Seminar 5: Tarski on Truth

In the 1930s, Alfred Tarski published two articles that soon became classics --"The Concept of Truth in Formalized Languages," in which he defines truth for formal languages and "On the Concept of Logical Consequence," in which he uses the definition of truth to provide the basis for the now standard model-theoretic, definition of logical consequence (and related notions).[1] His interest in truth arose from an interest in the expressive power of mathematical languages and theories, including the *definability* of metatheoretical notions in them. To say that a set s is *definable* in a language L is to say that there is some formula that is *true* of all and only the members of s. Since there were then no definitions of these concepts in terms of the concepts of logic and set theory, mathematicians regarded them with suspicion. Calling them "semantic" didn't help, in part because of their role in paradoxes like the Liar, and in part because it wasn't obvious how the concept *true sentence* could be treated with the same rigor and formality as *proof* and *provable sentence* (in a given system). Nevertheless, Tarski believed truth and definability to be essential to metamathematics, which led him to try to make them respectable.

Truth, Paradox, and Inconsistency

His first task was to insulate the truth predicate he wished to define from doubts stemming from the Liar paradox. His strategy was to identify features of the ordinary truth predicate responsible for the paradox, and to exclude them from his definition. The English predicate *is true* can be applied not only to sentences, but also to statements/propositions, and sentences. It correctly applies to a sentence S only if S is used to make a statement (express a proposition) that is true. Although *is a true sentence* is an English predicate, its application is universal. It applies to any sentence of any language that is used to make a true statement (or express a true proposition), and to only such sentences. The related predicate "is true" is capable of applying to any statement/proposition one might make or express, and to any sentence used to make or express it. Since "is true" is itself used to make or express statements/propositions, it is applicable to the statements/propositions it is used to make or express, and to the sentences used in doing so. This leads to the Liar paradox.

1. Sentence (1) is not true.

The expression "sentence (1)" is here used as an abbreviation of the singular term "the first numbered example on this page." So understood, (1) is a meaningful English sentence, as is shown by the fact that someone not familiar with this page would easily understand it. What (1) says would have been true if the first numbered example here had been "There are no even prime numbers". Since sentence (1) is used to say something that would have been true had certain facts obtained, it must be meaningful. It is paradoxical because a contradiction can be derived from seemingly incontrovertible assumptions about it.

The Liar

P1. 'Sentence (1) is not true' is a true iff sentence (1) is not true.
P2. Sentence (1) = 'Sentence (1) is not true'.
C1. Sentence (1) is true iff sentence (1) is not true.
C2. Sentence (1) is true and sentence (1) is not true.

C1 is derived by substituting 'Sentence (1)' for the quote-name 'Sentence (1) is not true' in P1 on the basis of P2. Given that the linguistic context *x is true iff sentence (1) is not true* is

[1] Alfred Tarski (1935 [1983]). "The Concept of Truth in Formalized Languages." Trans. of Tarski (1935) by J. H. Woodger. In Tarski (1983), *Logic, Semantics, Metamathematics*, 2nd ed., J. Corcoran. Indianapolis, IN,152–278; (1936 [1983]). "On the Concept of Logical Consequence." Trans. of Tarski (1936) by J. H. Woodger, in Tarski (1983), 409–20.

extensional, we derive C1 from P1 and P2. Given that C2 is a logical consequence of C1, we derive C2. Having derived a contradiction, we must reject P1 or P2. We can't reject P2, which is established by inspecting (1) above. Rejecting P1 is also difficult. Since P1 is an instance of *Schema True*, its correctness seems to be guaranteed by the meaning of the truth predicate.

> *Schema True*: 'X' is a true sentence of English iff P (where 'X' is replaced by a sentence S and 'P' is replaced either by S or by a sentence synonymous with S).

How could any instance of this schema be false? A claim ⌈P iff Q⌉ can be false only if P is true and Q is false, or Q is false and P is true. But when P is ⌈'A' is true⌉ and Q is A, these combinations seem impossible. Surely the claim that A is true can't be true when A is false, nor can the claim that A is true be false when A is true. But if no instance of *Schema True* can be denied, then P1 can't be denied. This is the paradox.

The problem arises from the idea that Schema True incorporates a linguistic rule essential to understanding the truth predicate. Suppose one were asked to explain the meaning of "is true" to someone who knew some English but wasn't yet acquainted the word 'true'. One might explain it by saying something like this:

> "The sentence *snow is white* is true iff snow is white, the *sentence the sun shines nearly every day in Seattle* is true iff the sun shines nearly every day in Seattle, and so on."

How, if this explains what 'true' means, can one deny P1? One is tempted to think that since understanding 'true' requires one to accept all instances of *Schema True*, which the Liar shows to lead to contradiction, the ordinary notion of truth is incoherent; its presence in English is a defect that needs to be corrected.

Since Tarski was influenced by this reasoning, he made sure that the truth predicates he required for his metamathematical work were insulated from paradox. He did so by using a metalanguage M to specify a formal object language L, and then defining, in M, a restricted truth predicate T applying to sentences of L. Since L doesn't contain a truth predicate, no Liar-paradoxical sentences are constructible in L. Since any sentence S of M containing the predicate T is not a sentence of L, the truth predicate in S doesn't apply to S itself. Thus S can't be seen as asserting or denying its own truth. If, one wants a truth predicate for M, the process can be repeated in a higher metalanguage M+.

Tarski's Criteria of Correctness for Defining Truth

Tarski was concerned with languages of the predicate calculus that were assumed to be already understood by working logicians or mathematicians. This was crucial. *His definition of truth doesn't provide an interpretation of sentences of L*. On the contrary, the fact that the sentences were already used to make claims about a given domain provided Tarski with the concept he wanted his defined truth predicate to express. Tarski insists that L doesn't contain predicates expressing semantic concepts; nor does it have the means of referring to, or quantifying over, arbitrary expressions or sets of expressions of L. The metalanguage M includes either the sentences of L or translations of them. M also has the resources to refer to, and quantify over, expressions, sentences, and sets of such in L plus arbitrary sets of n-tuples of objects about which sentences of L are used to make claims. Tarski then shows how to construct an explicit definition in M of a predicate that applies to all and only the true sentences of L.

Before giving his definition, he lays down criteria for success. The most important criterion is that the definition be *materially adequate*. A definition in M of a truth predicate 'T_L' of sentences of L is materially adequate *iff for every sentence S of L, the definition entails at least one instance of Schema T – i.e., a sentence of M gotten by replacing 'X' with a name, NS, of S and replacing 'P' with S itself (if S is a sentence of M) or with a sentence, PS, of M that is a paraphrase of S.*

Schema T: X is T_L iff P

The role of material adequacy is to guarantee that the defined predicate '$\mathrm{T_L}$' is coextensive with 'true$_{\mathrm{L}}$ ', and so applies to all and only true sentences of L.

The guarantee can be illustrated using Schema TM, instances of which are gotten by replacing 'X' with a transparent name of a sentence of L and 'P' with any sentence of M.

Schema TM: If X means in L that P, then X is true (in L) iff P

TM connects our ordinary notions of truth and meaning. All its instances are obviously true and assertable. Let S be a sentence of L and ⌜NS is T_L iff PS⌝ be an instance of Schema T. Since PS means the same as S, the corresponding instance of TM ⌜If NS means in L that PS, then NS is *true$_L$* iff PS⌝ has a true antecedent. This gives us ⌜NS is *true$_L$* iff PS⌝, which, along with ⌜NS is T_L iff PS ⌝, allows us to derive ⌜NS is T_L iff NS is *true$_L$*⌝. Hence, we establish that if the definition of '$\mathrm{T_L}$' is materially adequate, then 'true' and '$\mathrm{T_L}$' are coextensive over L.

The Illusion that Truth and Tarski-Truth Are More Than Coextensive

All Tarski needed for his metamathematical work was for 'true$_{\mathrm{L}}$' and '$\mathrm{T_L}$' to be coextensive. But he and Carnap were tempted to believe that '$\mathrm{T_L}$' and 'true$_{\mathrm{L}}$' are also conceptually connected. Consider (2) and (3).

2a. 'John gave the book to Mary' is true$_{\mathrm{L}}$ iff John gave the book to Mary.
 b. x knows that 'John gave the book to Mary' is true$_{\mathrm{L}}$ iff John gave the book to Mary.
3a. 'John gave the book to Mary' is $\mathrm{T_L}$ iff John gave the book to Mary.
 b. x knows that 'John gave the book to Mary' is $\mathrm{T_L}$ iff John gave the book to Mary.

It is tempting to think that merely understanding (2a) is all that is required to know (2a) and to satisfy (2b). Now suppose that 'John gave the book to Mary' is a sentence of L and that '$\mathrm{T_L}$' is a Tarskian truth predicate defined in a metalanguage that contains L. What is required to know (3a), to satisfy (3b)? If (3a) is a consequence of the materially adequate definition of '$\mathrm{T_L}$', all it takes *is* for one to understand (3a) (which includes understanding the definition). So just as understanding (2a) seems to warrant accepting it, and lead to one's satisfying (2b), so understanding (3a) warrants accepting it, and leads to one's satisfying (3b).

From this it might seem to follow that merely understanding (4a) warrants accepting it, and leads to one's satisfying (4b).

4a. 'John gave the book to Mary' is $\mathrm{T_L}$ iff 'John gave the book to Mary' is true$_{\mathrm{L}}$.
 b. x knows 'John gave the book to Mary' is $\mathrm{T_L}$ iff John gave the book to Mary' is true$_{\mathrm{L}}$.

Since this result—which doesn't seem to depend on having empirical information beyond that needed to understand 'true$_{\mathrm{L}}$' and '$\mathrm{T_L}$'—can be repeated for every sentence of L, the coextensiveness of 'true$_{\mathrm{L}}$' and '$\mathrm{T_L}$' may appear to be *conceptually guaranteed.* This is an illusion. But it is an illusion with a distinguished pedigree.

Tarski's Commitment to the Illusion

Tarski informally explained his definition of truth in Tarski (1944), where he claimed that his defined notion '$\mathrm{T_L}$' is conceptually connected to our ordinary notion of truth, restricted to L. He says the definition "*does not aim to specify the meaning of a familiar word used to denote a novel notion; on the contrary it aims to catch hold of the actual meaning of an old notion*."[2] Since he took '$\mathrm{T_L}$' to capture what is essential to the ordinary predicate 'true$_{\mathrm{L}}$', he thought it could play all theoretical roles for which we might need a notion of truth. So he says

[2] Alfred Tarski (1944 [1952]). "The Semantic Conception of Truth and the Foundations of Semantics." Reprinted in Leonard Linsky (1952). Originally published in *Philosophy and Phenomenological Research* 4:341–76, cited at p. 13 of Linsky (1952).

that his notion of truth can be used to define semantic notions including *consequence*, *synonymy*, and *meaning*. He would not have said this had he not believed that ‘T_L’ comes very close to capturing the ordinary notion *being a true sentence of L.*

His stance in Tarski (1969) is similar.[3] First he explains what he calls *partial definitions of truth* (applying to individual sentences); then he explains how a general definition of truth (for the language) is related to the partial definitions. He begins by discussing meanings of sentences used to predicate truth or falsity of other sentences.

> “Consider a sentence in English whose meaning does not raise any doubts, say the sentence ‘snow is white’. For brevity we denote this sentence by ‘S’, so that ‘S’ becomes the name of the sentence. We ask ourselves the question: What do we mean by saying that S is true or that it is false? The answer to this question is simple: in the spirit of Aristotelian explanation, by saying that S is true we mean simply that snow is white, and by saying that S is false we mean that snow is not white.” (p.64)

Sometimes when one says that ‘snow is white’ is true one may assert that snow is white (which is an obvious consequence of the claim that the sentence is true in a context in which the meaning of ‘snow is white’ is understood by all). But Tarski seems to suggest that the sentences *‘snow is white’ is true* and *‘snow is white’* mean the same thing, which is much stronger. He would also have said that ⌈‘snow is white’ is T_E⌉ means the same as ‘snow is white’, when ⌈‘snow is white’ is T_E iff snow is white⌉ is a consequence of a materially adequate definition of a Tarskian truth predicate ‘T_E’ for a fragment E of English. Combining all this, we get the questionable conclusion that ‘snow is white’, *‘snow is white’ is true*, and ⌈‘snow is white’ is T_E⌉ are paraphrases.

Next, Tarski describes (5a) and (5b) as partial definitions of truth and falsity.

5a. ‘Snow is white’ is true iff snow is white.
5b. ‘Snow is white’ is false iff snow is not white.

He then explains that the task of defining true-in-L consists of formulating a *materially adequate* definition D the logical consequences of which include, for each sentence S of L, a *partial definition* of the predicate ‘T_L’. He notes that in the imagined case of a language with only finitely many sentences, such a definition is trivial. Let E be the fragment of English consisting of the following ten sentences: *1 is one of Bill’s favorite numbers*, *2 is one of Bill’s favorite numbers…10 is one of Bill’s favorite numbers.*

6. Tarskian Definition: For all sentences S of E, S is T_E (true in E) iff S = ‘1 is one of Bill’s favorite numbers’ and 1 is one of Bill’s favorite numbers, or …, S = ‘10 is one of Bill’s favorite numbers’ and 10 is one of Bill’s favorite numbers.

From the definition we derive (7).

7. ‘1 is one of Bill’s favorite numbers’ is T_E iff ‘1 is one of Bill’s favorite numbers’ = ‘1 is one of Bill’s favorite numbers’ and 1 is one of Bill’s favorite numbers, or ‘1 is one of Bill’s favorite numbers’ = ‘2 is one of Bill’s favorite numbers’ and 2 is one of Bill’s favorite numbers, or … ‘1 is one of Bill’s favorite numbers’ = ‘10 is one of Bill’s favorite numbers’ and 10 is one of Bill’s favorite numbers.

Assume we can derive each instance of the schema ‘_S’ = ‘S’_ that results from replacing both occurrences of the letter ‘S’ with a sentence of E , and also derive each instance of ‘_S’ ≠ ‘S*’_ that results from replacing the occurrence of ‘S’ with a sentence of E and replacing the occurrence of ‘S*’ with a different sentence of E. Then we derive the partial definition (T1) from (7).

[3] Alfred Tarski (1969). “Truth and Proof.” *Scientific American,* June: 63–67.

T1. ‘1 is one of Bill’s favorite numbers’ is T_E iff 1 is one of Bill’s favorite numbers.

Since partial definitions for the other sentences of E are also derivable, the definition is materially adequate. Thus (6) is a materially adequate general definition. If each “partial definition” gives the meaning of the application of ‘T_E’ to a sentence of E and that meaning is the same as the application of our ordinary predicate ‘is a true sentence of E’, then Tarski’s defined predicate matches the meaning of the ordinary truth predicate over E. That is the logic of the explanation presented in Tarski (1969).

The problem faced in Tarski (1935) was to reproduce this result for languages L with infinitely many sentences. The problem, as he conceived it, was a technical -- to derive a *partial* “definition of truth” for each of *infinitely* many sentences of L from a finite definition of a predicate ‘T’. Since he thought of each partial definition as giving the meaning of the application of ‘T’ to a sentence S of L, and since substituting our ordinary predicate ‘true’ for ‘T’ would yield a partial definition of ‘true’, giving the same meaning of an application ‘true’ to S, he thought that ‘true’ and ‘T’ must mean the same thing when applied to any sentence of L.

Tarski placed two further requirements on its solution. First, the definition must be formally correct, by which he meant that it must satisfy the usual rules for constructing definitions—including the rule that the definiendum not be defined in terms of (or conceptually dependent upon) any expressions used in the definition. The definition of truth in L does use logical vocabulary—quantifiers, identity, and truth-functional connectives. Still the definition is formally correct, because this logical vocabulary is primitive. Tarski’s final requirement is that the truth definition not employ, or depend on, any semantic terms—like *denotes*, or *applies to*. Since they give rise to paradoxes similar to those involving truth, his goal of insulating his formally defined truth predicate from paradox led him to demand a definition free of semantic primitives. Thus if the truth definition requires *denotation* and *application,* they too must be defined from nonsemantic primitives. Tarski (1935) does this.

Dispelling the Illusion

S1. Homophonic instances of *Schema True,* which, like *‘snow is white’ is true (in a fragment E of English) iff snow is white, have the form ‘S’ is true iff S*, can be known just by understanding and reflecting on them.

S2. If the metalanguage E+ of a Tarskian truth definition contains E, and the definition in E+ of Tarski’s predicate ‘T_E’ entails homophonic instances of *Schema T_E* , they too can be known simply by understanding and reflecting on them.

S3. Thus, for each sentence S of E one can establish ⌜‘S’ is T_E iff ‘S’ is true_E⌝ simply by understanding ‘true in E’ and ‘T_E’.

S4. Since no empirical information is needed to establish S3, ⌜‘S’ is true_E⌝ and ⌜‘S’ is T_E⌝ are conceptually equivalent (in effect, synonymous). Each is conceptually equivalent to S.

S5. Similar results can be obtained for cases in which the metalanguage of a Tarskian truth definition does not contain the object language.

S6. So, materially adequate, formally correct definitions of truth predicates capture the ordinary concept of truth when restricted to those languages.

The only step in this argument that is correct is step 2. (8) illustrates the problem with S1.

8. ‘Snow is white’ is a true sentence of English iff snow is white.

Suppose I speak English, but don’t know that the name ‘English’ refers to my language. I understand the name and know several things about it—e.g., that it designates a language spoken in England, North America, Australia, and New Zealand. But I don’t know that it designates a language I speak. This is possible, just as it is possible for me to understand the name 'Japanese' without knowing it is the language I hear on channel 25. If I am in this situation and don't know that 'English' designates the language I am using when considering (8), then I may not be in a position to know that (8) is true and assertable.

S3 would be incorrect even if we could establish S1 and S2. If we could do that we could show that ⌜'snow is white' is true$_E$ iff snow is white⌝ and ⌜'snow is T$_E$' iff snow is white⌝ can be known simply by understanding them. *But we couldn't show that* ⌜*'snow is white' is T$_E$ iff 'snow is white' is true$_E$*⌝ can be known simply by understanding it, because one can understand that sentence without understanding the sentence 'snow is white'.

S4 is also incorrect (on independent grounds). Suppose, that merely understanding ⌜'S' is true in E iff S⌝ were sufficient to know it. This wouldn't establish the conceptual equivalence of ⌜'S' is true$_E$⌝ and S. For that to hold, the propositions they express would have to be necessary and a priori consequences of each other, which they aren't. It is a contingent matter which linguistic conventions endow a sentence with meaning. When p is the proposition expressed by S, there will be possible conventions that would have rendered S false had they governed S, without affecting the truth of p. Thus, the proposition expressed by ⌜'S' is true$_E$ iff S⌝ isn't necessary, and ⌜'S' is true$_E$⌝ and S aren't necessary consequences of each other. Also, learning the meaning of a sentence requires acquiring empirical evidence about the linguistic conventions governing it. Because of this, there can be cases in which understanding S involves having empirical information that provides justifying evidence required to warrant accepting the proposition S expresses. Without ruling out the possibility that ⌜'S' is true$_E$ iff S⌝ is such a sentence, one couldn't establish that it expresses an apriori truth even if it could be known to be true merely by understanding it.

The problem with S5 arises from the fact that the instance of *Schema T* for a given object-language sentence S pairs it with a metalanguage sentence S$_M$, which, although it may express the same proposition as S, can be understood by an agent who also understands both S and S$_M$, without knowing that the sentences must have the same truth value. (See the literature on direct reference.) This blocks the reasoning in S1-S4 whenever the object-language sentence on the left-hand side of the biconditional is different from the sentence on the right-hand side.

Why Tarski's Theory of Truth isn't an Analysis of Truth

For decades many philosophers thought that Tarski's definitions of truth for different object-languages were philosophically revealing analysis of our ordinary notion of truth, restricted to those languages. According to Tarski our ordinary notion of truth is defective precisely because its unrestrictedness generates paradox. By contrast, it was often maintained, Tarski's restricted truth predicate eliminates this defect while preserving the important and useful features of our ordinary notion. On this view, Tarski specified, not how 'true' is actually understood, but how it ought to be understood, if it is to function in our theories in the ways we have hoped it would. To analyze a pretheoretic concept C in this way is to define a related concept C* that (i) applies to clear, central instances of C, (ii) is precise and well-defined, (iii) is free of difficulties that plague C, and (iv) is capable of performing the function of C in all theoretical contexts in which some such notion is required. This is what is meant when it is said that Tarski gave an analysis of truth.

Criteria (i) and (ii) are met for object-languages on which Tarski focused. It would seem that the same can be said for (iii). Many of Tarski's contemporaries —including Carnap, Neurath, Hempel, and Reichenbach—had been skeptics about truth. Tarski was historically effective in sweeping away that skepticism. He showed how to define truth predicates for certain languages L, using only notions already expressible in L plus descriptive syntax and elementary set theory. So, if syntax, set theory, and L are all unparadoxical and philosophically unproblematic, then adding Tarski's predicate 'T$_{\mathrm{Tarski}}$' to a metalanguage for L can't lead to philosophically objectionable consequences. For any sentence S of L, $\ulcorner$'S' is $T_{\text{Tarski}}\urcorner$ is provably equivalent, in the presence of descriptive syntax and set theory, to S. So, if prior to Tarski one had been inclined toward truth-skepticism, without seeing how one could do without it, then Tarski's definition might have seemed to provide a liberating analysis of what had been a questionable notion.

The final criterion for assessing whether Tarski's definition is an explication of truth is theoretical fruitfulness. Truth is important, and arguably indispensible, for many metatheoretical investigations. Often, we want to know whether all the claims of a given theory are true, whether there are truths it doesn't capture, and whether other theories do better in telling the truth about a specific domain than it does. It was precisely this kind of question that Godel raised when he asked whether any consistent theory of first-order arithmetic is capable of proving all and only the first-order arithmetical truths. Although he was able to brilliantly answer this question in the negative prior to Tarski's definition of truth for such languages, Tarski's formalization put the icing on the cake. In addition, we often want to know precisely when the truth of a set of sentences logically guarantees the truth of other related sentences. Tarski's work on truth made a great contribution to this because it laid the foundations for the now standard notions of *truth in a model,* on which the modern definitions of *logical truth* (i.e. the truth of a sentence in every model) and *logical consequence* (Q is a logical consequence of P iff every model in which P is true is a model in which Q is true).

Thus, it might seem as if Tarski's definition of truth meets all the requirements for being an analysis of truth. However, if we need a notion of truth that is conceptually connected to the notion of meaning in theories of linguistic meaning, then Tarski's notion fails the test. Moreover, the fact that it fails the test will lead us to qualify our judgment about it's relation to the now accepted notions of *truth in a model, logical truth,* and *logical consequence.*

Truth and Meaning

It has been widely assumed, even by Tarski, that there is a conceptual connection between truth and meaning. That connection is provided by the observations (i) understanding a sentence involves knowing the conditions in which it is true, and (ii) knowing the conditions in which a sentence is true provides information about its meaning. On influential version of this view was held by Donald Davidson in his 1967 article "Truth and Meaning."

(T) s is T iff p

> What we require of a theory of meaning for a language L is that *without appeal to any (further) semantic 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 [a transparent Tarskian name] of a sentence of L and 'p' by that sentence.
>
> Any two predicates satisfying this condition have the same extension, so if the metalanguage is rich enough, nothing stands in the way of putting what I am calling a theory of meaning into the form of an explicit definition of a predicate "is T." But whether explicitly defined or recursively characterized, it is clear that the sentences to which the predicate "is T" applies will be just the true sentences of L, for the condition we have placed on satisfactory theories of meaning is, in essence, Tarski's Convention T that tests the adequacy of a formal semantical definition of truth.
>
> The path to this point has been tortuous, but the conclusion may be stated simply: *a theory of meaning* of a language L shows "how the meanings of sentences depend upon the meanings of words" if it contains a (recursive) *definition of truth-in L…*I hope that what I am doing may be described in part as defending the philosophical importance of Tarski's semantic concept of truth…
>
> There is no need to suppress, of course, the obvious connection between a *definition of truth of the kind Tarski has shown how to construct,* and the concept of meaning. It is this: the

> definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give the truth conditions is a way of giving the meaning of a sentence. *To know the semantic concept of truth for a language is to know what it is for a sentence-–any sentence—to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language.*[4]

If the view here expressed by Davidson were correct, then the notion of truth defined by Tarski could play the central role in a theory of meaning for the object language over which the predicate is defined. If such a result could be established, it would support the claims in Carnap (1942) and Tarski (1944) that Tarski's notion of truth can be used to define and study semantic notions such as meaning and synonymy, thereby providing further vindication for taking his definition to be an adequate explication of truth.[5] But no such result can be established. On the contrary, the idea that anything remotely along these lines could be correct was a widespread mistake.

Imagine that 'e' is a name of the earth, that 'R' is a predicate applying to all and only round things, that 'T_L' is a Tarskian truth predicate, and that (9) is an instance of schema T that is derivable in the metalanguage from an explicit Tarskian definition of 'T_L'.

9. 'Re' is T_L iff the earth is round.

Since 'T_L' is the *definiendum* of the definition, it can be replaced, with no alteration of content, by the *definiens* (which, is free of all semantic notions). Performing the replacement yields (10).

10. [There is a set T* such that 'Re' is a member of T*, and for all sentences s of L, s is a member of T* iff (i) s = ⌈Pt⌉ for some one-place predicate P and term t, and there is an object o such that P *applies*$_T$ to o and o is *denoted*$_T$ by t; or clauses for 2, 3, … n-place predicates (and terms); or (ii) S = … clauses for truth-functional connectives … ; or (iii) s = … clauses for quantifiers …] iff the earth is round.

Since 'Re' is a sentence consisting of a one-place predicate followed by a term, we can simplify (10) by dropping the extraneous clauses in (i), (ii), and (iii). This gives us (11).

11. [There is a set T* such that 'Re' is a member of T*, and for all sentences s of L such that s = ⌈Pt⌉ for some one-place predicate P and term t, s is a member of T* iff there is an object o such that P *applies*$_T$ to o and o is *denoted*$_T$ by t] iff the earth is round.

Next, we replace 'denotes$_T$' and 'applies$_T$' with the definitions of those terms provided by an explicit list-like non-semantic Tarskian definition of each. This yields (12).

12. [There is a set T* such that 'Re' is a member of T*, and for all sentences s of L such that s = ⌈Pt⌉ for some one-place predicate P and term t, s is a member of T* iff there is an object o such that (i) t = 'e' and o = the earth, or t = 'm' and o is Mars, or … (one disjunct for each name in L) … , and (ii) P = 'R' and o is round, or P = 'M' and o is massive, or … (one disjunct for each predicate of L) …] iff the earth is round.

Recognizing trivial identities and nonidentities about expressions of L, we can simplify (12) by eliminating the nonidentities. This gives us (13), which is trivially equivalent to (14).

13. [There is a set T* such that (i) 'Re' is a member of T*, and (ii) 'Re' = 'Re' and 'e' = 'e' and 'R' = 'R' and there is an object o such that o = the earth and o is round] iff the earth is round.

[4] Davidson, Donald (1967 [2001]). "Truth and Meaning." In Davidson, *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 2001. Originally published in *Synthèse* 17:304–23. The quoted passage comes from pp. 23-24 of the 2001 reprinting.

[5] Rudolf Carnap (1942). *Introduction to Semantics*. Cambridge, MA: Harvard University Press.

14. There is an object o such that o = the earth and o is round iff the earth is round.

(10-14) provide no information about the meaning of 'Re'. One could know the facts they express without knowing anything about what 'Re' does, or doesn't, mean. Suppose one didn't know that 'Re' means that the earth is round, and one was considering the hypothesis that it means that the earth is not round. Given (10)–(14) plus instances of the a priori schema that *If s means in L that P and 'T_L' is a truth predicate for L, then s is T_L iff P*, one could conclude that either 'T_L' isn't a truth predicate for L (and T* isn't the set of true sentences), or 'Re' *doesn't* mean that the earth is not round. But without knowing the meanings of the sentences of L in advance, one couldn't determine whether 'T_L' was a truth predicate, and without knowing that, one could determine *nothing* about the meaning of 'Re' from a statement of its "Tarski-truth conditions."

The key point is that instances of schema (15a), which contain our ordinary truth predicate, are obvious a priori truths, whereas instances of (15b), which contain a Tarskian truth predicate, are neither obvious nor knowable a priori.

15a. If s means in L that P, then s is true in L iff P.
b. If s means in L that P, then s is T in L iff P.

It is the obviousness and availability of (15a) that allows claims of the form *s is true in L iff P* to provide information about meaning. If one knew that 'Re' is true in L iff the earth is round, then one could immediately eliminate the hypothesis that 'Re' means in L that the earth isn't round—since that hypothesis plus (15a) would contradict one's knowledge of the truth conditions of 'Re'. The unavailability of (15b) prevents similar conclusions from being drawn from claims of the form *s is T_L iff P.* Consequently, those claims carry no information about meaning.

This result shows what should have been obvious all along: *Tarski's truth predicates aren't semantic*. The fact that he required them to be definable entirely from non-semantic concepts expressed in the object language plus logic, set theory, and the syntax of L guaranteed that they couldn't be semantic. Since no concepts definable from his non-semantic base are semantic, the ubiquitous label applied to the notion he defined—*the semantic conception of truth*—is an absurd misnomer. It is a testament to the monumental historical misunderstanding of Tarski (1935) (by Tarski, Carnap, and others) that the only major philosopher of the era who recognized this was Alonzo Church.[6]

What is the source of the conceptual connection between truth and meaning that is missing in Tarski's substitute for truth? The natural thought is that it is the primacy of propositions as bearers of truth. The bearers of truth are, in the first instance, what agents assert and believe when they assertively utter, or otherwise accept, sentences. Sentences are true only derivatively, when the linguistic rules governing their use determine a single proposition (the meaning of the sentence), which is, in fact, true. Thus, when we are told that a sentence is true, we are given information about its meaning and the proposition routinely expressed when it is used.

When a sentence contains no indexical or other semantically context-sensitive element, there is often a single proposition determined by its linguistic meaning that is reliably, though not invariably, a constituent of the illocutionary content of uses of the sentence. In these cases there is a close relationship between talk about meaning of the sentence and talk about the proposition it expresses. In such cases, instances of schema (15a) are tantamount to instances of schema (16).

16. If S means in L that P (i.e., is used by speakers of L to express the proposition that P), then

[6] Alonzo Church, (1944). *Introduction to Mathematical Logic*. Princeton, NJ: Princeton University, pp. 65-66.

the proposition expressed by S in L is true iff P.

The conceptual connection between truth and meaning is the result of the fact that to say of S that it means in L that P is to say that uses of S in accord with the conventions of L express the proposition that P. This explains why to say that S is true in L is to say that the proposition expressed by S in L is true. To this we add the instances of the schema *that the proposition that P is true iff P* are obvious, a priori, and necessary.

In sum, the information about meaning carried by statements specifying the truth conditions of sentences is due to the implicit commitment to propositions carried by ascriptions of our ordinary notion of truth to sentences. Since propositions play no role in the definition of Tarski's truth-substitute, predication of his concept to a sentence carries no information about the sentence's meaning. His predicate and the ordinary truth predicate of sentences do, of course, coincide in extension over the object language. But they don't express the same property, and so uses of sentences containing them don't encode remotely the same information.

The fact that Tarski's defined truth predicates are useless in semantics shows that his non-semantic notion of truth is not an adequate explication of our ordinary notion. But this doesn't mean that the recursive apparatus used in his characterization of truth, and truth in a model, is useless. Far from it. That apparatus is simply not part of a definition of truth. Rather, it is an essential part of theories or definitions that employ the ordinary notion of truth for special purposes. In logic and model theory the Tarskian formal apparatus is incorporated in defining what it means for a model to be taken as a genuine interpretation of the sentences of a formal language. In empirical theories of meaning that apparatus is part of the systematic assignment of the conditions in which a sentence is true in the ordinary sense. These are magnificent contributions. They simply aren't contributions of the sort that they have often been taken to be.

Tarski's Contribution to Model Theory

In order to understand the point just made about model theory a word must be said about (i) how the apparatus provided by Tarski's truth definition is used in defining the concept *truth in a model*, and (ii) how that notion is used to define *logical truth* and *logical consequence*.

To understand the relationship between Tarski's truth definitions and the concept *truth in a model,* it is helpful to state his truth definitions in a form that differs slightly from his, without affecting the final result. For this purpose we let L be a language the nonlogical vocabulary of which consists of the names, *a, b, c, d, e,* plus the predicates *F, G, H,…Q.* Atomic formulas of L are n-place predicates followed by n terms, where a term is either a name or a variable *w, x, y, z.* A sentence S is a non-atomic formula if and only if, for some variable *v* and formulas Φ and Ψ, S = (~Φ), or S = $(\Phi \; \& \; \Psi)$, or S = $(\Phi \vee \Psi)$, or S = $(\exists v \; \Phi v)$. Nothing else is a formula. Sentences are formulas with no free occurrences of variables. An occurrence of a variable *v* in a formula is free if it is not in the scope of any occurrence of $\exists v$. The scope of an occurrence of $\exists v$ in a formula is the smallest complete formula that immediately follows it.

We may divide a Tarski-style definition of truth for this language into two parts – one dealing with the non-logical vocabulary of L and the other, using the results of the first, defining truth for sentences. We begin with Tarski-like specifications of the reference of singular terms – names and variables – and the application of predicates. Assignments of objects as the values of variables are used to assign temporary referents to variables – treating them, in effect, as temporary names. Following Tarski, the specifications (which he, in effect, takes to be definitions) are simply lists.

The Reference of Names

A name n of L refers to an object o iff n = 'a' and o is Alberto, or n = 'b' and o is Bob, or n = c and o is Carmen, or n = d and o is Dolores (and so one, a clause for every name.

The Reference of Variables Relative to Assignments of Values to Variables
A variable *v* refers to an object o relative to an assignment A iff A assigns o to *v*.

The Application of Predicates
An 1-place predicate P of L applies to an object o iff P is 'F' and o is female, or P is 'M' and o is male,…and so on for each 1-place predicate.
There are similar clauses for 2, 3, … place predicates until all the predicates are given.

Next we define what it is for a formula to be true relative to an assignment of values to variables.

The Truth of a Formula Relative to an Assignment
An atomic formula consisting of an n-place predicate followed by n terms (names or variables) is true relative to an assignment A iff the predicate *applies* to the n-tuple of the referents of the terms, relative to A.

A formula $(\sim\Phi)$ is true relative to an assignment A iff Φ is not true relative to A.

A formula $(\Phi \,\&\, \Psi)$ is true relative to an assignment A iff Φ is true relative to A and Ψ is true relative to A.

A formula $(\Phi \text{ v } \Psi)$ is true relative to an assignment A iff Φ is true relative to A or Ψ is true relative to A.

A formula $(\exists v\ \Phi v)$ is true relative to an assignment A iff there is some assignment A* of values to variables that assigns *v* an object that makes Φv true relative to A* -- where A* must be either A itself, or an assignment that differs from A* on what it assigns to *v,* while agreeing with A on what is assigned to all other variables.

Last we define what it is for a sentence (which, of course, contains no free occurrences of variables) to be true.

The Truth of a Sentence
A sentence S of L is true iff it is true relative to some assignment (which, it turns out, is the same as saying that S is true relative to every assignment).

The result is a materially adequate, formally correct Tarski-style definition in which truth is "defined" in non-semantic terms. If the predicate being introduced were simply a previously uninterpreted symbol 'T', the fact that the definition is materially adequate would show that 'T' applies to all and only the truths of L.

To define truth in a model, we first introduce the idea of *a model for a language* corresponding to certain parts of the above "truth definition." A model is a selection of a set of objects, called the *domain*, that the language is used to talk about, plus an assignment of objects in the domain to the names, sets of objects in the domain to the 1-place predicates, and sets of n-tuples of objects in the domain to the n-place predicates. These assignments can be put in the form of list-like definitions of *the reference of names, the reference of variables relative to assignments of objects (in the domain) as values of the variables,* and *the application of predicates.* In short, we may take a model to be a collection of these definitions (of the sort illustrated above) in which it is stipulated that the objects mentioned in them are objects of the domain. *Truth in a model M* is then defined as follows:

Truth in a Model
An atomic formula $Pt_1 \ldots t_n$ is true in M relative to an assignment A iff *P applies* in M to the n-tuple of *denotations* $o_1 \ldots o_n$ of $t_1 \ldots t_n$ in M relative to A

$\sim\Phi$ is true in M relative to an assignment A iff Φ is not true in M relative to A.

Φ & Ψ is true in M relative to A iff Φ and Ψ are both true in M relative to A.

$\Phi \vee \Psi$ is true in M relative to A iff either Φ or Ψ (or both) are true in M relative to A.

$\exists v\ \Phi(v)$ is true in M relative to A iff there is an object o in D_M and assignment A* that assigns o to v that is identical with A or that differs from A only in what it assigns to v, and $\Phi(v)$ is true in M, relative to A*.

A sentence is true in a model M iff it is true in M relative to all assignments.

Finally we define logical truth and logical consequence.

S is *logically true* iff S is true in every model of M (that assigns S any truth value).
Q is a logical consequence of P iff Q is true in every model in which P is true.

Although the definition of *truth in a model*, used in defining these logical notions, employs the technical apparatus found in Tarski's original definition of truth, the notion of *truth* occurring in *truth in a model* is standardly taken to be our ordinary, pretheoretic notion, rather than any Tarskian substitute. When *truth in a model* is so understood, to say that S is logically true is to say that no matter how the nonlogical vocabulary of S is *interpreted*, and no matter what objects are talked about, S will come out true.

Carnap's Flawed Tarskian Epiphany

Prior to Tarski, Carnap, Neurath, Hempel, Reichenbach, and other logical empiricists either identified *truth* with *being highly confirmed,* or rejected the former in favor of the latter. Tarski immediately changed this for Carnap, and eventually for most of the others as well. Upon learning Tarski's views, Carnap become convinced of their philosophical importance and the need to communicate them to a philosophical audience. So he suggested to Tarski that he lecture on truth at the International Congress for Scientific Philosophy held in Paris in September of 1935. He reports Tarski as being skeptical that many philosophers would be interested, a skepticism that Carnap countered by promising to deliver his own lecture on the importance of Tarski's "semantic" conception. So Tarski agreed to speak.

Carnap's Congress paper distinguished truth from confirmation.

> The difference between the two concepts 'true' and 'confirmed' ('verified', 'scientifically accepted') is important and yet frequently not sufficiently recognized. 'True' in its customary meaning is a time-independent term. … For example, one cannot say "such and such a statement is true today (was true yesterday; will be true tomorrow)" but only "the statement is true." 'Confirmed', however, is time-dependent. When we say "such and such statement is confirmed to a high degree by observations" then we must add: "at such and such a time."[7]

A statement (proposition) that is highly confirmed at one time may not be highly confirmed at another time, even though it is true throughout. Hence, Carnap argued, the *truth* of a proposition is different from its being *highly confirmed.* Unfortunately, the passage doesn't identify statements with propositions. Problems arise when Carnap removes the unclarity by identifying statements with sentences, which, by the inclusion of tense, may express different propositions with different truth values at different times. In such cases a *sentence* said to be true at one time would be said to be false at another time, even though for Carnap the truth is supposed to be timeless.

The problem stems from his persistent conflation of sentences, uses of sentences in accord with the linguistic conventions of a language, and propositions. Being an opponent of propositions as nonlinguistic entities that are meanings of sentences, it was natural for him to

[7] Rudolf Carnap (1949). "Truth and Confirmation." In H. Feigl and W. Sellars, eds., *Readings in Philosophical Analysis,* New York: Appleton-Century-Crofts, 119–27, cited at p. 119.

use ‘proposition’ and ‘statement’ for *uses of sentences*. The idea that certain uses of sentences are propositions is justifiable, but he didn’t systematically explore it. Since he also took a sentence to be true iff uses of it are true, he conflated sentences as syntactic structures with uses those of structures in accord with linguistic conventions. This spelled trouble. Whereas Tarski’s truth predicate applies directly to syntactic structures, abstracted from the semantic conventions governing them, our ordinary notion of truth applies directly to uses of sentences in accord with their governing conventions, and only indirectly to sentences individuated syntactically.

Carnap’s failure to notice this infects his argument for distinguishing truth from confirmation, which was based on the following sentences.

17a. The substance in this vessel is alcohol.
17b. The sentence ‘the substance in this vessel is alcohol’ is true.
18a. X knows (at the present moment) that the substance in this vessel is alcohol.
18b. X knows that the sentence ‘the substance in this vessel is alcohol’ is true.

The argument assumes that Tarski showed us that since (17a) and (17b) are logically equivalent, (18a) and (18b) are also be equivalent. If one adds (19) to this, one gets the absurd result that sentence (17a) is logically equivalent to (18a).

19. To say that S is true is to say that S has been confirmed to a degree high enough to warrant accepting the provisional claim *that X knows that S.*

That was Carnap’s argument against (19).

What is interesting about his argument is not its obviously correct conclusion that (19) should be rejected, but the theses about truth and Tarski-truth that he makes in giving the argument.

T1. Our ordinary predicate *true* of sentences of a language L means essentially the same thing as the predicate *true*$_{Tarski}$ that Tarski defines.

T2. ⌜‘P’ is true⌝ and ⌜‘P’ is true$_{\text{Tarski}}$⌝ are logically equivalent to sentence P, and so to each other. They are “different formulations of the same factual content”; “nobody may accept the one and reject the other”; and “they convey the same information.”

T3. ⌜John knows that ‘P’ is true⌝ and ⌜John knows that ‘P’ is true$_{\text{Tarski}}$⌝ are logically equivalent to ⌜John knows that P⌝, and hence to each other.

To evaluate these theses, one must realize that both truth predicates can meaningfully be predicated of sentences one doesn’t understand. I can say, of a Japanese sentence, or a sentence of English that contains a word I don’t understand, that I know, from the testimony of others, that it is true. I can do this using ⌜‘P’ is true⌝, which I understand perfectly well. I can also understand and accept ⌜‘P’ is true$_{\text{Tarski}}$⌝ without understanding P. The falsity of T1–T3 is now obvious. Suppose that ⌜‘P’ is true$_{\text{Tarski}}$ iff Mary gave Bill the book⌝ is a logical consequence of the Tarskian definition of *true*$_{Tarski}$. Then ⌜‘P’ is true$_{\text{Tarski}}$⌝ and ‘Mary gave Bill the book’ will be logical consequences of each other, and anyone who understands both will be in a position to logically derive one from the other. By contrast, ‘Mary gave Bill the book’ is not a logical, necessary, or a priori consequence of ⌜‘Mary gave Bill the book’ is true⌝, nor is ⌜‘Mary give Bill the book’ is true⌝ such a consequence of ‘Mary gave Bill the book’. This falsifies all three theses.

Carnap’s errors concerning T1-T3 were closely connected with what became his long-standing failure to see that statements of the Tarski-truth conditions play no role in endowing sentences with meaning, interpreting them, or describing their meanings once they have acquired them. Thus, after announcing in the preface of *Introduction to Semantics* (1942), that he was using Tarski’s notion of truth, he characterized the rules of a semantical system S (which are really stipulated conventions governing the use of its expressions) as constituting of "nothing else than a definition of certain semantical concepts with respect to S, e.g., 'designation in S' or 'true in S'."[8] In section 7 he says:

> A *semantical system* is a system of rules which state *truth-conditions* for the sentences of an object language and thereby determine the meaning of those sentences. A semantical system S may consist of *rules of formation,* defining 'sentence in S', *rules of designation*, defining 'designation in S', and *rules of truth,* defining 'true in S'. The sentence in the metalanguage ⌈P is true in S⌉ means the same as the sentence P itself. This characteristic constitutes a condition for the *adequacy* of the definition. (p. 22)

Note Carnap's insistence that the rules of the semantical system constitute *definitions* of 'designation in S' and 'true in S', exactly as Tarskian definitions define *denotation*$_{Tarski}$ and *true*$_{Tarski}$. Carnap adds that a metalanguage sentence that predicates truth of a sentence *means the same as* the sentence itself. This is false if by 'true in S' means *true in the ordinary sense.* It is true if (i) he means *true*$_{Tarski}$, (ii) what he calls "logical equivalence" is sufficient for sameness of meaning, and (iii) the metalanguage definition of *true*$_{Tarski}$ pairs P and ⌈'P' is true$_{\text{Tarski}}$⌉. Carnap's final remark about the *adequacy* of the *definition* being provided by the equivalence of P and ⌈'P' is true⌉ leaves no room for doubt; by 'true' he means 'true$_{\text{Tarski}}$'.

Carnap continues, describing semantic rules that

> determine a *truth-condition* for every sentence of the object language, i.e. a sufficient and necessary condition for its truth. In this way the sentences are *interpreted* by the rules, i.e. made understandable, because *to understand a sentence, to know what is asserted by it, is the same as to know under what conditions it would be true.* To formulate it in still another way: the rules determine the *meaning* or *sense* of the sentences. (p. 22)

Here Carnap connects claims about truth conditions to claims about meaning and understanding. By contrast with the preceding paragraph just cited, this paragraph makes sense only if the truth conditions are stated using the ordinary truth predicate of sentences. If Tarski's defined truth predicate is intended, the remarks are absurd.

Four pages later he reveals an important source of his error.

> A remark may be added as to the way in which the term '*true'* is used in these discussions. … We use the term here in such a sense *that to assert that a sentence is true means the same as to assert the sentence itself*; e.g. the two statements "The sentence 'the moon is round' is true" and "The moon is round" are merely two different formulations of the same assertion. (The two statements mean the same in a logical or semantical sense.) (p. 26)

Let S be the English sentence 'Five is a prime number'. Imagine it being used in an ordinary context in which speaker and hearer (i) understand the sentence, (ii) know that it expresses the proposition that five is a prime number and hence is true iff five is a prime number, (iii) presuppose this about each other, and (iv) realize that they both presuppose this. In this context, an agent who assertively utters *The sentence 'Five is a prime number' is true* can correctly be reported either as having asserted *that 'Five is a prime number' is true,* or as having asserted *that five is a prime number,* or as having asserted both. In many contexts, one who assertively utters 'Five is a prime number' also could be correctly taken to assert *that 'Five is a prime number' is true*. In these contexts it is transparent that to commit oneself to the truth of the sentence is to commit oneself to five's being a prime number, and to commit oneself to five's being a prime number is to commit oneself to the truth of the sentence. Carnap was sensitive to this fact about assertive commitments, but he misdiagnosed its source. The sentences don't mean the same thing.

[8] Rudolf Carnap (1942). *Introduction to Semantics*. Cambridge, MA: Harvard University Press, p. xii.