of Philosophical Logic* 28, 455–488.

Tarski, Alfred (1935). "The concept of truth in formalized languages", in A. Tarski, *Logic, Semantics, Metamathematics*, second edition, J. Corcoran (ed.), Indianapolis: Hackett, 152–278.

Tennant, Neil (2002). "Deflationism and the Gödel phenomena", *Mind* 111, 551–582.