Facta Philosophica 1, 1999: 167-195
© Peter Lang, Switzerland

Deflationism and Logic

Christopher Gauker

1 Introduction

What we say and think depends on what happens around us. Many philosophers have hoped that this dependence would be characterizable in terms of truth. In striving to speak and think truths, we strive for words and thoughts whose structure corresponds to structure in the world. Unfortunately, no one has been able to explain in a fully persuasive way what the pertinent correspondence relation might be. In light of this, some philosophers have concluded that we ought to seek elsewhere for our explanation of the connection between thought and the world. As for truth, there is nothing more to say about it beyond what can be summed up in a few simple principles such as that *p is true if and only if p*. This is the message of *deflationists*, who aim to *deflate* our expectations of the concept of truth.

Besides being a locus of control between the world and our thought, the concept of truth has traditionally played a further role in philosophy. Truth is supposed to be that which logically valid inference preserves. Of course, an argument does not qualify as valid just because it happens to have a false premise or a true conclusion. Validity is supposed to abstract from actual truth value. According to the modern definition, an inference is logically valid if and only if for every interpretation of the language on which the premises are true, the conclusion is true as well. The question arises whether a deflationist can retain this conception of logical validity while rejecting the correspondence conception of truth.

I will argue that deflationism is incompatible with standard conceptions of logic in two ways. First, deflationism is incompatible with the classical rules of inference. This need not be a problem in itself but it highlights the need for the deflationist to give some definite theory of logical validity. The second, more serious problem is that deflationism is incompatible with the model theoretic definition of logical validity (classical or otherwise). My primary target will be *inference rule deflationism*. Subsequently, I will consider two further forms of deflationism, namely, *T-schema deflationism* and Horwich's *minimal theory*, and I will argue that they fare no better.[1]

2 Inference rule deflationism

Inference rule deflationism claims that the nature of truth can be explained by explaining something about the use of the word "true" (and its synonyms in other languages) and that what has to be understood about the word "true" is just that instances of the following two inference rules are logically valid (and that instances of the corresponding rules in other languages are valid):

Semantic Ascent:

$$\frac{p}{\text{“}p\text{” is true.}}$$

Semantic Descent:

$$\frac{\text{“}p\text{” is true.}}{p}$$

(Of course, the "*p*" in quotation marks is a schematic letter.) This much of the theory deals only with the truth of sentences in our own language. The theory might be extended to sentences of other languages by appealing to translation relations, but I will not consider how exactly this extension might be achieved. Alternatively, truth might be treated as first of all a property of propositions and Semantic Ascent might be defined as the inference from *p* to "The proposition that *p* is true" and Semantic Descent as the converse of this.

A virtue of inference rule deflationism is that we can recognize it as an explanation of meaning by means of an analogy with a certain common explanation of the meaning of logical operators. What does the logical operator "or" mean? One way to explain it is in terms of the truth conditions of sentences containing "or". An alternative is to identify the inference rules that govern the use of the word "or", formulated, perhaps, as Gentzen-style introduction and elimination rules. To say that the meaning of "true" is given by the rules of Semantic Ascent and Semantic Descent is to say the meaning of "true" can be explained in the way that the meaning of "or" is explained when we explain it in terms of inference rules. Such an explanation of the meaning of "or" does not overlook the meaningful use of "or" in sentences other than those to which the inference rules immediately apply. Although the meaning-giving inference rules apply only to sentences of the form "*p* or *q*", the word "or" can meaningfully occur in sentences such as "Every tree in the park is either an oak or a maple" inasmuch as other inference rules take us from (or to) that sentence to (or from) sentences of the form "*x* is an oak or *x* is a maple". Similarly, we can adopt the inference rule deflationist's account of the meaning of "true" and still allow that "true" can meaningfully occur in sentences such as "Everything John believes is true".

On this theory, the expression "is true" is treated as in some ways like a logical operator and in some ways like a predicate. It is a logical operator insofar as the explanation of its meaning is analogous to one sort of explanation of the meaning of logical operators such as "or". At the same time it is, or is like, a predicate, in that it can be used to form sentences such as "Everything John believes is true". I do not think that there is anything objectionable in the inference rule deflationist's attribution of this dual nature to the expression "is true". Compare the identity sign. Someone might claim that the meaning of "=" is given in the logical laws of identity formulated as rules of inference. Regardless of whether that is right or wrong, it would be too simple to object that the identity sign is a two-place predicate and not a logical operator.

I should emphasize that inference rule deflationism (as well as T-schema deflationism, to be examined below) proposes to explain the *nature of truth* by explaining the *meaning of "true"*. Other forms of deflationism, such as Horwich's minimal theory, may aim directly at an account of the nature of truth, but inference rule deflationism, as I am defining it, explicitly states that an account of the nature of truth may take the form of a theory of the meaning of "true" and its synonyms in other languages. One way to criticize inference rule deflationism might be to criticize its substitution of a question about a meaning for a question about a nature. However, my criticisms will concern only the proposed explication of meaning.

Inference rule deflationism is basically the form of deflationism defended (but not unreservedly endorsed) by Hartry Field (1994a, 1994b). Field defines deflationism as "the view that truth is at bottom disquotational" (1994b, p. 405). He explains that this means that in its primary use "true", as understood by a given person, applies only to utterances that that person understands and that for any utterance *u* that a person understands, the claim made by uttering "*u* is true" is cognitively equivalent for that person to the claim made by uttering *u* itself. Field intends his thesis to be not merely a psychological claim about what other people think but a theory of what truth really is; so we may add that the claim we make by uttering *u* and the claim we make by uttering "*u* is true" are, or ought to be, cognitively equivalent *for us*. Field adds that he intends this thesis to apply not so much to overt utterances but to "internal analogs" of them, which he calls "sentence-readings". He expects that that particular refinement will help the

[1] I am omitting Dorothy Grover's *prosentential theory* (1992) since it is very different in form from the others to be discussed here.

deflationist accommodate ambiguity and indexicals (1994a, pp. 278-281). Moreover, the cognitive equivalence between the claim made by uttering *u* and the claim made by uttering "*u* is true" is supposed to be relative to the existence of *u*, since the claim made by uttering "*u* is true" claims that *u* exists and the claim made by uttering *u* itself does not claim that *u* exists.

That Field's deflationism is essentially a form of inference rule deflationism becomes apparent when he explains what he means by cognitive equivalence. Field writes: "I take cognitive equivalence to be a matter of conceptual or computational role: for one sentence to be cognitively equivalent to another for a given person is for that person's inferential rules to license (or, license fairly directly) the inference from either one to the other" (1994b, p. 405, note 1). In this passage, Field takes the relation of cognitive equivalence to be a relation between sentences, whereas in defining deflationism in terms of cognitive equivalence, he took it to be a relation between *claims*. I will assume, although Field does not say this, that the relation between sentences is the primary one and that claims may be cognitively equivalent inasmuch as they are made by means of sentences that are cognitively equivalent. Thus, the central claim of Field's deflationism - subject to the qualifications I have noted - is that "true" means *true* because we should regard any utterance of a sentence *p* as licensing in the utterer an utterance of a sentence of the form "'*p*' is true", and conversely. Thus Field's theory is a form of what I am calling inference rule deflationism.[2]

I have defined inference rule deflationism as holding that Semantic Ascent and Semantic Descent are *valid*, and so when I interpret Field as an inference rule deflationist, he might wish to object that he did not mean to say that Semantic Ascent and Semantic Descent are logically *valid* in any technical sense. I take it, though, that these inferences must have the same normative status as, for instance, the rules of inference governing "or"; otherwise, I just do not know what the content of the theory is supposed to be. This analogy is all I intend in describing the rules of Semantic Ascent and Semantic Descent as valid. In saying that, I am not already presuming any technical sense for the word "valid". Moreover, I will not at any point take for granted any particular definition of logical validity. In section 4, however, I will argue that we must provide *some* kind of precise definition of logical validity.

3 Deflationism and classical rules of inference

Deflationists typically brush aside the semantic paradoxes as if they posed no serious problem. For instance, Field says only that as a person "comes to terms with the paradoxes he revises his standard of cognitive equivalence on conceptual grounds" (1994a, p. 251, note 2). He does not consider that the revision that is called for on the part of that person might be the rejection of the cognitive equivalence claim altogether and that the revision that is called for on the part of us theorists might be the rejection of deflationism. The serious problem, as I will show, is that Semantic Ascent and Semantic Descent, together with the classical rules of inference, enable us to derive contradictions from plain facts. Consequently, inference rule deflationism is incompatible with the classical rules of inference. I do not think my argument in this section is in any way clever or original, but what I say here needs to be said since so many people seem to be overlooking the obvious.[3] My purpose here is just to demonstrate the need for the deflationist to address the question of the nature of logical validity before I go on, in the rest of the paper, to argue that the deflationist cannot blithely endorse the standard model theoretic conception of logical validity.

Consider the classic liar:

(α) α is not true.

Given that α = "α is not true", a plain fact, we seem to be able to derive a contradiction by the following reasoning:

> Suppose α is true. In that case (by Identity), "α is not true" is true. In that case (by Semantic Descent), α is not true. So (by Indirect Proof), α is not true. Suppose α is not true. In that case (by Identity), "α is not true" is not true. In that case (by the contraposed form of Semantic Ascent), it is not the case that α is not true. So (by Indirect Proof), α is true. Contradiction!

Or consider the following sentence, sometimes called Löb's paradox:

(β) If β is true, then $\perp$.

[2] In one place, Field gives a slightly stronger definition of cognitive equivalence: Two sentences are cognitively equivalent for a person if and only that person's inferential procedures license substitution of one for the other in all but quotational and intensional contexts (1994a, p. 251, note 2). I should acknowledge that I have somewhat twisted Fields words in order to simplify exposition. In both of the papers cited here, Field writes of an "utterance u", where "the claim that u is true" is cognitively equivalent to "u itself" (1994a, p. 251; 1994b, p. 405). Thus cognitive equivalence is treated as a relation between a claim and an utterance. I assume that what Field really means in these passages is that cognitive equivalence holds between two claims.

[3] My conclusion has been drawn before me, by Alan Weir (1996), who in turn relies heavily on a paper by McGee (1992), but I think I can bring out the nature of the difficulty a little more clearly.

(Here "$\bot$" is some contradiction.) Given that β = "If β is true, then $\bot$", another plain fact, we seem to be able to derive a contradiction by the following reasoning:

> Suppose β is true. In that case (by Identity), "If β is true, then $\bot$" is true. In that case (by Semantic Descent), if β is true, then $\bot$. In that case (by Modus Ponens), $\bot$. So (by Conditional Proof), if β is true, then $\bot$. So (by Semantic Ascent), "If β is true, then $\bot$" is true. So (by Identity), β is true. So (by Modus Ponens), $\bot$.

These arguments will be sound if the inference rules that justify the individual steps are all valid. But the inference rule deflationist accepts Semantic Ascent and Semantic Descent. So the inference rule deflationist must deny the validity of one of the other inference rules. But the others are all classical rules of inference. So the inference rule deflationist must deny the validity of some classical rules of inference. (The concept of validity applies primarily to inferences from sets of sentences to sentences. But in a secondary sense we can apply the concept of logical validity to rules like Indirect Proof and Conditional Proof, which merely tell us that if a certain inference from one set of sentences to a sentence is valid in the primary sense then some other inference from a set of sentences to a sentence must be valid too.)

The most dubious candidates are probably Conditional Proof and Indirect Proof. Suppose we have a three valued semantics and find that the classic liar is neither true nor false on any interpretation and that likewise the negation of the liar is neither true nor false on any interpretation. Vacuously then, for any sentence *p* and any interpretation, if the liar is true on that interpretation, then *p* is true on that interpretation. *p* might even be a contradiction or the negation of the liar. Yet this does not mean that the negation of the liar is true on every interpretation. Our assumption was that it is neither true nor false on any interpretation. So Indirect Proof will be invalid. Similarly, Conditional Proof will be invalid.

In arguing in this way, I have assumed that certain facts are plain, such as that α = "α is not true". However, these plain facts involve self-reference; for instance, α refers to itself. Someone might claim that the way to evade the paradoxes is to deny that such self-reference is possible. But in light of other paradoxical sets of sentences, this is not a very helpful suggestion. Consider, for example, the note card paradox. On side A is written, "Every sentence on side B is true", and on side B is written, "No sentence on side A is true". These sentences generate contradictions in much the same way as the classic liar, but they do not refer to themselves in any way that can generally be forbidden. (In this case, the pertinent plain facts are that the sentence on side A = "Every sentence on side B is true" and that the sentence on side B = "No sentence on side A is true".)

A deflationist might think that the problem can be solved simply by stipulating that only certain instances of Semantic Ascent and Semantic Descent qualify as valid, but this thought does not withstand scrutiny. Consider, for instance, the idea that the instances of Semantic Ascent and Semantic Descent that result from substituting some sentence for "*p*" in these rules are valid if and only if we cannot use them to derive a contradictory conclusion from "plain facts" (derive by classical means, that is). The problem is that we sometimes require several different instances to derive the contradictory conclusion. Consider again the note card paradox. The pertinent instances of Semantic Ascent and Semantic Descent are these:

(1) No sentence on side A is true.
"No sentence on side A is true" is true.

(2) "No sentence on side A is true" is true.
No sentence on side A is true.

(3) Every sentence on side B is true.
"Every sentence on side B is true" is true.

(4) "Every sentence on side B is true" is true.
Every sentence on side B is true.

Even given that the sole sentence on side A is "Every sentence on side B is true" and the sole sentence on side B is "No sentence on side A is true", no contradiction can be derived using just the inferences in (1) and (2). To derive a contradiction, we need (3) and (4) as well. So if we simply declared instances of Semantic Ascent and Semantic Descent to be valid whenever we could not use those instances on their own to derive a contradiction from plain facts, then we would find ourselves declaring a number of instances to be valid that we could use in conjunction with one another to derive a contradiction from plain facts.[4]

[4] McGee's 1992 critique of deflationism may be viewed as a generalization of the argument in this paragraph. If his critique were reformulated as a critique of inference rule deflationism, then the point would be that there is no uniquely acceptable maximal consistent set of instances of Semantic Descent and Semantic Ascent (where by "maximal consistent" set of instances, I mean a set that does not permit the derivation of contradictions from plain facts, while the addition of any further instances would permit this). The rest of my argument may be viewed as addressing the suggestion that the deflationist might settle for something less than such a maximal consistent set of instances.

In response to this, it might be suggested that an instance of Semantic Ascent or Semantic Descent should be declared invalid if it can be used to derive a contradiction from plain facts in conjunction with any other instances. The problem is that by this criterion no instance will qualify as valid whatsoever. Consider a version of the note card paradox in which side B contains in addition the sentence “The moon is the moon”. In this case, the apparently innocent inference from “The moon is the moon” to “‘The moon is the moon’ is true” will have a hand in the derivation of a contradiction from plain facts. Plainly it will not do to say that an instance is valid if and only if it does not *in fact* lead from truths to falsehoods. By that criterion, every inference whatsoever could qualify as valid provided only it had a true conclusion or at least one false premise.

Instead of trying to maintain that the valid instances of Semantic Ascent and Semantic Descent are just those that do not get us into a trouble, we might try to identify the valid instances independently of the trouble caused by the invalid ones. We might approach this either syntactically or semantically. The syntactic approach will say that Semantic Ascent and Semantic Descent are simply not articulate enough - that it is only when the sentences we substitute for the schematic letter “*p*” have certain syntactic characteristics that Semantic Ascent and Semantic Descent will take us from plain facts into contradictions. The problem is that any general prohibition against certain kinds of instances based wholly on syntactic structure is bound to be too strong. Suppose we declare that an instance of Semantic Ascent or Semantic Descent is not valid if the sentence we substitute for “*p*” contains semantic vocabulary such as “true”. That might indeed block all of the derivations of contradictions from plain facts that concern us here, but it would be too strong. We could no longer plausibly contend that all we have to understand about the meaning of the word “true” (in order to understand the nature of truth) is that it figures into such inferences. That was plausible only insofar as all other legitimate applications of the word “true” could be mediated by such inferences. The sentence “‘“The moon is the moon” ist true’ is true” is a legitimate application of the word “true”, but it cannot be reached from “The moon is the moon” without applying Semantic Ascent to “‘The moon is the moon’ is true”, which is a sentence that contains the word “true”.

The semantic approach would be to try to identify the valid instances of Semantic Ascent and Semantic Descent semantically. For instance, we might say that an instance of one of these rules is valid if and only if for any interpretation of the language on which the premise is true the conclusion is true as well. Such a definition will not validate Semantic Ascent and Semantic Descent if an interpretation can assign an arbitrary extension to the predicate “is true”, however. So before we could offer such an account of the validity of these rules, we would have to resolve some difficult issues concerning truth, such as whether sentences such as the liar are true, false, or neither. Moreover, if in order to identify the valid instances of Semantic Ascent and Semantic Descent we have to appeal to truth theoretic semantics in this way, then the deflationary explanation of truth in terms of Semantic Ascent and Semantic Descent will rest on a prior conception of truth on an interpretation that cannot itself be explained in terms of Semantic Ascent and Semantic Descent. The notion of truth on an interpretation is not yet the notion of truth simpliciter, but we should not expect to understand the former without an understanding of the latter. So on this approach the deflationist's theory of truth will beg the question.

We should not assume that semantic definitions have to be formulated in terms of truth on an interpretation. Perhaps the valid instances of Semantic Ascent and Semantic Descent could be identified without any question-begging reliance on the concept of truth. We might say that an instance is valid if and only if it never leads from a premise that is acceptable in some other sense to a conclusion that is not acceptable in that sense. But this begs the question in different way. If we could explain in a general way when the application of the word “true” to a sentence or utterance were acceptable in the pertinent sense, then we could take that explanation as our theory of truth and we would not need inference rule deflationism.

I conclude that the inference rule deflationist has no alternative but to maintain that Semantic Ascent and Semantic Descent are valid without qualification and then, to evade the paradoxes, to deny the validity of some of the other inferences on which the paradoxes depend. That means that the inference rule deflationist must abandon some classical rules of inference. One should not imagine that “technical” approaches to the semantic paradoxes will get around this conclusion somehow. Serious attempts to explain how a language can consistently contain its own semantic vocabulary (or at least “true”) fall into two categories. Some do indeed abandon classical rules of inference (for example, Kripke 1975, and Barwise and Etchemendy 1987). Others preserve the classical rules of inference but deny the validity of Semantic Ascent and Semantic Descent (for example, Gupta and Belnap 1993).

I should stress that I have not argued that Semantic Ascent and Semantic Descent are incompatible with classical *logic* but only that they are incompatible with the classical *rules of inference*. One might maintain that Semantic Ascent and Semantic Descent are valid while maintaining that every classically valid inference from a set of sentences A to a conclusion q is in fact valid. A counterexample to, say, Indirect Proof, will show that for some set of sentences

A and some sentence *p*, while the union of *A* and {*p*} implies a contradiction, *A* does not imply the negation of *p*. This will not show that the validity of any classically valid argument has to be denied unless the inference from the union of *A* and {*p*} to a contradiction is classically valid. Where the semantic paradoxes are used to produce counterexamples to Indirect Proof, that will not be the case, since the inference from the union of *A* and {*p*} to a contradiction will employ Semantic Ascent or Semantic Descent.

I do not take the result of this section to be in itself an objection to inference rule deflationism, for I do not take the classical rules of inference to be sacrosanct. However, it shows that the deflationism must admit as a serious question, "What are the valid rules of inference?" For present purposes, it is not necessary for me to survey the possibilities or select a best option. The point has only been to challenge the deflationist to decide on a conception of logical validity and then to make sure that the conception selected is one that is compatible with deflationism.

4 Why logic needs semantics

Suppose that the inference rule deflationist accepts that Semantic Ascent and Semantic Descent are in general valid (and thus rejects some classical rules of inference) or finds some nonsemantic means of identifying the valid instances (which I have claimed is impossible). Still, it might be objected that any viable definition of logical validity will appeal to the concept of truth in a manner incompatible with deflationism. One response to this objection might be to deny that we need any semantic definition of logical validity at all. Semantics cannot justify our rules of inference, it might be said, because the semantics can be doubted as readily as the inference rules. The validity of our inference rules, it might be argued, is a matter of social practices, or of how people tend to reason, or is an a priori consequence of the meanings of the logical constants (and words like "true"). Against this response, I now want to explain why, though semantics does not justify our inference rules and validity may indeed be a matter of such things, we require a semantic definition of logical validity nonetheless.

Suppose we try to use our axioms and inference rules to derive a certain conclusion from certain premises and are unable to do so. Our failure to produce a derivation is no proof that no derivation exists. To prove it, we must employ a semantic definition of logical validity. For instance, on the usual model theoretic sort of definition, we will be able to demonstrate that an argument is logically invalid by constructing an interpretation of the language such that on that interpretation all of the premises are true and the conclusion is false. The reason we need a semantic definition of logical validity is to have a means of demonstrating that underivable arguments are indeed underivable.

More precisely, we need a semantic definition of logical validity relative to which our system of axioms and inference rules is sound and complete. Our axioms and inference rules have to be sound with respect to it so that we can be sure that any argument that is not valid according to the definition is not derivable by means of the axioms and rules. Our system has to be complete with respect to the definition of validity so that we can be sure that whenever an argument is underivable we can show that it is so using the definition. This is not to say that derivability must be decidable (since there may be no algorithm for constructing a countermodel even when we know there is one). It is also not to say that our semantic definition of logical validity must be model theoretic, only that there must be some kind of semantic definition.

In reply it might be said that, having constructed a method of demonstrating, for any underivable argument, that it is underivable, it is not necessary to interpret that method as talking about truth or about anything else. It may be just a meaningless technique for demonstrating underivability. To see that this is wrong, suppose we have a set of axioms and inference rules that long experience tells us will never lead us into contradiction and will never fall short of our inferential needs. Then any definition of logical validity with respect to which these axioms and inference rules are sound and complete will have to be regarded as defining a valuable property that all and only derivable arguments possess. Otherwise, the fact of soundness and completeness with respect to it would have to be regarded as an absurd miracle.

5 Logical validity and reference

The model theoretic definition of logical validity for a first order language begins with the definition of an *interpretation* of the language. (Interpretations are sometimes called *models*. I call them *interpretations* because the expression "model of a theory" is sometimes used to describe specifically those interpretations on which the sentences in a given set are true.) An interpretation includes, minimally, a universe and an assignment. An assignment is a function that takes each individual constant (or each member of a given subset of the individual constants), each predicate and each function symbol into an *extension*. Typically, the extension for an individual constant is a member of the universe, the extension for an *n*-ary predicate is a set of *n*-tuples of members of the universe, and the extension for an *n*-ary function symbol is an *n*-ary function from *n*-tuples of members of the universe into members of the universe. An interpretation may also assign to each predicate an antiextension.

If the language contains operators for modalities or tenses, the interpretation may also contain sets of parameters such as possible worlds or times and the assignment will be relativized to parameters so that a given constant or predicate may be assigned different extensions at different parameters. If the language contains context-relative elements, then the assignment might be relativized to contexts as well. The devices by which linguists attempt to construct model theoretic interpretations of natural languages are a little different, but for simplicity I will keep up the usual fiction that the languages we have to account for are just those having the grammar of the languages of formal logic.

The next step is to explain how an interpretation of the language determines a truth value for each of the sentences of the language. This determination will be based on the assignment and universe somehow and will employ some kind of recursion whose clauses correspond to the logical operators in the language. (More may be involved as well, such as the identification of fixed point interpretations of semantic vocabulary, as in Kripke 1975.) Typically, an argument is said to be logically valid if and only if, for every interpretation of the language (and every parameter), if every premise is true on that interpretation (at that parameter), then the conclusion is true on that interpretation (at that parameter) as well. (For certain many-valued systems, substitute "has a designated value" for "is true".) Such a definition need not validate all the classical rules of inference; so there is room to hope that within this framework we might evade the semantic paradoxes while preserving the validity of Semantic Ascent and Semantic Descent.

If validity is defined in terms of truth on an interpretation, then inevitably we will suppose that there is one special interpretation of the language, call it the *intended interpretation*, such that truth on that interpretation is truth simpliciter in the language. This will be an interpretation that assigns to each individual constant and predicate the extension that it really refers to. The intended interpretation for a language is the reference relation restricted to that language. (Here of course I am speaking of reference as a relation between each well-formed expression of the language and an appropriate value. The reference of a name will be an individual. The reference of an *n*-ary predicate, whether simple or compound, will be a set of *n*-tuples or perhaps an *n*-ary property. I use the term "intended interpretation" only in deference to tradition. I do not wish to suggest that people's intentions have anything to do with the identity of the intended interpretation.) The reason this is inevitable is that the definition of logical validity in terms of truth on an interpretation is supposed to capture the fact that a valid argument guarantees the truth of the conclusion given the truth of the premises without regard for the reference of the nonlogical words, and the definition of validity in terms of truth on an interpretation can be regarded as abstracting from the reference of the nonlogical words only if one of these interpretations is the one that gives the referents that the nonlogical words really have.[5]

Can an inference rule deflationist accept this identification of truth simpliciter for a language with truth on the intended interpretation? There is no immediate circularity in espousing inference rule deflationism while endorsing the model theoretic definition of validity. At least, there is no immediate circularity if the valid instances of Semantic Ascent and Semantic Descent can be identified without appeal to the semantic definition of logical validity (as they will be if we simply say that they are all valid and deny the validity of some of the classical rules). On the other hand, the characterization of truth offered by the inference rule deflationist and the characterization of truth as truth on the intended interpretation do not seem to be in any sense equivalent. Certainly we cannot easily reconcile the thesis of Field's "Tarski's Theory of Truth" (1972) with both deflationism and the characterization of truth as truth on the intended interpretation. It was Field who, in that paper, first persuaded many people that if we accept such a characterization of truth for a language, then we ought to explain the nature of the reference relation in a naturalistically acceptable way. Nonetheless, if there really is an incompatibility between deflationism and the characterization of truth as truth on the intended interpretation, then that incompatibility ought to be demonstrable in some way. What follows, in the next two sections, is my attempt to demonstrate it.[6]

6 Specifying the intended interpretation

If there is an intended interpretation, grounded in the reference relation, then presumably it must be possible to distinguish it from the other interpretations, and thus to distinguish the reference relation, as it pertains to the language in question, from the assignments that figure in the other interpretations. This is not to say that it should be possible to say explicitly what each term refers

[5] A doubt about this conclusion may be raised that is independent of the issue about deflationism. For many important languages, it may be said, there cannot be an intended interpretation because we know that terms such as "set" and "ordinal number" cannot refer to sets. But in light of the argument just given I do not see how we can take this as evidence that we can define logical validity in terms of truth on a model and yet deny that there is any intended interpretation. It is, rather, an independent difficulty for the model theoretic conception of semantics.

[6] I have presented much briefer versions of this argument twice before, in my 1990 and my 1994.

to, but only that the reference relation, as it pertains to a given language, should be uniquely describable. It would not do for the deflationist, or anyone else, to maintain that *there is* a unique intended interpretation but nothing that we can say about it distinguishes it from infinitely many other interpretations of the language. Imagine I say, “There is a unique intended interpretation”. You ask, “Well, which interpretation is it?” Perhaps you could excuse an answer like this: “We don’t know yet; we need more funding”. But you should not excuse an answer like this: “It is impossible to answer that question”.

The point is not merely that we can generalize from the questions we might ask about particular terms. The deflationist will not reject as illegitimate a question on the order of “What does ‘rabbit’ refer to?”, even if the answer is never more than a syntactically definable transformation of the question, on the order of “‘Rabbit’ refers to rabbits”. Thus, it might seem that the deflationist should not reject as unanswerable a general question on the order of “For each referring term t of L, what is x such that t refers to x?” But this is not the basis for the obligation I am ascribing to the deflationist. If it were, then the deflationist might deny that obligation on the grounds that the attempt to generalize aims at something impossible.

The demand for a specification of the intended interpretation arises specifically from the commitment to its existence. The basis for that obligation is just the presumption that *there is* a definite interpretation that is the intended interpretation. If there is one, then for every other interpretation of the language, there must be something that distinguishes the intended interpretation from that other one. That much is simply a consequence of the fact that the intended interpretation is unique. But what is more (this is logically stronger), there must be something about the intended interpretation that distinguishes it from every other interpretation of the language. Given that there is something that distinguishes the intended interpretation from other possible interpretations, it should be possible, at least in principle, to say what it is that distinguishes it. This is not in itself a refutation of deflationism, however, because the account of what distinguishes the intended interpretation from other interpretations might itself be deflationary, as we will see.

To see how the intended interpretation of a language might be specified, let us distinguish between two kinds of theories of reference. One kind of theory would be a *substantive* theory of reference as it pertains to the language in question. This would be a theory having the form,

(R) t refers to e in L if and only if ... t ... e ... L ... ,

where the ellipsis is filled out in some suitably non-question-begging way. The other kind of theory would be a *deflationary* theory of reference. For instance, an inference rule deflationist might hold that all we have to explain about the reference relation as it pertains to a given language is that inference rules such as the following are valid:

Reference Rules:

$$\text{A. } \frac{\text{“}a\text{” refers to } b.}{a = b.} \qquad \text{B. } \frac{a = b.}{\text{“}a\text{” refers to } b.}$$

$$\text{C. } \frac{\text{“}F\text{” refers to } G\text{’s.}}{\text{All and only } F\text{’s are } G\text{’s.}} \qquad \text{D. } \frac{\text{All and only } F\text{’s are } G\text{’s.}}{\text{“}F\text{” refers to } G\text{’s.}}$$

(Actually, such rules would constitute at most a theory of reference for our own language; but I will assume that the theory could be extended to deal with the reference relation for other languages as well.) Clearly, the deflationist with respect to truth has to favor a deflationary theory of reference. If we had a substantive theory of reference for a given language, then we could give a substantive definition of truth for the language in terms of it. Understanding that truth for a language can be defined in that way would be essential to understanding the nature of truth. So deflationism would be wrong.

There really is a difference between these two approaches to reference, because the Reference Rules cannot be expressed in the form of what I have called a substantive theory of reference. If we try to express A and B in the form of (R), we might get something like this:

For all a and b, “a” refers to b if and only if $a = b$.

This is not well formed. The quantifier “For all a” cannot have both of the subsequent occurrences of “a” within its scope. If we introduce a naming function N that takes an object into its name (never mind that not every object has a name), we might write:

For all a and b, $N(a)$ refers to b if and only if $a = b$.

This is well formed, but now we have to identify the function N (as it pertains to the language) and this is not essentially different from having to identify the reference relation (as it pertains to the language).

The question whether an inference rule deflationist can accept the definition of validity in terms of truth on an interpretation thus becomes the question whether a deflationary theory of reference can be used to identify the intended interpretation. Toward answering this, I want to draw a distinction between *primitive referrers* and other, nonprimitive referring expressions. In some cases, the reference of an expression will be a function of the reference of its components. For example, the reference of "the successor of zero" might be explained as a function of the reference of a functional expression "the successor of" and the reference of the singular term "zero". What I call a *primitive referrer* is an expression whose reference is not in this way a function of the reference of more basic components. It is an expression whose reference, in a specification of the intended interpretation, is not assigned on the basis of any recursion, but is assigned, as I will say, *separately*. The primitive referrers of a language must not be confused with the primitive vocabulary, for a language might contain primitive referrers constructed from a number of vocabulary items that occur in other combinations in other constructions.

If a language contained at most finitely many primitive referrers, then a deflationary theory of reference might indeed suffice for specifying the intended interpretation of a language; but not otherwise. If there are only finitely many primitive referrers, then, for each such term *t*, we might use the inference rule deflationist's theory of reference to generate a sentence of the form "*t* refers to *e*". For instance, since we know that Mark Twain is Samuel Clemens, we might use rule B above to infer that "Mark Twain" refers to Samuel Clemens. Then employing these finitely many results we might write a general specification of the reference relation for the language. Suppose, for example, that we have a language with just one name, α, which refers to *o*, just one predicate, Φ, which refers to the set of *G*'s and just one function symbol, π, which refers to the function *f*. Then we might write a recursive specification of the reference relation for that language as follows:

> *t* refers to *e* in *L* if and only if either (i) $t = \alpha$ and $e = o$, or (ii) $t = \Phi$ and $e =$ the set of *G*'s, or (iii) $t = \pi$ and $e =$ the function *f*, or (iv) for some t' and e', $t = \pi(t')$, $e = f(e')$, and t' refers to e'.

This would still not be an analysis or explanation of the reference relation, but it would do what I have said is required, namely, specify the intended interpretation of the language in question. On the other hand, if there are infinitely many terms that need to have a reference assigned to them separately, that is, at the basis, then a deflationary theory of reference cannot in this manner be used to specify the intended interpretation. An advantage of a substantive theory of reference over this deflationary theory of reference would be that it could be used to define the reference relation for infinitely many terms all at once.

Here I have assumed that in order to use the inference rules A–D to specify the intended interpretation, we would actually have to apply the rules at least once for each term that separately has to be assigned a reference; so if infinitely many terms have to have a reference assigned to them separately, the deflationary theory of reference will not suffice. But why must we suppose that the deflationary theory of reference is capable of specifying the intended interpretation only insofar as the inference rules are *applied*? Why cannot the sheer list of rules be considered an adequate specification of the intended interpretation? This question will be taken seriously only by those who fail to grasp the distinction between an inference rule and a statement. An inference rule does not answer any question. If the question arises, "Which animals are mammals?", then it might be taken as part of an answer to say that the inference rule "*x* is an armadillo; therefore, *x* is a mammal" is valid. However, that is an answer only because this inference rule allows us to infer "For all *x*, if *x* is an armadillo, then *x* is a mammal". From the validity of that inference rule we can infer the truth of a general statement. Inference rules such as A–D do not answer the question, "What refers to what?", because, as we have seen, we cannot in the same way move from these inference rules to general statements.

A simpler deflationary theory of reference might be that the reference of a singular term is *given* by the following *reference schema*:

> "*a*" refers to *a*.

In addition, it might be said that the reference of a common noun is given by the following schema:

> "*F*" refers to the set of *F*'s.

(Further such schemata would be required for other types of referring expression.) But this theory of reference is not essentially different from the one based on inference rules. What it means to say that the reference of a term is "given" by the schema is that any instance can always be taken for granted. So in effect, these schemata are also inference rules, or, more precisely, axiom schemata. In any case, the challenge is the same. If infinitely many terms have to have a reference assigned to them separately, then mere schemata cannot be used to produce a specification of the intended interpretation.

Although this theory and the challenge to it are essentially the same as before, in this guise the theory might invite the following defense: While of course an infinite list of instances of the schemata cannot be produced, still we might specify the intended interpretation by describing it as the interpretation specified by the infinitely many instances of the reference schemata. The answer to this reply is that in merely positing the existence of an infinite list of sentences, without actually producing it, we would not fulfill our obligation to specify the intended interpretation. I do not doubt that we could uniquely identify the list as the list generated by substituting terms of the language for schematic letters in the schemata, but I do deny that we could use that list to identify the intended interpretation. We do not answer a question if we merely describe a sentence that answers the question. To answer a question we have to make a *statement* by means of a sentence that answers the question. For instance, if the question is, "Where is Scruffy now?", then the question is not yet answered by saying, "The answer to that question is given by the first sentence that Joe spoke when he walked in the door." The infinitely many instances of the reference schemata cannot be used to identify the intended interpretation because we could not actually make a statement by means of every sentence on the list.

There is an ambiguity in the word "theory" that might obscure this last point. In logic we speak of *theories* containing infinitely many sentences, such as the theory of Peano arithmetic, which is specified not by stating it, but by listing finitely many axioms and finitely many axiom schemata of which infinitely many axioms are instances. (In particular, there are infinitely many axioms having the form of the induction scheme.) But an infinite collection of sentences is not a theory in the sense of an answer to a question, something one can undertake to defend against objections. A defender of Peano arithmetic undertakes to defend the proposition that the axioms of Peano arithmetic are all true, not the proposition that ..., where the ellipsis is thought of as occupied by an infinite statement of Peano arithmetic itself. What we may require of someone who says that there is an intended interpretation is a specification of the intended interpretation that can be stated and defended against objections, and such a thing cannot be literally infinite.

Still, the demand for a specification of the intended interpretation - that can be stated - might be rejected on the grounds that the infinitely many instances of the reference schemata constitute a theory of reference for a language in the same sense in which the infinitely many axioms of Peano arithmetic constitute a theory of natural numbers. But on the contrary, a theory in that sense is not theory enough. The sense in which we can say that the infinitely many of axioms of Peano arithmetic constitute a theory of natural numbers is just that with reference to the axioms of Peano arithmetic we can formulate a theory of natural numbers as follows: For any interpretation of the language of arithmetic that is model of Peano arithmetic, the things in the domain of that interpretation are natural numbers relative to that model. But such a theory of natural numbers does not yet tell us what natural numbers *are* but only what they are *relative* to a model of the axioms of Peano arithmetic. A theory of natural numbers proper would tell us which of those models contains as its domain the set of natural numbers. Similarly, a specification of the intended interpretation for a language would not just tell us that it is the relation mapped to the predicate "refers" in a model of an infinite number of metalinguistic sentences containing that predicate, but would tell us which of those relations that might be mapped to "refers" in some interpretation of the semantic metalanguage really is the reference relation for the language in question.

If infinitely many expressions in a given language had to be assigned a reference separately, then a deflationary theory of reference would not suffice for specifying the intended interpretation of the language. If the intended interpretation could not be specified, then it would not be reasonable to define truth simpliciter for the language as truth on the intended interpretation. Since such a definition of truth for a language is inevitable if logical validity is defined in terms of truth on an interpretation, a deflationist would not be free to define logical validity in that way. So the question whether a deflationist is free to accept a standard model theoretic definition of validity for a language comes down to the question of how many terms in that language would have to be separately assigned a reference.

The question is not whether the inference rule deflationist can identify the reference relation as it pertains to all possible languages. No doubt, if we counted the expressions in all possible languages that would separately have to have a reference assigned to them (if they existed), then the number would be (denumerably) infinite. That is not an objection to inference rule deflationism because the inference rule deflationist's claim is precisely that all that can be said about reference *in general* is that certain rules of inference pertain to "refers" and its synonyms in other languages. The question is whether the inference rule deflationist can identify the reference relation as it pertains to a specific language, for that, I have argued, is what one must be able to do, at least in principle, if one wishes to define logical validity in the manner of standard model theory. That question, we have seen, comes down to how many terms of a single language would have to be separately assigned a reference - a question I will now take up in greater detail.

7 Compositionality and finitude

In fact, a specification of the intended interpretation of a natural language will have to assign distinct referents to infinitely many terms, for natural languages contain infinitely many primitive referrers. Consider, for instance, the predicate "believes that Ortcutt is a spy". Its reference is the set of people who believe that Ortcutt is a spy. But the reference of this predicate is not a function of the reference of "believes", "Ortcutt" and "is a spy". Similarly, there are infinitely many other belief-predicates in English whose reference is not a function of the reference of their parts. Perhaps a natural language must contain at most finitely many primitive vocabulary items, but a natural language will contain infinitely many primitive referrers.

In taking for granted that the reference of a belief-predicate is not a function of the reference of its components I am taking for granted only that "believes that *p*" is, in at least some of its occurrences, referentially opaque, for the former assumption is an immediate consequence of the latter. For example, suppose that "Ralph believes that Ortcutt is a spy" is true and "Ortcutt" and "Snodgrass" corefer. Then if the extension of "believes that Ortcutt is a spy" were a function of the reference of its component expressions, then, since the reference of "Ortcutt" and "Snodgrass" is the same, this function would determine that Ralph is in the extension of "believes that Snodgrass is a spy" as well. So "Ralph believes that Snodgrass is a spy" would be true. Generalizing, we may infer that if the reference of "believes that *p*" were a function of the referents of its component expressions, then "believes that *p*" would always be referentially transparent. But "believes that *p*" is not always referentially transparent, and so its extension is not a function of the reference of its component expressions.

Someone might hope (following Frege) that the reference of a belief-predicate might be understood as a function of the reference of the components if the referents of the components are the *intensions* that the components have in nonintensional contexts. But first, it is doubtful whether a deflationist can take the referent of an expression to be an intension. (If we can explain in a substantive way what intensions are, then should we not be able to explain in a substantive way what truth is?) Second, it is doubtful whether a deflationist can derive the reference of a complex expression from the intensions of its components. (The usual method assumes that intensions can be defined as infinite functions from parameters to referents. The doubt is whether a deflationist can explain what it means for an expression to have a reference at a parameter.) In any case, this strategy does not obviate the conclusion that a natural language must contain infinitely many primitive referrers. On this theory, the reference of "Ortcutt" in "believes that Ortcutt is a spy" must be different from the reference of "Ortcutt" in "believes that Ralph believes that Ortcutt is a spy", and similarly we have to suppose that each term takes a different reference at each different level of embedding. Thus even if only finitely many different expressions serve as primitive referrers, each of them will have a different reference at each of infinitely many levels.[7]

Of course, there are many proposals for the semantic analysis of attributions of propositional attitude. According to some of them, the "that"-clauses are referentially transparent and the expressions that occur within them take their normal reference. So strictly speaking I cannot show that the deflationist must reject model-theoretic semantics, but only that the deflationist must either reject model-theoretic semantics or accept the referential transparency of intensional locutions generally (with standard reference of component expressions). But I think that the latter option will be acceptable to few, and so in the remainder I will ignore it.

In arguing in this way that a natural language will contain infinitely many primitive referrers, I am not also arguing, or taking for granted, that natural languages lack a compositional semantics. Even if we allow that there are infinitely many primitive referrers, we might have good reasons for restricting the class of primitive referrers in one way or another to a proper subset of the expressions of the language as a whole. The nontrivial task of a compositional semantics would then be to show that interpretations of the remaining expressions of the language can be generated from an interpretation of the (infinitely many) primitive referrers. Some philosophers and linguists have probably conceived of compositionality as entailing a finite basis (for example, Schiffer 1987), which of course they are free to do, but this conception is not universal (see, for instance, Janssen 1997), and in any case the question of compositionality can be separated from the question of a finite basis, for one might question the possibility of a non-trivial compositional semantics even while allowing that the basis might be infinite.

These arguments may seem to produce only paradox rather than a demonstration of the incompatibility of deflationism and model-theoretic semantics. They will be paradoxical to anyone who thinks that a natural language must have a compositional semantics and for whom the primary motives for seeking a compositional semantics equally motivate a restriction to finitely many primitive referrers. There are basically two different motivations for compositional semantics that we need to consider, which I will call the logical motivation and the psychological motivation. Neither one, I will now argue, gives us a good reason to resist the prima facie evidence that a natural language contains infinitely many primitive referrers.

The purely logical motivation to seek a compositional semantics is primarily to obtain a definition of logical validity that can be used to demonstrate of valid arguments that they are valid and of invalid arguments that they are invalid. Suppose that English contains only finitely many primitive referrers, and imagine a language a lot like English except that it contains infinitely many primitive referrers. Call it Infinite English. Would there be any forms of argument that were valid for Infinite English but not valid for English or valid for English but not valid for Infinite English (i.e., such that exclusively valid arguments in Infinite English had that form while some invalid

[7] A possibility that ought to be mentioned is that the deflationist might try to escape my argument by exploiting the fact that the levels are systematically related to one another in some way. However, I will not try to anticipate and answer the objections that a deflationist might make on this basis.

arguments of English had that form, or the other way around)? If so, then the disparity would have to concern argument forms containing infinitely many distinct schematic letters. (Since sentences are finite, such forms would be forms of arguments with infinitely many premises, which we can allow. We can allow such forms even for English.) Any counterexample in English would be equally a counterexample in Infinite English; so the only possibility is that there are forms of argument that contain infinitely many sentence forms and are valid for English but not valid for Infinite English. I do not know of any such forms of argument. I conclude that a compositional semantics might perfectly well serve a definition of logical validity without restricting itself to finitely many primitive referrers.

I said that the logical motivation to seek a compositional semantics is *primarily* to obtain a definition of logical validity. I put it that way because I think that some semantic theorists have conceived of their goal somewhat differently, as being to explain how linguistic structures can correspond to reality or fail to correspond. Thinking of the goal in this way, there might be reason to draw distinctions that the relation of logical validity for the language does not depend on. No deflationist, whose aim is to provide an alternative to the correspondence theory, should offer this conception of the goal of compositional semantics as a reason for resisting the prima facie evidence that a natural language may contain infinitely many primitive referrers. Moreover, I do not see how such a conception of the goal of compositional semantics, apart from any other motive, might lead us to insist that a natural language can contain at most finitely many primitive referrers.

The psychological motivation to seek a compositional semantics is to explain how it is possible for a person to learn a language – how it is possible for a hearer to understand a speaker's words and how it is possible for a speaker to choose his or her words. On this conception, a compositional semantics for a language will posit at most finitely many primitive referrers if at most finitely many primitive referrers could be learned and understood. For instance, Donald Davidson has famously argued that a language will be learnable only if it contains at most finitely many "semantical primitives" (1984/1965). In contemplating this, it is important not to confuse primitive vocabulary with primitive referrers. I do not doubt that a person can learn and understand at most finitely many vocabulary items. It does not follow that a person can learn and understand at most finitely many primitive referrers, for it may be that infinitely many primitive referrers are formed from finitely many primitive vocabulary items. But perhaps we should conclude that a language will be learnable only if it contains at most finitely many primitive referrers on the grounds that a person learns a language in part by separately grasping the reference of each of the primitive referrers in the language and that a person can engage in at most finitely many acts of grasping a reference.

A deflationist, of all people, should be especially inclined to doubt that a language is learnable and understandable only if it contains at most finitely many primitive referrers. To believe this would be to suppose that the relation of reference will play an important role in our explanation of the nature of cognition or linguistic communication. If the relation of reference has this kind of explanatory utility, then the definition of truth simpliciter as truth on the intended interpretation, which assigns to each term what it really refers to, will be a substantive theory of truth that competes with the deflationist's own.

In any case, I do not see how the nature of language learning and understanding is made more tractable by supposing that there are only finitely many primitive referrers. Inasmuch as the range of the composition is infinite, if there are only finitely many primitive referrers, then the principles of composition (if there are only finitely many of these) will have to include functions having an infinite domain and an infinite range. Thus, the restriction to finitely many primitive referrers would not remove the need to suppose that the mind can in a sense grasp infinities. Consider one's grasp of the reference of infinitely many numerals. Perhaps it will be said that we grasp the reference of infinitely many numerals only in the sense that we grasp the reference of a finite number of numerals (maybe just "0") and grasp various operations on numbers, such as addition (or maybe just the successor function). I do not understand how a grasp of these operations, with all of their infinitely many applications, is supposed to be psychologically more tractable than a grasp of the reference of infinitely many distinct numerals.[8]

Perhaps we may conclude that there can be at most finitely many primitive referrers on the grounds that languages are learned by means of a kind of translation into a prior language of thought and that there can be at most finitely many primitive referrers in the language of thought (although it is hard to believe that this could have been any part of Davidson's motivation). Even supposing that there is a prior language of thought, I do not see any reason additional to those already considered to suppose that there are at most finitely many primitive referrers in the language of thought.

In sum, I do not think there is any very persuasive reason to resist the prima facie evidence that a natural language such as English will contain infinitely many primitive referrers. As we have seen in the previous section, if the possibility of infinitely many primitive referrers is allowed, then the inference rule deflationist must reject the standard model theoretic

[8] This is of course the subject of Kripke 1982. My own contribution to this debate is Gauker 1995.

definition of logical validity. I conclude that the inference rule deflationist must reject the standard model-theoretic definition of logical validity.

8 T-schema deflationism

Inference rule deflationism may not be what people usually think of as deflationism. More often, deflationism is supposed to say that the meaning of the word "true" is expressed in a *T-schema* such as,

> "*p*" is true if and only if *p*,

or

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

Call this *T-schema deflationism*. T-schema deflationism, as here defined, is what deflationism is understood to be by some of its critics (see Gupta 1993a and David 1994). A characteristic formulation is this one from Frederick Schmitt (also a critic): The deflationist proposes "that the notion of truth in a given language is completely captured by the trivial truth conditions or 'T-sentences'" (Schmitt 1995, p. 124).

The first question we have to ask about T-schema deflationism is: what does it mean to say that the meaning of "true" is *captured* or *expressed* by the T-schema? What is special about instances of the T-schema that distinguishes them from other sentences containing the word "true"? If any part of the answer is that the instances of the T-schema are all *true*, then T-schema deflationism may be charged with circularity. Perhaps part of what is special about instances of the T-schema is that they are all *assertible*. But what does that mean? Certainly it does not mean that in any conversation it is appropriate to assert any instance of the T-schema. If part of the explanation is that any such assertion would at least be true, then again the theory is circular. If assertibility could be explained in some other way, then perhaps we could explain the meaning of "true" by giving the assertibility conditions of "*x* is true", and in that case we would not need T-schema deflationism at all.

Another possibility would be to say that what is special about instances of the T-schema is that one is permitted to treat them as axioms in logical derivations. So construed, T-schema deflationism is at least as strong as inference rule deflationism (given Modus Ponens), because any conclusion that can be derived using the rules of Semantic Ascent and Semantic Descent can be derived as well using instances of the T-schema (and Modus Ponens). But inference rule deflationism is not as strong as T-schema deflationism because these rules do not yield instances of the T-schema without the help of less obvious rules (in particular, Conditional Proof).

There is no reason to prefer the stronger thesis and there is some reason to prefer the weaker. The greater strength of T-schema deflationism plainly does not afford any means of identifying the reference relation for a language. However, inference rule deflationism has an advantage over T-schema deflationism with respect to the paradoxes. The inference rule deflationist can hope to evade the paradoxes by denying the validity of Conditional Proof and Indirect Proof. This strategy will not help T-schema deflationism. Consider again the classic liar. By virtue of the plain fact that α = "α is not true", we know:

(i) α is true if and only if "α is not true" is true.

Instantiating the T-schema with the liar, we obtain:

(ii) "α is not true" is true if and only if α is not true.

By almost everyone's lights, (i) and (ii) cannot both be true (or, more generally, cannot both have whatever semantic property must be preserved in valid arguments). So by almost everyone's lights, if we have a complete system of axioms and inference rules (one such that for every valid argument the conclusion can be derived from the premises), then it will be possible to derive a contradiction from (i) and (ii).

It would be a mistake for a T-schema deflationist to react to the semantic paradoxes simply by acknowledging that the T-schema must be subjected to "restrictions" in order to avoid the paradoxes. As we saw in connection with inference rule deflationism, one cannot simply rule out those instances that on their own permit the derivation of contradictions from plain facts, since the problem with an instance may be that a contradiction can be derived by means of it together with other instances. Moreover, no syntactic condition on the T-schema will draw the boundary between those instances that contradict plain facts and those that express part of the meaning of "true". One might try to identify the acceptable instances semantically, but if logical validity is defined in terms of truth on an interpretation, then the question concerning the nature of truth will be begged. If one's semantic theory is formulated otherwise than in terms of truth, then one should expect to be able to use it to explain the meaning of the word "true" and, in that case, T-schema deflationism will have been supplanted.[9]

[9] Gupta and Belnap (1993) have put forward a theory of truth that at first glance seems to combine the classical rules of inference with the acceptance of every instance of the T-schema. Gupta and Belnap do not themselves offer this as a kind of deflationism, however, and it is doubtful whether the Gupta-Belnap deductive calculus can be appealed to as a way of combining deflationism with the classical rules of inference. (Gupta himself is a stern critic of deflationism; see his 1993a and 1993b.) In the Gupta-Belnap deductive calculus, the instances of the T-schema have a special status as partial definitions and one cannot say that they are all valid or even simply true. A definition can be appealed to at any point in a proof, but not in the manner of a logically valid sentence. When "*p* is true if and only if *p*" is merely

9 Horwich's minimalism

A further form of deflationism is Paul Horwich's *minimal* theory of truth. Since this is often regarded as the paradigm of deflationism, separate discussion of it is warranted. Horwich's minimal theory is supposed to include every proposition of the form *The proposition that p is true if and only if p* (1990, pp. 18–19; 1998, pp. 17–18). As Horwich acknowledges, this would mean that the minimal theory were literally infinite. In logic we may speak of infinite sets of sentences as "theories", but it is questionable whether an infinite theory is the sort of thing one can adopt as a solution to a problem and defend against objections. Since the theory is infinite, it cannot even be stated. Indeed (as Gupta has observed in his 1993b, p. 360), what Horwich proposes to defend against objections is not the minimal theory per se but various claims about it, such as that the minimal theory contains all the essential facts about truth. In any case, whether Horwich's theory of truth *is* the minimal theory or is only certain claims *about* the minimal theory, he should tell us exactly what the minimal theory is. Can he do so?

In fact, Horwich does not tell us which propositions there are of the form *The proposition that p is true if and only if p*. He introduces a function symbol "E*" and writes as if it denoted a function from propositions that *p* into propositions of the form *The proposition that p is true if and only if p*. Then he says (1990, p. 21; 1998, pp. 21–22) that a proposition x is an axiom of the minimal theory if and only if for some y, $x = E^*(y)$. But this attempt to describe a theory is a sham since Horwich has not defined or in any way identified the function E*. The illusion that we understand the minimal theory is engendered by the belief that we can take any well formed sentence of English and put it in place of "*p*" in the sentence schema "The proposition that *p* is true if and only if *p*" and the result will be a sentence expressing an axiom of the minimal theory, so that we will know what belongs to the minimal theory at least in so far as it concerns our own language. In fact we cannot always form a sentence expressing an axiom of the minimal theory in this way. For instance, if the sentence we substitute for "*p*" is the classic liar this will not be the case.

What Horwich says about this is that such instances are not supposed to express axioms of the minimal theory because they "engender 'liar-type' contradictions". Moreover, instances are not to be excluded unnecessarily and the specification of exceptions should be simple (1990, p. 42; 1998, p. 42).[10] He does not seem to consider that what might be required in order to produce this specification is precisely an alternative theory of truth. I cannot guess what manner of specifying the exceptions Horwich might have in mind, but it will be worthwhile to demonstrate that it will not do simply to say that an instance of the T-schema expresses an axiom of the minimal theory unless it expresses a contradictory proposition. Call this the *selective description* of the minimal theory. The problem is that this selective description itself implies that the minimal theory contradicts plain facts. Suppose that α = the proposition that α is not true. Let *s* be the sentence "The proposition that α is not true is true if and only if α is not true". Two cases: Case 1: "α is not true" expresses a noncontradictory proposition. Then we should expect that likewise *s* expresses a noncontradictory proposition, namely, the proposition giving the truth conditions of the proposition that α is not true. So by the selective description, *s* expresses an axiom of the minimal theory. Case 2: "α is not true" expresses a contradictory proposition. Then *s* is not contradictory because both sides of the biconditional express contradictory propositions. So by the selective description, *s* again expresses an axiom of the minimal theory. In either case, *s* expresses an axiom of the minimal theory, and so the minimal theory is inconsistent with the fact that α = the proposition that a is not true.

The preceding demonstration takes for granted that "α is not true" and, therefore, *s* express propositions. So, alternatively, Horwich might maintain that an instance of the T-schema expresses an axiom of the minimal theory unless it fails to express a proposition at all. The reason *s* does not express an axiom of the minimal theory, he might say, is just that it does not express a proposition at all. But if this is how Horwich proposes to identify the minimal theory, then he owes us a general account of which instances of

definitional, one cannot infer from that biconditional and "*p* is true" to *p* or from that biconditional and *p* to "*p* is true". Rather, lines of proofs bear superscripts and from $[p \text{ is true}]^i$ one may infer only $[p]^{(i+1)}$ and conversely. The suggestion would be that a form of T-schema deflationism might likewise preserve the classical rules of inference by declaring that the instances of the T-schema have the status of definitions and not that of logically valid sentences. I think it is doubtful whether this would qualify as a form of deflationism, however, precisely because of the distinction it draws between the T-biconditionals taken as partial definitions of truth and the T-biconditionals taken as material or necessary biconditionals. A sharp distinction has to be drawn because we are supposed to be able to draw sound inferences from the partial definitions even when the corresponding material or necessary biconditionals are false. The T-biconditionals taken as partial definitions do not themselves provide an explanation of this distinction, and so it is doubtful whether the T-biconditionals taken as partial definitions can be supposed to explicate the meaning (i.e., the cognitive significance, to use Gupta and Belnap's phrase) of "true".

[10] In the second edition Horwich gives an example of a liar sentence and uses it to derive a contradiction having the form of a biconditional (1998, pp. 40–41). He then lists four classical inference rules he has used and says that one option, which he rejects, would be to deny "classical logic". But he does not list Conditional Proof, which he has also used, and he does not point out that the conclusion can also be drawn in just one step by just one identity substitution.

the T-schema express propositions. By the same sort of arguments that we encountered in section 3 above (and rehearsed in section 8), it should be evident that he cannot succeed at this. One cannot say merely than an instance fails to express a proposition if a contradiction can be derived from it together with plain facts; no purely syntactic criterion will identify the instances that express elements of the minimal theory; and a semantic characterization of the acceptable instances will either beg the question of the nature of truth or offer us alternative resources for explaining what truth is.

10 Conclusion

I conclude from these arguments that the deflationist must abandon some classical rules of inference and must abandon standard model theory. I do not conclude that deflationism is mistaken on the grounds that we should cling to the classical rules of inference. The validity of the problematic instances of classical rules of inference can comfortably be denied. The incompatibility with standard model theory is a more serious problem for deflationism, for as I argued in section 4, we do require a semantic definition of logical validity. However, I do not assume that we should cling to standard model theory. The primary motivation for deflationism remains, namely, the failure of all attempts heretofore to explain the nature of the correspondence relation. The conclusion I draw is that we should formulate our semantics using concepts other than truth and should explain the meaning of "true" in terms of those.[11]

Christopher Gauker
Philosophy Department
University of Cincinnati
Cincinnati, Ohio 45221-0374
christopher.gauker@uc.edu

References

Barwise, Jon and John Etchemendy. 1987. *The Liar: An Essay on Truth and Circularity*. Oxford: Oxford University Press.

David, Marian. 1994. *Correspondence and Disquotation*. Oxford: Oxford University Press.

[11] I gratefully acknowledge the support of the Taft Memorial Fund of the University of Cincinnati for this research. I also thank Bradley Armour-Garb for discussion that led to some improvements.

Donald Davidson. 1984. "Theories of Meaning and Learnable Languages." In his *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press. Pp. 3–15. (Originally published in 1965.)

Field, Hartry. 1972. "Tarski's Theory of Truth." *Journal of Philosophy* 69: 347-75.

Field, Hartry. 1994a. "Deflationist Views of Meaning and Content." *Mind* 103: 249–285.

Field, Hartry. 1994b. "Disquotational Truth and Factually Defective Discourse." *Philosophical Review* 103: 405–452.

Gauker, Christopher. 1990. "Semantics without Reference." *Notre Dame Journal of Formal Logic* 31: 437-461.

Gauker, Christopher. 1994. *Thinking Out Loud: An Essay on the Relation between Thought and Language*. Princeton, NJ: Princeton University Press.

Gauker, Christopher. 1995. "A New Skeptical Solution." *Acta Analytica* 14: 113–129.

Grover, Dorothy. 1992. *A Prosentential Theory of Truth*. Princeton, NJ: Princeton University Press.

Gupta, Anil. 1993a. "A Critique of Deflationism." *Philosophical Topics* 21: 57–81.

Gupta, Anil. 1993b. "Minimalism." In *Philosophical Perspectives 7, Language and Logic*, edited by James Tomberlin. Atascadero, CA: Ridgeview Publishing Co. Pp. 359–369.

Gupta, Anil and Nuel Belnap. 1993. *The Revision Theory of Truth*. Cambridge, MA: MIT Press.

Horwich, Paul. 1990. *Truth*. Cambridge, MA: Blackwell.

Horwich, Paul. 1998. *Truth*, Second Edition. Oxford: Oxford University Press.

Janssen, Theo M. V. 1997. "Compositionality." In *Handbook of Logic and Language*, edited by Johan van Benthem and Alice ter Meulen. Cambridge, MA: MIT Press. Pp. 417–464.

Kripke, Saul. 1975. "Outline of a Theory of Truth." *Journal of Philosophy* 72: 690–716.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Cambridge, MA: Harvard University Press.

McGee, Vann. 1992. "Maximal Consistent Sets of Instances of Tarski's Schema (T)", *Journal of Philosophical Logic* 21: 235–241.

Schmitt, Frederick. 1995. *Truth: A Primer*. Boulder, CO: Westview Press.

Stephen Schiffer. 1987. *Remnants of Meaning*. Cambridge, MA: MIT Press.

Weir, Alan. 1996. "Ultramaximalist minimalism!" *Analysis* 56: 10–22.