THE JOURNAL OF PHILOSOPHY

VOLUME LXXXI, NO. 8, AUGUST 1984

WHAT IS A THEORY OF TRUTH?*

ALFRED TARSKI'S theory of truth and its successors enjoy a perplexing double status. On the one hand, they are mathematical theories characterized by a rich class of mathematical results. On the other hand, they are commonly believed by philosophers to provide analyses of the nature of truth and, hence, to be philosophically significant. With this broader significance comes a kind of controversy not normally associated with mathematical theorems. No one disputes the correctness of Tarski's formal results. In that sense, there is no doubt that his theory is true. However, there is considerable doubt about whether, or in what sense, it is a theory of truth.

One main reason for this uncertainty is the difficulty of determining what a theory of truth ought to be. Generally, theories of truth have tried to do one or the other of three main things:

(i) to give the meaning of natural-language truth predicates;
(ii) to replace such predicates with substitutes, often formally defined, designed to further some reductionist program; or
(iii) to use some antecedently understood notion of truth for broader philosophical purposes, such as explicating the notion of meaning or defending one or another metaphysical view.

In order to do the first of these things, a theory must analyze the content of paradigmatic examples in which what is said to be true is a proposition, rather than a sentence or utterance.

(1) a. The proposition that the earth moves is true.
    b. Church's theorem is true.
    c. Everything he said is true.

*This paper was presented in the spring of 1983 at Yale and Dartmouth, where I profited from useful discussion. Thanks are also due to Paul Benacerraf, Axel Buhler, John Burgess, Gilbert Harman, David Lewis, and Walter Sinnott-Armstrong for reading and commenting on an earlier draft.

0022-362X/84/8108/0411$01.90

There are theories that try, in my opinion unsuccessfully, to do just this.[1] Tarski's theory, which restricts itself to cases in which truth is predicated of sentences of certain formal languages, is not one of them. Thus, Tarski cannot be seen as even attempting to give the meaning of natural-language truth predicates.

Nor can he be seen as attempting to use the notion of truth for broad philosophical purposes. In order to do that, one must regard truth as essentially unproblematic and philosophically productive. For Tarski, truth itself is what has to be legitimated. Once it is, it turns out to be useful for certain primarily technical purposes, but useless for ambitious philosophical programs. For example, Tarski recognized that his notion of truth could not be used to give the meanings of logical constants (or, I suspect, anything else).[2] He also thought of it as epistemologically and metaphysically neutral. Thus, in "The Semantic Conception of Truth" he says:

> . . . the semantic definition of truth implies nothing regarding the conditions under which a sentence like . . .
>
> *snow is white*
>
> can be asserted. It implies only that, whenever we assert or reject this sentence, we must be ready to assert or reject the correlated sentence . . .
>
> *The sentence "snow is white" is true.*
>
> Thus, we may accept the semantical conception of truth without giving up any epistemological attitude we may ever have had; we may remain naive realists, critical realists or idealists, empiricists or metaphysicians—whatever we were before. The semantic conception is completely neutral toward all these issues (33/4).

It is helpful, in understanding this remark, to focus on something that the truth predicate is good for—namely, what W. V. Quine has called "semantic ascent."[3] The simplest example of this is provided by Tarski:

(2) a. Snow is white.
    b. The sentence 'Snow is white' is true.

Any speaker of English knows that these sentences are at least materially equivalent. Because of this, they can often be used to convey essentially the same information. To choose (2b) is to use a semantic statement to convey information that could have been

[1] Various versions of the "redundancy theory" fall into this category. Although these versions deal with "propositional" contexts like those in (1), they deny that 'true' is predicated of propositions, or anything else. Arguments against these approaches are given in my *Semantic Theories of Truth* (New York: Oxford, in preparation).

[2] "The Semantic Conception of Truth," sec. 15, in Leonard Linski, ed., *Semantics and the Philosophy of Language* (Urbana: U of Illinois Press, 1952).

[3] *The Philosophy of Logic* (Englewood Cliffs, N.J.: Prentice-Hall, 1970), pp. 10-13.

conveyed nonmetalinguistically. To do this is to engage in semantic ascent.

The importance of semantic ascent is illustrated by cases like (3), in which we want to generalize.

(3) a. Snow is white → (Grass is blue → Snow is white)
b. The earth moves → (The sun is cold → The earth moves)
⋮

Each of these examples is something one could feel safe in asserting. However, if one wanted to get the effect of asserting all of them, one would have to quantify, replacing sentences with variables. In English such quantification is most naturally, though not inevitably, construed as first-order and objectual. Thus, if the variables are taken to range over sentences we need a metalinguistic truth predicate. Semantic ascent gives us

(4) For all sentences p, q (p is true → (q is true → p is true))[4]

That which is conveyed by (4) is closely related to that which is conveyed by (3). However, here the truth predicate is especially handy, since we don't have the alternative of asserting each member of (3).

Truth predicates can be used in the same way in more obviously philosophical cases. For example, consider:

(5) There is a duplicate of our sun in some remote region of space, but we will never find (sufficient) evidence that there is.

Someone who asserted this would, by contemporary standards, be counted a metaphysical realist—i.e. as being someone who thinks that what there is doesn't depend in any way on what we may rationally believe. Of course, one can be a realist without believing (5). One may think that what there is doesn't depend on us, while believing that there is no duplicate of our sun, or being uncertain whether there is, or while holding that evidence will someday be found to settle the matter.

What, then, distinguishes realism from anti-realism? One is tempted to answer that it is the belief that

(6) Either there is a duplicate of our sun in some remote region of space, but we will never find (sufficient) evidence that there is; or
there is no duplicate of our sun in any remote region of space, but we will never find (sufficient) evidence that there isn't; or

[4] Or, equivalently, (4′):
(4′) For all sentences p, q $\ulcorner(p \rightarrow (q \rightarrow p))\urcorner$ is true.

there is intelligent life elsewhere in the universe, but we will never find (sufficient) evidence that there is; or

⋮

But this is awkward. We ought to be able to state the realist's position without having to gesture toward an infinite list. Semantic ascent provides a convenient way of doing this. With the help of a truth predicate and quantification over sentences, we can characterize the realist as believing, and the anti-realist as denying:

(7) There is at least one sentence *s* such that *s* is true (in English), but we will never find (sufficient) evidence supporting *s*.

The relationship between (6) and (7) is like that between (2a) and (2b) and between (3) and (4). In each case, the semantic sentence may not say exactly the same thing as its nonsemantic counterpart; but if knowledge of English is assumed, the two can be used to convey essentially the same information.[5]

The utility of the truth predicate in stating this dispute has led some to believe that the dispute is about truth and, hence, that truth is a deeply metaphysical notion. However, there is no reason to suppose this. The realist and anti-realist may agree about truth; they may even accept something like Tarski's definition. Where they differ is in their conceptions of reality. Since statements about truth mirror direct statements about nonlinguistic reality, semantic ascent makes the truth predicate a convenient vehicle for expressing competing metaphysical views. But a convenient vehicle is all it is. As Tarski puts it, the notion of truth is completely neutral toward all these issues.

The upshot of this is that Tarski's definition of truth is neither an attempt to analyze the meaning of natural-language truth predicates nor an attempt to use the notion of truth for broad philosophical purposes. Rather, Tarski's goal was to replace natural-language truth predicates with certain restricted, but formally defined substitutes. He thought such replacements were needed both to remove the doubts of certain scientifically minded truth skeptics and to eliminate what he took to be the incoherence in our ordinary notion brought out by the liar paradox.

For Tarski, these two motivations were connected, since the par-

[5] There are, I presume, many versions of realism, of which (6) and (7) represent only one. A different version might hold that some sentence of English is such that it is metaphysically possible for it to be true (keeping the semantics fixed) in cases in which the proposition it expresses cannot (ever) be known or rationally believed. This thesis is no more linguistic than (6) and (7) are.

adoxes constituted one source of skepticism about truth.[6] However, the truth skeptics of his day also had other, more broadly philosophical grounds for their doubts. These included the frequent use of truth in metaphysical discussions, the tendency to confuse truth with epistemological notions like certainty and confirmation, and the inability to see how acceptance of a truth predicate could be squared with the doctrine of physicalism and the unity of science.[7] Although Tarski's work was historically effective in alleviating each of these worries, the only one discussed by Tarski was the final one, involving physicalism.[8]

Tarski's version of physicalism was a moderate one, allowing both physical and mathematical elements, without requiring the latter to be reduced to the former. Roughly, this "moderate physicalism" asserts that

(i) all facts are physical or mathematical facts;

(ii) all scientific (or descriptive) claims are reducible to claims about the physical and mathematical characteristics of things; and

(iii) all scientific (or descriptive) concepts are definable in terms of physical and mathematical concepts.[9]

Tarski took this doctrine to require that truth be eliminable via an explicit, physicalistic definition. Anything else—for example, taking truth to be a primitive whose extension is fixed by a set of axioms—was deemed to be undesirable.

It is worth pointing out that this emphasis on definition is primarily philosophical rather than technical. What is at issue is

[6] And also of skepticism about related notions like definability. Tarski cites the paradoxes as a source of skepticism in "The Establishment of Scientific Semantics" and, as John Burgess has pointed out to me, in "On Definable Sets of Real Numbers," reprinted in *Logic, Semantics, Metamathematics* (New York: Oxford, 1956).

[7] See C. G. Hempel, "On the Logical Positivists' Theory of Truth," *Analysis*, II, 4 (January 1935): 49-59; Rudolf Carnap, "Intellectual Autobiography," in P.A. Schilpp, ed., *The Philosophy of Rudolf Carnap* (LaSalle Ill.: Open Court, 1963), p. 63; Carnap, "Truth and Confirmation," in H. Feigl and W. Sellars, eds., *Readings in Philosophical Analysis* (New York: Appleton-Century-Crofts, 1949); Hans Reichenbach, *Experience and Prediction* (Chicago: University Press, 1938), sec. 22; Otto Neurath, "Sociology and Physicalism" and "Protocol Sentences", both reprinted in A. J. Ayer, *Logical Positivism* (Glencoe Ill.: Free Press, 1959); Karl Popper, *The Logic of Scientific Discovery* (New York: Harper Torchbook, 1965), p. 274; and Hartry Field, "Tarski's Theory of Truth," this JOURNAL, LXIX, 13 (July 13, 1972): 347-375; page references to Field will be to this article.

[8] "The Establishment of Scientific Semantics," *op. cit.*, p. 406.

[9] Tarski's physicalism countenanced both physical science and "logic," where the latter was construed as including set theory and everything obtainable from it. In what follows, I will use the term 'physicalism' in the moderate sense of (i-iii) above. In particular, physicalism, in my sense, does not require the reduction of set-theoretical facts, mathematical facts, or syntactic facts about expression types.

not the technical results achievable, but the philosophical significance of those results.[10] It is possible to view a Tarski truth characterization for a language *L* as simply specifying the extension of 'true' for *L*, explaining how the truth value of a sentence depends on the semantic properties of its parts, and providing the basis for accounts of logical truth and logical consequence. Even if the truth characterization is put in the form of what is technically an explicit definition, it doesn't have to be viewed as an explication of truth in any interesting philosophical sense. If one's philosophical views differ from Tarski's, one can accept his formal results while taking truth to be primitive.

I will not comment directly on this way of viewing Tarski, but will instead concentrate on his own view of his work. I do this not out of any commitment to physicalism, but rather out of a sense that his deflationist attitude toward truth is interesting and worth defending. Tarski's attitude needs defense because his definition of truth fails to satisfy certain initially plausible demands one might place on an explication of truth. His attitude is defensible because these demands turn out to be dubious or illegitimate. The importance of this defense extends beyond Tarski to the general question of what ought, and what ought not, to be expected from a theory of truth.

II

Tarski's basic idea is that for certain languages *L*, adequate for natural science, one can define a truth predicate using only notions already expressible in *L*, plus certain syntactic and set-theoretic apparatus. Thus, if *L* is physicalistically pure and if syntax and set theory are unproblematic, then defining a metalanguage truth predicate can't introduce any difficulties.

Following Hartry Field, we can think of such a definition as divided into two parts. The first part is concerned with what Field calls "primitive denotation"; here one defines what it is for a name to refer to an object and for a predicate to apply to one or more objects. The second part of the definition defines truth in terms of primitive denotation. The end result is a metalanguage sentence:

(8) For all sentences s of L, s is true iff $T(s)$.

[10] There are, of course, technical issues as well. When the metalanguage contains quantifiers ranging over arbitrary subsets of the domain of the object language, an explicit definition of object-language truth is possible in the metalanguage. On the other hand, if a classical object language containing set theory has quantifiers ranging over all sets, then an explicit metalanguage definition of truth is impossible. Tarski's emphasis on explicit definition is philosophical in the sense that he saw significant philosophical advantages in explicit definitions of truth, *where they are possible.*

in which $T(s)$ is a formula with only 's' free, made up entirely of logical, set-theoretic, and syntactic apparatus, plus translations of the primitives of L. If these translations are (extensionally) correct, then $T(s)$ will be coextensive with 'true' over L.

Tarski's technique can be illustrated using a particularly simple example. Let L be a language whose only logical constants are '$\vee$' and '$-$', and whose nonlogical constants consist of finitely many names and one-place predicates. (R) and (A) define reference and application for L; (T) uses these notions to define truth:

(R) For all names n of L and objects o, n refers in L to o iff n = 'a' and o = Arizona, or n = 'b' and o = Boston, . . . (and so on for each of the names of L).

(A) For all one-place predicates P of L and objects o, P applies in L to o iff P = 'C' and o is a city, or P = 'S' and o is a state, . . . (and so on for each one-place predicate of L).

(T) For all sentences S of L, S is true in L iff $S \varepsilon$ the set K such that for all x, $x \varepsilon K$ iff

(i) $x = \ulcorner Pn \urcorner$ for some predicate P and name n of L, and there is an object o such that *n refers in L to o* and *P applies in L to o*; or

(ii) $x = \ulcorner (A \vee B) \urcorner$ for some formulas A and B of L, and $A \varepsilon K$ or $B \varepsilon K$; or

(iii) $x = \ulcorner -A \urcorner$ for some formula A of L and $A \not\varepsilon K$.

Let (T′) be just like (T) except for containing the right-hand sides of (R) and (A) where (T) contains *n refers in L to o* and *P applies in L to o*, respectively. (T′) is then an explicit Tarskian truth definition for L, with '$T(s)$' in (8) representing the right-hand side of (T′).[11]

Although truth definitions for richer languages are technically more interesting, their philosophical status as putative physicalistic reductions of truth is essentially the same as that of the simple definition just given. On the basis of such definitions, Tarski concluded that he had shown truth, reference, and application to be physicalistically acceptable terms.

In a well-known critique of Tarski, Hartry Field (*op. cit.*) argues that this conclusion is unjustified. The problem, according to Field, is that the proposed replacements for the notions of primitive denotation are not physicalistically acceptable reductions of

[11] Note, since there are only finitely many atomic formulas in L, that we could have got an equivalent result by substituting (i′) for (i) in (T).

(i′) x = 'Ca' and Arizona is a city, or x = 'Cb' and Boston is a city, or x = 'Sa' and Arizona is a state, or x = 'Sb' and Boston is a state, . . . (and so on for each atomic formula).

our pretheoretic notions of reference and application. Because Field takes Tarski to have reduced truth to primitive denotation (350), he concludes that Tarski has not legitimated the notion of truth for physicalists.

Field does not, of course, dispute the fact that Tarski's definitions are extensionally correct. He maintains, however, that extensional correctness is not enough. In addition, any genuine reduction must show semantic facts about expressions to be supervenient on physical facts about their users and the environments in which they are used. Tarski's definitions don't do this.

This can be seen by considering a simple example. Suppose that '*Cb*' is a sentence of *L* and that the relevant semantic facts about it are given in (9):

(9) a. '*b*' refers (in *L*) to Boston.
b. '*C*' applies (in *L*) to cities (and only cities).
c. '*Cb*' is true (in *L*) iff Boston is a city.

If Tarski's definitions really specify the physicalistic content of semantic notions, then, in each case, we ought to be able to substitute the physicalistic definiens for the semantic definiendum without changing the physical fact thereby specified. Performing this substitution and simplifying results, we obtain

(10) a. '*b*' = '*b*' and Boston = Boston.
b. For all objects *o*, '*C*' = '*C*' and *o* is a city, iff *o* is a city.
c. '*Cb*' = '*Cb*' and there is an object *o* such that *o* = Boston and *o* is a city, iff Boston is a city.

But there is a problem in identifying these facts with those in (9). As Field points out, it is natural to suppose that the expressions of a language have semantic properties only in virtue of the ways they are used by speakers. Thus, he holds that the facts given in (9) wouldn't have obtained if speakers' linguistic behavior had been different.[12] Since the facts in (10) are not speaker-dependent in this way, Field concludes that they are not semantic facts and that Tarski's attempted reduction fails. Tarski's truth predicate is both physicalistic and coextensive with 'true in *L*'; but it is not, according to Field, a physicalistic conception of truth.

On Field's view, Tarski's truth characterization inherits its inadequacy as a reduction from the pseudo-reductions that constitute its base clauses. Thus, Field's strategy for solving the problem is to

[12] I use the phrase 'linguistic behavior' in a broad sense to include all facts about speakers relating to their use of language.

provide genuine reductions for the notions of primitive denotation, on something like the model of the causal theory of reference. The picture that emerges from his discussion is one in which an adequate definition of truth is a two-stage affair. Stage 1 is Tarski's reduction of truth to primitive denotation. Stage 2 is the imagined causal theory-like reduction of the notions of a name referring to, and a predicate applying to, an object in a language.[13] If the physical facts that determine denotation in one language do so in all, then these relations will hold between expressions and objects, for variable '*L*'. When logical vocabulary and syntax is kept fixed, the result is a notion of truth that is not language-specific, but is itself defined for variable '*L*'.

Although the resulting picture appears rosy, there are several problems with it. One concerns reference to abstract objects, for which a causal account seems problematic. Another involves Quinean worries about ontological relativity and referential indeterminacy. These, of course, are obstacles to a physicalistic reduction of primitive denotation. However, there are other difficulties which become clear when one notices that Field has understated his objection to Tarski. If the alleged dependence of semantic facts on facts about speakers shows that Tarski has not reduced primitive denotation to physical facts, then the very same point shows that he has not reduced truth to primitive denotation.

This can be seen by considering a pair of elementary examples. Imagine two languages, L_1 and L_2, which are identical except that in L_1 the predicate '*R*' applies to round things, whereas in L_2 it applies to red things. Owing to this difference, certain sentences will have different truth conditions in the two languages.

(11) a. '*Re*' is true in L_1 iff the earth is round.
b. '*Re*' is true in L_2 iff the earth is red.

Under Tarski's original definition, this difference will be traceable to the base clauses of the respective truth definitions, where the applications of predicates are simply listed.

Field's objection to this is that although Tarski's definitions correctly *report* that '*R*' applies to different things in the two languages, they don't *explain* how this difference arises from the way in which speakers of the two languages use the predicate. What Field fails to point out is that exactly the same objection can be

[13] Field also includes the notion of a function sign being fulfilled by a pair of objects. In the interest of simplicity, I am ignoring this.

brought against Tarski's treatment of logical vocabulary and syntax in the recursive part of his definition.

This time let L_1 and L_2 be identical except for their treatment '$\vee$'.

(12) a. A formula $\ulcorner(A \vee B)\urcorner$ is true in L_1 (with respect to a sequence s) iff A is true in L_1 (with respect to s) or B is true in L_1 (with respect to s).
b. A formula $\ulcorner(A \vee B)\urcorner$ is true in L_2 (with respect to a sequence s) iff A is true in L_2 (with respect to s) and B is true in L_2 (with respect to s).

Owing to this difference, sentences containing '$\vee$' will have different truth conditions in the two languages. In order to satisfy Field's requirements on reduction, it is not enough for a truth characterization to report such differences. Rather, such differences must be explained in terms of the manner in which speakers of the two languages treat '$\vee$'.[14] Since Tarski's truth definitions don't say anything about this, their recursive clauses should be just as objectionable to the physicalist as the base clauses.

This means that Field's strategy of achieving a genuine reduction of truth by supplementing Tarski with nontrivial definitions of primitive denotation cannot succeed. The reason it can't is that, given Field's strictures on reduction, Tarski has not reduced truth (for standard first-order languages) to primitive denotation. At best he has reduced it to the class of semantic primitives listed in (13):[15]

(13) the notion of a name referring to an object
the notion of a predicate applying to objects
the notion of a formula being the application of an n-place predicate P to an n-tuple of terms $t_1 \ldots t_n$
the notion of a formula A being a negation of a formula B
the notion of a formula A being a disjunction of formulas B and C
the notion of a formula A being an existential generalization of a formula B with respect to a variable u and a domain D of objects

[14] Presumably, speakers of L_1 differ in some way from speakers of L_2 regarding their beliefs, intentions, attitudes, brain states, or conditioned responses involving '$\vee$'.

[15] Field partially anticipates this point in footnotes 5 and 10 of his paper. In fn 5 he notes that in model theory quantifiers are given an "unusual" semantics in which they range over the members of some specified set, rather than over all (actually existing) things. In such a case, Field claims, Tarski has reduced truth to primitive denotation, plus the notion of the range of the quantifiers. (For Tarski this constituted the usual case, since it is only when the range of quantifiers is restricted that explicit truth definitions can be given—for languages with a certain minimal richness.)

In fn 10 Field notes, without specifying, the existence of problems that must be faced when the definition of truth is generalized so as not to contain a particular logical vocabulary.

This way of looking at things requires a restatement of every clause in Tarski's truth definition. For example, the recursive clause for negation, which had been given by (14a), is now given by (14b).

(14) a. If $A = \ulcorner -B \urcorner$, then A is true in L (with respect to a sequence s) iff B is not true in L (with respect to s).
b. If A is a negation of a formula B, then A is true in L (with respect to a sequence s) iff B is not true in L (with respect to s).

The resulting abstraction extends the generality of the truth definition to classes of first-order languages that differ arbitrarily in syntax, plus logical and nonlogical vocabulary.

Although this generality is appealing, it has a price. Whereas the original definitions simply stipulated that $\ulcorner -A \urcorner$ is a negation, $\ulcorner A \vee B \urcorner$ is a disjunction, and $\ulcorner \exists x A x \urcorner$ is an existential generalization over a range D of objects, the revised definition doesn't provide a clue about which formulas fall into these categories. Moreover, Field's physicalist now has to provide reductions of each of these semantic notions.

How might this be done? We are accustomed either to using truth to explain the logical notions or to taking them as primitive, while stipulating that certain symbols are to count as instances of them. Neither of these policies is open to Field. He cannot characterize negation as a symbol that attaches to a formula to form a new formula that is true (with respect to a sequence) iff the original is false (with respect to the sequence); for that would make the reduction of truth to the notions in (13) circular. Nor can he take negation to be primitive and stipulate that $\ulcorner -s \urcorner$ is to be the negation of S; for that would fail to give the facts about speakers that explain the semantic properties of $\ulcorner -s \urcorner$. Although there are alternative approaches, none that I know of is clearly successful.[16] For example, in *The Roots of Reference*[17] Quine attempts to characterize truth-functional operators in terms of community-wide dispositions to assent and dissent. He ends up concluding that indeterminacy between classical and intuitionist construals of the connectives is inevitable. Although I do not accept Quine's argument for this,[18] I do think that the task confronting Field's physicalist is nontrivial. The problems involved in reducing primitive denotation to

[16] The most interesting, in my opinion, is briefly sketched in Gilbert Harman, "Beliefs and Concepts," *Proceedings of the Philosophy of Science Association*, II (1982).

[17] LaSalle, Ill.: Open Court, 1974.

[18] For an excellent critique of Quine, see Alan Berger, "Quine on 'Alternative Logics' and Verdict Tables," this JOURNAL, LXXVII, 5 (May 1980): 259-277.

physical facts are hard enough; adding the logical notions makes the job that much harder.

As I have stressed, the source of this difficulty is the demand that semantic facts be supervenient on physical facts about speakers. In effect, this demand limits adequate definitions to those which legitimate substitution for semantic notions in contexts like (15) and (16).

(15) If *L*-speakers had behaved differently (or been differently constituted), then '*b*' wouldn't have referred (in *L*) to Boston, and '*C*' wouldn't have applied (in *L*) to cities, and '$Cb \vee Ca$' wouldn't have been true (in *L*) iff Boston was a city or Arizona was a city.

(16) The fact that *L*-speakers behave as they do (and are constituted as they are) explains why '*b*' refers (in *L*) to Boston, etc.

Field's critique of Tarski is based on the conviction that there ought to be a way of spelling out (15) and (16) so that they come out true when physicalistic substitutes replace semantic terms and their initial clauses are construed as expressing contingent physical possibilities.[19] As we have seen, Tarski's definition doesn't have this character.

III

It is helpful in understanding the issues at stake to compare this criticism of Tarski to a parallel objection. Whereas Field's critique is based upon a view about the relationship between speakers and semantic properties like truth, the parallel objection is based on a view about the relationship between meaning and truth. It is widely held that the meaning of a sentence is closely related to its truth conditions and that knowledge of the one constrains knowledge of the other. Thus, many philosophers would accept arbitrary instances of (17) and (18):

(17) If '*S*' had meant in *L* that *p*, then '*S*' would have been true in *L* iff *p*.

(18) If *x* knows that it is not the case that '*S*' is true in *L* iff *p*, then *x* knows (or has sufficient grounds for concluding) that '*S*' does not mean in *L* that *p*.[20]

[19] In stating this requirement in terms of the replacement of a semantic term by its physicalistic definiens, I have tacitly relied on Tarski's emphasis on explicit definition. However, I don't think the philosophical point of the requirement depends on this. In cases in which only an axiomatic treatment is possible, Field could require that the axioms governing 'true', together with empirical facts about speakers and their environments, have statements of type (15) and (16) as consequences.

[20] (18) is considerably weaker than the claim that knowledge of truth conditions is sufficient for knowledge of meaning. (18) says only that knowledge of truth condi-

A natural demand growing out of this view is that substituting an adequate explication for 'true in *L*' in (17) and (18) should result in true sentences with contingent antecedents.[21]

As before, it is obvious that Tarski's definition does not satisfy this demand. For example, let '*Ws*' be a sentence of *L* meaning that snow is white. Using Tarski's definition of truth, we can produce the following counterparts of (17) and (18):[22]

(17$_T$) If '*Ws*' had meant in *L* that snow is black, then it would have been the case that snow was white iff snow was black.

(18$_T$) If *x* knows that it is not the case that snow is white iff snow is black, then *x* knows (or has sufficient grounds for concluding) that '*Ws*' does not mean in *L* that snow is black.

These are clearly not what the defender of (17) and (18) has in mind. The reason they aren't is that Tarski's set-theoretic truth predicate doesn't impose any conditions on the meanings of the sentences to which it applies. To be sure, Tarski wouldn't count any predicate *T* as a truth predicate unless $\ulcorner\alpha$ is $T\urcorner$ were materially

tions is capable of providing some information about meaning. In effect, it says that even if knowledge that

(i) '*S*' is true in *L* iff *q*.

is not sufficient for knowing that

(ii) '*S*' means in *L* that *q*.

it should be sufficient for knowing that

(iii) '*S*' does not mean in *L* that *p*.

(where the sentences replacing '*p*' and '*q*' are obviously incompatible).

[21] Although the contexts in question are intensional, this demand does not require that an adequate explicatum for the pretheoretic notion of truth be intensionally equivalent to an ordinary, pretheoretic truth predicate. Rather, it requires that all legitimate theoretical purposes served by the explicandum (truth), be equally well served by the explicatum. For example, if knowledge of that expressed by

(i) . . . is true . . .

is used to help explain the nature of some capacity (say, the capacity to understand sentences), then knowledge of that expressed by

(ii) . . . T . . .

(where the explicatum T is substituted for 'is true') should be sufficient for the same purpose. An explication that meets this requirement of theoretical productivity will allow the explicandum to be eliminated from one's total scientific and philosophical theory without loss of explanatory power. Thus, substitution of explicatum for explicandum in intensional contexts contained in one's total explanatory theory must be countenanced, even if such substitution is not always countenanced in ordinary discourse.

The qualification in fn 19 above regarding substitution and explicit definition also applies here.

[22] (17$_T$) and (18$_T$) are simplifications of the sentences that would result from substituting Tarski's explicatum [the right-hand side of (T') in section II] for 'true in *L*' in (17) and (18). The simplifications are based on the fact that, where T is Tarski's explicatum, $\ulcorner$'Snow is white' is T$\urcorner$ and 'Snow is white' are necessarily equivalent (in the presence of elementary set theory). In light of this equivalence, replacing one with the other should not affect the philosophical issues at stake in (17) and (18).

equivalent to any metalanguage paraphrase of the object-language sentence named by α. On the basis of this, one might interpret Tarski as implicitly supposing that instances of (19) are necessary or apriori.

(19) If 'T' is a truth predicate for L, and 'S' means in L that p, then 'S' is T iff p.

However, this is quite different from maintaining that if 'T' in (20) is replaced with a truth predicate for L, then the resulting instances of the schema will be necessary or apriori:

(20) If 'S' means in L that p, then 'S' is T iff p.

It is this that the advocate of (17) and (18) demands and that Tarski appears not to provide.[23]

[23] Hilary Putnam has used a version of the argument involving (17)/(17_T) against Tarski (in a lecture at Princeton, fall 1982). Michael Dummett has used a version of the argument involving (18)/(18_T) against Tarski [in the preface to *Truth and Other Enigmas* (Cambridge, Mass.: Harvard, 1978), and in "Truth," reprinted there].

The arguments given above are intended as stand-ins for a variety of related arguments, all designed to show that Tarski's notion of truth has nothing to do with semantic interpretation or understanding. For example, it is probably best to understand Davidson not as attempting to analyze meaning in terms of truth, but rather as eliminating the notion of meaning in favor of the notion of truth. Since (18) utilizes the notion of meaning, a defender of the Davidson of "Truth and Meaning" might want to trade it for something like (i):

(i) If x knows that which is expressed by the relevant instance of
'S' is true in L iff p
for each sentence of L, then x is a competent speaker of L.

If 'true in L' is understood as short for the definiens provided by Tarski, (i) is as absurd as (18_T).

Just this sort of absurdity is present in familiar and often repeated remarks like the following (which would allow Tarski's definiens to be the central notion in a theory of meaning):

> (T) s is T if and only if p
>
> What we require of a theory of meaning for a language L is that without appeal to any (further) semantical notions it place enough restrictions on the predicate 'is T' to entail all sentences got from schema T when 's' is replaced by a structural description of a sentence of L and 'p' by that sentence. . . .
>
> It is worth emphasizing that the concept of truth played no ostensible role in stating our original problem. That problem, upon refinement, led to the view that an adequate theory of meaning must characterize a predicate meeting certain conditions. It was in the nature of a discovery that such a predicate would apply exactly to the true sentences. I hope that what I am doing may be described in part as defending the philosophical importance of Tarski's semantical concept of truth [Donald Davidson, "Truth and Meaning," in J. F. Rosenberg and C. Travis, eds., *Readings in the Philosophy of Language* (Englewood Cliffs, N.J.: Prentice-Hall, 1971), pp. 455/56].

Earlier statements of essentially the same absurdity can be found in Rudolf Carnap's *Meaning and Necessity* (Chicago: University Press, 1947), pp. 5/6; and in section 7 of his *Introduction to Semantics* (Cambridge, Mass.: Harvard, 1943).

IV

We have, then, two major objections to Tarski. Field demands that semantic properties be dependent on speakers in a way in which Tarski's substitutes are not. A familiar sort of semantic theorist demands that meaning and truth conditions be contingent, but analytically connected, properties of a sentence in a manner incompatible with Tarski. The only way to defend Tarski's philosophical interpretation of his work is to reject these demands.

Although this might initially seem to be a desperate strategy, it is not. Think of a standard first-order language L as a triple $\langle S_L, D_L, F_L \rangle$, where S_L is a family of sets representing the various categories of well-formed expressions of L; D_L is a domain of objects; and F_L is a function that assigns objects in D_L to the names of L, subsets of the domain to one-place predicates of L, and so on.[24] Let J be a class of such languages. Truth can now be defined in nonsemantic terms for variable 'L' in J in a straightforward Tarskian fashion. The only significant change from before is that the notions of primitive denotation are no longer given language-specific list definitions, but rather are defined for variable 'L' using the "interpretation" functions built into the languages. In particular, a name n refers to an object o in a language L iff $F_L(n) = o$.[25] The resulting truth predicate is just what is needed for metatheoretical studies of the nature, structure, and scope of a wide variety of theories.

What the truth definition does not do is tell us anything about the speakers of the languages to which it applies. On this conception, languages are abstract objects, which can be thought of as bearing their semantic properties essentially. There is no possibility that expressions of a language might have denoted something other than what they do denote; or that the sentences of a language might have had different truth conditions. Any variation in semantic properties (across worlds) is a variation in languages. Thus, semantic properties aren't contingent on anything, let alone speaker behavior.

[24] This sort of construction is familiar from model theory. However, its use here is different from model-theoretic treatments. Here we are not defining *truth in L relative to a model,* but rather *truth in L* (simpliciter) for an enriched conception of a language. This way of looking at things was suggested to me from two sources: David Lewis's "Languages and Language," in K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science,* VII (Minneapolis: U of Minnesota Press, 1975), pp. 3-35; and one of Saul Kripke's seminars on truth, Princeton, 1982.

[25] Note, F_L is a purely mathematical object—a set of pairs, if you like. Thus, it does not incorporate any undefined semantic notions. This was one of the points noted by Kripke in the seminar mentioned in fn 24.

What is contingent on speaker behavior is which language a person or population speaks and which expression a given utterance is an utterance of. Let L_1 and L_2 be two languages in J which are identical except for the interpretations of certain nonlogical vocabulary—perhaps the color words in L_1 are shape words in L_2. We can easily imagine a situation in which it is correct to characterize L_1, rather than L_2, as the language of a given population. To ask what such a characterization amounts to, and what would justify it, is to ask not a semantic question about the languages, but a pragmatic question about their relation to speakers.

Although Tarski had nothing to say about this relation, other philosophers have. David Lewis, using a different, but equally abstract, conception of language has proposed (*op. cit.*) an analysis in terms of a convention of truthfulness and trust. Discussions of what Donald Davidson calls "radical interpretation" can also be reconstructed as dealing with this issue. For physicalists, the interesting question is whether any purely physical explication can be given. If so, then the physicalist can accept both semantic notions that apply to sentences and those which apply to utterances. If not, then either the latter or physicalism itself must be abandoned.

It is interesting to note that much of Hartry Field's concern is with the semantic properties of utterances rather than sentences. In describing the physicalist's position he says:

> People utter the sounds 'Electrons have rest mass but photons don't' . . ., and we apply the word 'true' to their utterances. We don't want to say that it is a primitive and inexplicable fact about those utterances that they are true, a fact that cannot be explicated in non-semantic terms; this is as unattractive to a physicalist as supposing that it is a primitive and inexplicable fact about an organism at a certain time that it is in pain (*op. cit.*, 359).

In effect, Field criticizes Tarski for not providing a physicalistically acceptable truth predicate of utterances. But Tarski wasn't concerned with utterances. Thus, confronted with the question

(i) In virtue of what are certain sounds utterances which are true in L?

Tarski's response ought to be to break it up into two subsidiary questions:

(ii) In virtue of what are certain sounds utterances in L of its sentences?
(iii) In virtue of what are sentences of L true (in L)?

Whereas Tarski answered the second question, the first was no part of his task.

It is hard to see how Field himself could avoid this division of labor. At one point he suggests that in order to handle ambiguous and indexical expressions, truth definitions should be formulated in terms of tokens rather than types (351-353). The idea is that utterances are contextually disambiguated and that semantic notions should apply to unambiguous entities. This means that all clauses in a truth definition must be formulated as applying to tokens. To this end, Field reformulates the clause for negation as (21):

(21) A token of $\ulcorner -e \urcorner$ is true (with respect to a sequence) iff the token of e that it contains is not true (with respect to the sequence) (352).

However, this won't do. As I indicated earlier, Field can't accept any truth definition in which a certain syntactic form is simply stipulated to be a negation; for to do this would be to fail to explicate the facts about speakers in virtue of which negative constructions have the semantic properties they do. Instead, (21) must be replaced with something along the lines of (22).

(22) A token of a formula A, which is a negation of a formula B, is true (with respect to a sequence) iff some designated token of B is not true (with respect to the sequence).

But now there is a problem. Even if the notion of a formula A being a negation of a formula B can be given a physicalistic definition in terms of the behavior of speakers, there is no clear way of specifying the relevant token of B needed in (22); indeed, there is no way of ensuring that it will exist.

If we could count on utterances of negative sentences always containing, as proper parts, utterances of the sentences they are negations of, then the problem would not arise. Although this is a feature of certain artificial languages, it is not a characteristic of natural languages actually spoken by people. In order to avoid arbitrarily restricting truth definitions to (utterances involving) this subset of artificial languages, we need some way of eliminating undue dependence on empirically unreliable tokens. The most straightforward way of doing this is to define truth for types, thereby acknowledging the theoretical division of labor I have attributed to Tarski.[26] Once this is done, the physicalist is free to accept Tarski-like truth definitions applying to sentences, while leaving it open

[26] Ambiguity can then be treated as a case of homonomy. For example, instead of thinking that English contains a single (ambiguous) word type 'bank', one can take English to contain two different words, 'bank_1' and 'bank_2', whose tokens are phonologically identical. The contextual factors that Field relies on to disambiguate tokens can then be thought of as determining whether particular utterances are tokens of the type 'bank_1' or the type 'bank_2'.

whether the pragmatic relations between languages, expressions, speakers, and utterances are purely physicalistic.[27]

It should be emphasized that although the linguistic threat to physicalism has been moved from semantics to pragmatics, it is still a serious one. It is by no means evident that physicalistic reductions of the crucial relations can be given. One physicalist who seems to think they cannot be given is Quine. Although he doesn't conceptualize matters in just the way that I have, it is illuminating to interpret him as accepting Tarski's semantic definitions while rejecting any physicalistic reduction of the pragmatic notions. On this interpretation, there is no indeterminacy about the claim that 'rabbit' refers to rabbits in a certain specified language, call it "English", or about the claim that 'gavagai' refers to rabbits in another language, call it "Junglese". What is indeterminate is whether I speak English, as opposed to some related rabbit-stage language, and whether the native speaks Junglese, as opposed to some similar counterpart.

The upshot of this is that it is all right for a Quinean physicalist to *use* a Tarskian language to describe the world, and even to attribute Tarskian semantic properties to expressions in that language. What he cannot do is identify the language he is using. When it comes to describing linguistic behavior—even one's own—identifiable Tarskian languages are excluded in favor of dispositions to verbal behavior. The strain in this position is a measure of the challenge that language use presents to physicalism. What is not problematic is the physicalist's acceptance of Tarski.

V

This discussion illustrates a general strategy for answering Tarski's critics. Field's objection was that Tarski's semantic properties are not dependent on facts about speakers. The Tarskian reply is that nothing is lost by thinking of semantics abstractly and relegating the interpretation of speakers' behavior to pragmatics. In so doing, one gains the advantages of a truth predicate for metatheoretical discussions, while retaining the ability to raise deep philosophical problems in other areas.

As I pointed out earlier, Field's is not the only objection to Tarski. Any theory of semantic competence that makes knowledge of truth conditions the central notion implicitly rejects Tarski's claim to have provided a notion of truth adequate for all theoreti-

[27] Acknowledging the need to formulate truth definitions in terms of types does not force one to think of the semantic properties of sentences as invariant from world to world and not dependent on the properties of tokens. However, it does make this a natural position.

cal purposes. The defense against this objection is that such theories are flawed in any case.

The problem with these theories lies in specifying what truth conditions are in such a way that knowledge of them is necessary and sufficient for understanding. If we assume that truth conditions involve the notion of truth, then it is natural to suppose that they are given by T-sentences of the form (23):

(23) '*S*' is true in $L \equiv P$.

(Instances are formed by replacing '*P*' with a sentence that means the same as the sentence replacing '*S*'.) However, it is easy to show that knowing the propositions expressed by T-sentences is neither necessary nor sufficient for understanding meaning (where 'true' is taken to be a non-Tarskian primitive and '≡' represents either material or necessary equivalence). Thus it is not obvious that what one knows when one understands a language involves the notion of truth at all. If it doesn't, it may be that nothing is lost by adopting a Tarski-like explication of truth together with an independent account of semantic competence.

Although I won't try to show it here, I think that this is the right approach for both truth and semantic competence. This does not mean that Tarski's semantic predicates really are adequate for all theoretical purposes. Saul Kripke's theory of truth is a genuine advance on Tarski's treatment of the liar.[28] In addition, semantic predicates for richer languages, as well as for propositions, are needed. What does seem right about Tarski's approach is its deflationist character. Theories of truth for sentence types need not specify the facts about speakers in virtue of which their utterances have content; nor should such theories be seen as issuing in theorems knowledge of which is necessary and sufficient for semantic competence. Instead, theories, or definitions, of truth should provide accounts of the content of familiar truth predications, while resolving the semantic paradoxes (and their propositional variants). Beyond this, and the attendant dissolution of confusions, it is best not to expect too much. Truth is a useful notion, but it is not the key to what there is, or to how we represent the world to ourselves through language.

SCOTT SOAMES

Princeton University

[28] "Outline of a Theory of Truth," this JOURNAL, LXXII, 19 (Nov. 6, 1975): 690-716.