Truth III: Paradox & Tarski's Solution

THE LIAR

(1) Sentence (1) is false

That gives us a puzzle. But Tarski (and many of his contemporaries) thought it was much worse than that: they were concerned that a language containing a truth predicate was actually *inconsistent*. How can a *language* be inconsistent? Consider Prior's connective 'tonk' which has the introduction rule of disjunction, and the elimination rule of conjunction. So from *p*, one can go to *p tonk* ~*p*, and from there to ~*p*. Adding 'tonk' to the language makes the language inconsistent, in that one can immediately derive a contradiction: from *p,* ~*p* follows. Tarski thought that the truth predicate did much the same thing, though of course we need to add some rules to see this. Rather than doing this with introduction and elimination rules, Tarski did it with a biconditional, or more accurately a schema for a biconditional. He thought that an understanding of the truth predicate in ordinary English commits one to accepting each instance of:

The T-schema: 'p' is true iff p

where 'p' is replaced with a sentence of English in each occurrence, so in each case the first instance is quoted, and the second is used. (Actually Tarski did this with structural descriptions rather than quotation, but this won't matter for our purposes.) Then we have the following argument

P1: Sentence (1) is 'Sentence (1) is false'. (from (1))
P2: 'Sentence (1) is false' is true iff sentence (1) is false. (instance of the *T-schema*)
C1: Sentence (1) is true iff sentence (1) is false. (instantiating identity from P1 into left side of P2)
C2: Sentence (1) is true and sentence (1) is false. (Assuming that (1) is either true or false, and deriving consequences from C1.)

Which assumptions did we need to get that? Not many: just the universal applicability of the T-schema, which is exactly what a grasp of the notion of truth seemed to require, elementary application of a principle of substitution, and bivalence: the claim that every sentence is either true or false.

Perhaps we could reject bivalence and say that (1) is neither true nor false (though that might be a bad way to put it for reasons we'll look at later). Perhaps it has some third status. Let's just say it's *bad*. But then we get the Strengthened Liar:

(2) Sentence (2) is either false or bad.

We can now run though the same reasoning, provided that, in place of the principle of bivalence, we have the principle that every sentence is either true, false, or bad.

P1: Sentence (2) is 'Sentence (2) is either false or bad'. (from (2))
P2: 'Sentence (2) is either false or bad' is true iff sentence (2) is false or bad. (instance of the *T-schema*)
C1: Sentence (2) is true iff sentence (2) is either false or bad (instantiating identity from P1 into left side of P2)
C2: Sentence (2) is true and sentence (2) is either false or bad (Assuming that (2) is true or either false or bad, and deriving consequences from C1.)

Languages and Metalanguages

Tarski thought that ordinary languages like English are hopeless. The only hope is to construct new formal languages which don't risk paradox. There his central device for avoiding paradox was to make a distinction between the object language—the language of whose sentences truth is predicated—and the metalanguage—the language in which the predication is made. The truth predicate is in the metalanguage, not in the object language. Object language/metalanguage suggests a simple binary division. But it is really a relative notion. So we can think in terms of a hierarchy of languages, each of which contains the truth predicates of the languages below it, but not its own.

Traditionally liar sentences involve self reference. e.g.

(A) This sentence is not true

But truth must be relative to a language. Suppose that (A) is in the language L_1. So, trying to capture the intuitive sense of (A) we might reformulate it as:

(A*) This sentence is not true in L_1

But the truth predicate for L_1 is not part of L_1. So sentence (A*) is simply ungrammatical. Alternatively, the truth predicate might be for the language below it:

(A**) This sentence is not true in L_0

But now (A**) is true, i.e. true in L_1, since it is not in L_0, and so *a fortiori* is not one of the true sentences of L_0.

Paradoxical sentences need not involve immediate self-reference; and Tarski's approach can certainly sometimes help here. Consider

(B) Sentence (C) is true
(C) Sentence (B) is not true.

But it can't be that each of these sentences is in the metalanguage for the other. So at least one of them must be ungrammatical. For a harder case though, consider Yablo's paradox, the infinite sequence of sentences:

S_0 All later sentences in this sequence are not true
S_1 All later sentences in this sequence are not true
S_2 All later sentences in this sequence are not true ...

Clearly this gives rise to paradox if all of the sentences contain the same truth predicate. (Suppose the first sentence is true; then the second must be false, but since it is entailed by the first, it can't be. Suppose then that the first sentence is false; then at least one of the sentences below it must be true; but now reapply the earlier reasoning to that sentence.) What happens if we use a hierarchy of truth predicates in Tarski's fashion. Suppose that, going down the list, each sentence is in a language one level lower than the one before, and each truth predicate is correspondingly for a language one level lower than the one it is formulated in. Can they now get a consistent assignment of truth values? Apparently not for the same reasons as before. Tarski doesn't just need a hierarchy of truth predicates; he also needs there to be a lowest level. (See Yablo, 'Circularity and Paradox')

THE DEFINITION OF TRUTH

Tarski held that a good definition of truth must meet two conditions. First, it must be 'materially adequate': it must generate, in the metalanguage L_1 each instance of the T-schema for the relevant language L_0:

X is true in L_0 iff P

where 'X' is replaced by a name of a sentence in L_0, and 'P' is replaced by a translation of that sentence into the metalanguage. The metalanguage may be an extension of the object language (it may be the object language together with the truth predicate of that language) but it need not be. So, for instance, we might get T-sentences like:

'La neige est blanche' is true in $French_0$ iff snow is white.

This is clearly not trivial. Similarly, the name of the sentence may be formed by placing quotation marks around the sentence, but it need not be.

Secondly, the definition must be 'formally correct'. This is a harder notion to get clear on. For a start the definition must contain, on the right hand side, no undefined semantic notions. If truth is a dubious concept, the only way to legitimate it is to define it in less dubious terms. For Tarski, who defined himself as a physicalist, this meant that ultimately truth should be defined in purely physical terms. If there were only a finite number of sentences for which the truth predicate were defined, this would be fairly straightforward. Consider a language L_A that contained only the sentences:

snow is white
grass is green
snow is green
grass is white

then we could define truth for that language as follows:

S is true in L_A iff

S = 'snow is white' and snow is white
or S = 'grass is green' and grass is green
or S = 'snow is green' and snow is green
or S = 'grass is white' and grass is white

The right hand side of the definition doesn't contain any semantic terminology: it doesn't contain any mention of truth or reference or anything like that; it refers only to notions of snow, grass, white and green. (These may not be physicalistically acceptable; but if not, then they should be reduced to physicalistically acceptable notions, or banished from the language; at any rate, it isn't the notion of truth that is causing the problem.) Of course, it doesn't tell you *which* of these sentences is true; but a theory of truth isn't meant to teach you what colours things are. And although it may look trivial, that is only because the metalanguage and the object language are the same. Suppose that the object language was a fragment of Russian, and the metalanguage English. Then this would tell you something about Russian.

But our languages aren't finite: whilst the basic elements are, they can be used to generate an infinite number of sentences. So no list will do the job. Moreover, lists leave out important generalizations. For instance, we want our theory of truth to register the fact that a conjunction will be true iff both of the conjuncts are true, that the predicate 'is white' applies to white things.

A DEFINITION OF TRUTH FOR AN INFINITE LANGUAGE L_B

Non-logical vocabulary
The one-place predicates 'is white' and 'is green'
The terms 'snow' and 'grass'

Logical vocabulary
'&'

Sentences

If P is a predicate and t is a term, ‹Pt› is a sentence

If A and B are sentences, ‹A & B› is a sentence

Denotation
'snow' denotes snow
'grass' denotes grass

Application
The predicate 'is white' applies to an object iff it is white

The predicate 'is green' applies to an object iff it is green

Truth

An atomic sentence ‹Pa› is true in L_B iff the predicate ‹P› applies to the object denoted by ‹a›

A sentence ‹A & B› is true in L_B iff A is true in L_B and B is true in L_B

Tarski's definition was for a language containing quantification. That makes things considerably more complicated, but the basic ideas remain the same.

A SIMPLE TARSKI-STYLE DEFINITION OF TRUTH FOR A QUANTIFIED LANGUAGE L_C

This account (broadly following Soames in *Understanding Truth*) is different in its details from Tarski's own approach, which is rather harder to follow. But in its essence it gives a good idea.

Non-logical vocabulary
A two-place predicate '='
A one place function symbol 'S'
A two place function symbol '+'
A name '0'

Logical vocabulary

'&', '~', '∃', together with infinitely many variables v_1, v_2, v_3, ...

Terms
Names and variables are terms

If t is a term, ‹St› is a term

If t_1 and t_2 are terms, ‹A + B› is a term

Formulas

If t_1 and t_2 are terms, ‹$t_1 = t_2$› is an atomic formula

If A and B are formulas, ‹A & B› is a formula

If A is a formula, ‹ ~ A› is a formula

If A is a formula and v is a variable, ‹∃vA› is a formula, with a quantifier containing v.

Sentences
A sentence is a formula containing no free occurrence of a variable, where a free occurrence is one that is not within the smallest formula following a quantifier that contains that variable.

Denotation
A variable free term t denotes the number n iff
(i) t is '0' and n is zero
(ii) t is ‹Sa› for some term a, and n is the successor of the number denoted by a

(iii) t is ‹a + b› for terms a and b, and n is the sum of the numbers denoted by a and b.

Application
The predicate '=' applies to two numbers n and m iff n is the same number as m.

Truth
An atomic sentence ‹a = b› is true in L_C iff the predicate '=' applies to the two numbers denoted by a and b.

A sentence ‹~A› is true in L_C iff A is not true

A sentence ‹A & B› is true in L_C iff A is true and B is true

A sentence ‹∃vA› is true in L_C iff there is some true sentence A* that is obtained by erasing '‹∃v›' and replacing every free occurrence of v in A with the same variable free term.

Evaluation

These definitions meet the material adequacy condition: it is clear that every T-sentence for the language L_A and L_B can be derived from this definition; this is a bit less obvious for L_C (see Soames p. 73 for a discussion of this.)

Do they meet the formal correctness condition? That is controversial. Certainly truth is defined in terms of denotation and application, and they do not have ineliminable semantic vocabulary on the right hand sides of their definitions. But they are defined by means of a list (this is especially clear in the definition of L_B).

Is that to understand the notion of denotation? There is a parallel problem for the idea of application, which is also defined by means of a list. This is the heart of Field's complaint. His parallel: suppose we had an account of valence that said:

> E has valence n iff E is potassium and n = +1, … or E is sulphur and n is -2 etc.

We wouldn't think that was adequate to explain valence. Perhaps Tarski's definitions, if they are to meet his notion of correctness, require supplementation; but perhaps we should drop this requirement.

A second worry: what does this account tell us about truth in natural language? Tarski thought that he had captured the intuitive idea of truth (with a little bit of experimental philosophy), but he insisted that his notion only applies to artificial hierarchical languages, to avoid the risk of paradox. If it is the same notion, shouldn't it apply to natural languages? If it doesn't, is it the same notion?