10

Truth, the Liar, and Tarski's Semantics

GILA SHER

1 Tarski's Theory of Truth

The most influential (and arguably, the most important) development in the modern study of truth was Tarski's 1933 essay "The Concept of Truth in Formalized Languages." The theory formulated in this essay distinguished itself from earlier theories in a number of ways: (1) it was a formal, that is mathematical (or quasi-mathematical) theory; (2) it offered a detailed, precise, and rigorous definition of truth; (3) it confronted, and removed, a serious threat to the viability of theories of truth, namely, the Liar Paradox (and other semantic paradoxes); (4) it made substantial contributions to modern logic and scientific methodology; (5) it distanced itself from traditional philosophical controversies; and (6) it raised a spectrum of new philosophical issues and suggested new approaches to philosophical problems.

Historically, we may distinguish two goals of Tarski's theory: a *philosophical* goal and a (so-called) *metamathematical* goal. Tarski's philosophical goal was to provide a definition of the ordinary notion of truth, that is the notion of truth commonly used in science, mathematics, and everyday discourse. Tarski identified this notion with the classical, *correspondence* notion of truth, according to which *the truth of a sentence consists in its correspondence with reality*. Taking Aristotle's formulation as his starting point – "To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true" (Aristotle: $1011^{b}25$) – Tarski sought to construct a definition of truth that would capture, and give precise content to, Aristotle's conception.

Tarski's second goal had to do with logical methodology or, as it was called at the time, metamathematics. Metamathematics is the discipline which investigates the formal properties of theories (especially mathematical theories) formulated within the framework of modern logic (first- and higher-order mathematical logic) as well as properties of the logical framework itself. Today we commonly call this discipline 'metalogic.' The notion of truth plays a crucial, if implicit, role in metalogic (e.g. in Gödel's completeness and incompleteness theorems), yet this notion was known to have generated paradox. Tarski's second goal was to demonstrate that 'truth' could be used in metalogic in a consistent manner (see Vaught 1974).

2 Tarski's Solution to the Liar Paradox

One of the main challenges facing the theorist of truth is the Liar Paradox. There are many versions of the paradox. (In antiquity, it was formulated in terms of 'lie,' whence its name, 'the liar paradox.') Tarski formulates the paradox as follows:

Let *c* abbreviate the expression 'the sentence printed on line 10 of the present page'. Consider the sentence:

c is not true.

It is clear that:

(1) c = '*c* is not true',
(2) '*c* is not true' is true iff (if and only if) *c* is not true.

Using the laws of classical logic, we derive a contradiction from (1) and (2):

(3) *c* is true iff *c* is not true.

What is the source of the paradox? Tarski's premises appear innocuous: (1) is an easily verified empirical statement, and (2) is an instance of an uncontroversial schema, namely, the *Equivalence Schema,*

(E) x is true iff *p*,

where '*p*' represents a sentence and 'x' a name of this sentence. (A simple instance of this schema is *'Snow is white' is true iff snow is white.*) Assuming the laws of classical logic are not the source of the paradox, it is natural to look for its source in *c*. One special feature of *c* is its predicating a property involving truth of itself. Tarski identifies this feature as responsible for the paradox. A language which contains its own truth predicate as well as names of all its sentences Tarski calls *semantically closed.* (More generally, any language which has the resources for describing its own syntax and contains its own semantic predicates (see below) is semantically closed.) Provided that such a language has a reasonable logical apparatus, it generates paradoxical sentences. Tarski concludes that semantically closed languages are inconsistent, that is they generate sentences that cannot be consistently given either the value True or the value False. In particular, the notion of truth (and other semantic notions) cannot be consistently defined for such languages. This conclusion is far from trivial: Natural languages are *universal* in the sense that anything that can be said by a speaker in any language can be said by him/her in his/her natural language. As such, natural languages are (generally) semantically closed, and truth (and other semantic notions) cannot be defined for such languages.

Not all languages, however, are semantically closed. Most mathematical and scientific languages are not. Such languages Tarski calls *semantically open.* Tarski's solution to the Liar Paradox is to restrict the definition of truth to open languages. This solution

requires that we think of languages as placed in a *hierarchy*: To define truth for a given open language L (our 'target language' or, in Tarski's terminology, 'object language'), we ascend to a higher (open) language, ML or meta-L, which has the resources for referring to all expressions (in particular, sentences) of L, and we formulate our definition of truth for L in ML. Truth for ML is defined in a third open language, MML, still higher in the hierarchy, and so on. This solution to the Liar Paradox is commonly called the *hierarchical solution*.

Tarski directs his attention to a particular family of open languages, namely, languages formalized within the framework of modern mathematical logic. Each such language includes (1) a set of logical constants containing a complete collection of truth-functional connectives (classically interpreted), the existential and/or universal quantifier, and possibly identity; (2) an infinite set of variables; and (3) a set (possibly empty) of nonlogical constants: individual constants, functional constants, and predicates. (Note: If L is a Tarskian language of order n, then for each $1 \leq i \leq n$, L has an infinite set of variables of order i, and the number of its symbols and well-formed expressions of order i is countable, that is it does not exceed the number of positive integers.) Since only interpreted sentences can be said to be true or false, Tarski restricts his attention to *interpreted* languages, that is languages whose primitive constants (logical and nonlogical) are fully interpreted. Such languages are naturally viewed as formalizations of scientific and mathematical languages as well as of open segments of natural languages. Tarski refers to such languages as "formalized languages" (or "formalized languages of the deductive sciences"). His goal is to construct a general method for defining truth for formalized languages.

3 Tarski's Method of Defining Truth for Formalized Languages

General principles

Given a formalized language L, the definition of truth for L is formulated in a metalanguage of L, ML. To define truth for L in ML we introduce an uninterpreted 1-place predicate, 'T,' into ML, and define it as a truth predicate for L, that is as a predicate satisfied by all and only true sentences of L. The definition of T is required to satisfy two conditions: (1) it has to be formally correct, that is avoid paradox, and (2) it has to be materially adequate, that is capture the idea that truth is correspondence with reality.

Formal correctness

To define T in a formally correct manner we follow the usual procedures for formally correct definitions, and in particular we make sure that the circumstance responsible for the Liar Paradox, namely, the truth for L being defined in L itself, does not arise. To this end we construct ML as an essentially stronger language than L, that is ML has expressions which are not translatable to L. In particular, the definition of T in ML is not translatable to L.

Material adequacy

To ensure that the definition of T is materially adequate, we require that it satisfy the following criterion ("convention," in Tarski's terminology):

Criterion (T)

A definition of T (in ML) is a materially adequate definition of truth for L iff it implies, for every sentence σ of L, an ML-sentence of the form

T(s) iff p,

where 's' stands for an ML name of σ and 'p' for an ML sentence with the same content as σ (a translation of σ to ML).

The idea is that given a sentence σ of L, an adequate definition of truth for L implies that σ has the property T just in case things in the world are as σ says. For example, if σ is the sentence 'Snow is white,' the definition of T implies that σ has the property T iff the stuff snow has (in reality) the property of being white. To satisfy this requirement, ML is required to contain, for each sentence σ of L, a sentence with the same content as σ. Using the notational conventions that 'snow is white' is an ML-name of the L-sentence 'Snow is white,' and 'snow is white' is an ML sentence with the same content as 'Snow is white,' the definition of T implies the ML-sentence:

T('snow is white') iff snow is white.

In constructing a definition of truth for L in ML we have to take into account the fact that the number of sentences in any language formalized within the framework of modern logic is infinite. A definition like

T(s) iff (s = 'snow is white' and snow is white, or s = 'grass is red' and grass is red, or . . .),

will not do, since such a definition would be infinitely long. To avoid this difficulty Tarski uses the *recursive* method. The recursive method enables us to define predicates ranging over infinitely many objects in a finite manner, provided certain conditions are satisfied. Such definitions are finitely long and they determine whether a given object falls under a given predicate in finitely many steps. I will not specify the conditions for recursive definitions here (for a good account see Enderton 1972, section 1.4), but the idea is that if every sentence of L is uniquely generated from finitely many atomic sentences (of L) by finitely many logical operations, and if the atomic sentences and logical operators of L are finitely specifiable, then truth for L can be recursively defined. Such a definition determines the truth value of each sentence of L based on (1) the truth values of its atomic constituents, and (2) its logical structure. For example, if the only logical constants (operators) of L are Negation and Disjunction, then truth for L is definable by specifying (1) the truth values of the atomic sentences of L, (2) a rule for determining the truth value of a Negation given the truth value of the negated sentence, and (3) a rule for determining the truth value of a Disjunction given the truth values of its disjuncts.

If L contains quantifiers, however, truth for L cannot be defined in this way. Sentences involving quantifiers are generated not from atomic sentences but from atomic formulas, including formulas with free variables (variables which are not in the

scope of any quantifier), and such formulas do not have a truth value. (For example, '(∀x)Px' is generated from the atomic formula 'Px' which, having a free variable, has no truth value.) But truth for L can be recursively defined via an auxiliary notion, *satisfaction*, applicable to formulas. The notion of satisfaction is an intuitive notion: The atomic formula 'x is even' is satisfied (in the domain of the natural numbers) by 0, 2, 4, More generally, '$Rx_1, \ldots x_n$' is satisfied by an n-tuple of objects, $\langle a_1, \ldots, a_n \rangle$, iff $a_1, \ldots, a_n$ (in that order) stand in the relation $\underline{R}$ (the relation referred to by 'R'). The definition of truth for L proceeds in two steps: (1) a recursive definition of satisfaction for L, and (2) a (nonrecursive) definition of truth for L based on (1).

Tarski's example

Tarski explained his method through an example. Using contemporary terminology, his example can be concisely described as follows.

Object language: L_C

The target language is the language of the calculus of classes (an interpretation of the language of Boolean algebra). I will refer to it as 'L_C.' L_C is an interpreted first-order language whose primitive vocabulary consists of the logical constants '~' (negation), '∨' (disjunction) and '∀' (the universal quantifier), the nonlogical constant '⊆' (a 2-place predicate interpreted as class inclusion), and variables, 'x_1', 'x_2', 'x_3', . . . , ranging over all objects in the domain, D_C, of L_C. D_C is a set of classes.

Meta-language: ML_C

Truth for L_C is defined in a meta-language, ML_C. ML_C relates to L_C in the way described above. In particular: (1) the syntax of L_C is describable in ML_C; (2) each constant of L_C has both a name and a translation (a constant with the same meaning) in ML_C; (3) ML_C has an undefined 1-place predicate, 'T,' designated as the truth predicate of L_C, as well as other predicates definable as semantic predicates of L_C; and (4) ML_C has variables of a higher-order than those of L_C (or a set-theoretical apparatus richer than that of L_C).

Definitions (in ML_C*)*

Notation: Let 'v_i' and 'v_j' be schematic symbols representing arbitrary variables, x_i and x_j, of L_C, and let 'Φ,' 'Ψ' and 'σ' be schematic symbols representing arbitrary expressions of L_C. Let '$\ulcorner$' and '$\urcorner$' be square quotes, where '$\ulcorner \Phi \vee \Psi \urcorner$' stands for 'the result of concatenating the formula Φ, the symbol '∨' and the formula Ψ, in that order' (see Quine 1951). For each primitive constant c of L_C, let $\underline{c}$ be a name of c in ML_C and $\underline{\underline{c}}$ a translation of c to ML_C.

Formula (of L_C*) – Inductive Definition*

1. $\ulcorner v_i \subseteq v_j \urcorner$ is a formula.
2. If Φ is a formula, $\ulcorner \sim \Phi \urcorner$ is a formula.
3. If Φ and Ψ are formulas, $\ulcorner \Phi \vee \Psi \urcorner$ is a formula.

4. If Φ is a formula, $\ulcorner \underline{\forall} v_i \, \Phi \urcorner$ is a formula.
5. Only expressions obtained by 1–4 are formulas.

Sentence (of L_C)

σ is sentence iff σ is a formula with no free occurrences of variables.

Let g be any function which assigns to each variable of L_C an object in the domain, D_C, of L_C. We will call g 'an assignment function for L' and refer to $g(v_i)$ as 'g_i'.

Satisfaction (of a Formula of L_C by g) – Recursive Definition

1. g satisfies $\ulcorner v_i \subseteq v_j \urcorner$ iff $g_i \subseteq g_j$
2. g satisfies $\ulcorner \simeq \Phi \urcorner$ iff $\simeq$ (g satisfies Φ).
3. g satisfies $\ulcorner \Phi \veebar \Psi \urcorner$ iff [(g satisfies Φ) $\veebar$ (g satisfies Ψ)].
4. g satisfies $\ulcorner \underline{\forall} v_i \, \Phi \urcorner$ iff $\underline{\forall} g'$ (if g′ differs from g at most in g_i, then g′ satisfies Φ).

T (Truth of a Sentence of L_C)

T(σ) iff: (1) σ is a sentence, and (2) $\underline{\forall} g$(g satisfies σ).

4 Tarskian Semantics

Semantics and correspondence

Truth, for Tarski, is (as we have seen above) a correspondence notion. But truth is not the only correspondence notion. The discipline which studies correspondence notions in general Tarski calls 'semantics':

> We shall understand by semantics the totality of considerations concerning those concepts which, roughly speaking, express certain connexions between the expressions of a language and the objects and states of affairs referred to by these expressions. (Tarski 1936a: 401)

Some semantic notions express correspondence directly: *reference*, *satisfaction*, and *definition* are such notions: the name 'Mount Everest' *refers* to a mountain in the Himalayas; the formula 'x was assassinated' is *satisfied* by John Kennedy; the expression 'x^2' (where 'x' ranges over the natural numbers) *defines* the set $\{0,1,4,9,16, \ldots\}$. Other semantic notions, for example 'truth', express correspondence indirectly. Truth is a property of sentences rather than a relation between sentences and objects, but truth holds of a given sentence only if the *objects* referred to by this sentence possess the *properties* (*relations*) attributed to them by it. (To apply this principle to sentences containing logical constants we either construe the logical constants as referential constants – that is Identity as referring to the identity relation, Negation as referring to complementation, the Existential quantifier as referring to the higher-order property of nonemptiness, etc. – or we construe statements containing logical constants as *reducible* to statements (or formulas) satisfying the correspondence principle.)

Correspondence and disquotation

Some philosophers regard semantic notions as *disquotational* notions: a sentence enclosed in quotation marks has the property of being true iff this sentence, its quotation marks removed, holds (Ramsey 1927). Tarski, however, views the two analyses as equivalent:

> A characteristic feature of the semantical concepts is that they give expression to certain relations between the expressions of language and the objects about which these relations speak, or that by means of such relations they characterize certain classes of expressions or other objects. We could also say (making use of the *suppositio materialis*) that these concepts serve to set up the correlation between the names of expressions and the expressions themselves. (Tarski 1933: 252)

We can explain Tarski's view as follows: There are two modes of speech, an *objectual mode* and a *linguistic mode* ('material' mode, in Medieval terminology). The correspondence idea can be expressed in both modes. It is expressed by

'Snow is white' is true iff snow is white,

as well as by

'"Snow is white" is true' is equivalent to 'Snow is white.'

In the objectual mode we say that a sentence attributing the (physical) property of whiteness to the (physical) stuff snow is true iff the (physical) stuff snow has the (physical) property of whiteness; in the linguistic mode we say that a sentence attributing (the semantic property of) truth to a sentence attributing whiteness to snow is equivalent to a sentence attributing whiteness to snow.

Logical semantics

One of the most important achievements of Tarskian semantics is its contribution to the definition of meta-logical notions ('logical consequence,' 'logical truth,' 'logical consistency,' etc.). Shortly after completing his work on truth, Tarski turned his attention to the notion of *logical consequence*. Prior to Tarski, 'logical consequence' was defined in terms of proof (the sentence σ is a logical consequence of the set of sentences Γ iff there is a logical proof of σ from some sentences of Γ). Gödel's incompleteness theorem showed, however, that the proof-theoretic definition of 'logical consequence' is inadequate: Not all theories formulated within the framework of modern logic can be axiomatized in such a way that all their true sentences are provable from their axioms. Using the resources of semantics on the one hand and set theory on the other, Tarski developed a general method for defining 'logical consequence' for formalized languages:

Semantic Definition of 'logical consequence'

σ is a logical consequence of Γ (in a formalized language L)

iff

there is no model (for L) in which all the sentences of Γ are true and σ is false. (Tarski 1936b)

This definition (which can easily be converted to a semantic definition of other metalogical notions – 'logical truth,' 'logical consistency,' etc.) played a critical role in turning *logical semantics*, or *model theory*, into one of the two main branches of contemporary (meta-)logic.

5 Three Criticisms of Tarski's Theory

While Tarski's theory of truth is widely viewed as one of the prime achievements of twentieth-century analytic philosophy, its philosophical significance has been repeatedly questioned. Among the main criticisms of Tarski's theory are: (A) Tarski's hierarchical solution to the Liar Paradox is applicable to artificial languages but not to "natural" languages; (B) Tarski's theory relativizes truth to language; (C) Tarski's definitions of truth are trivial.

Limitations of the hierarchical solution

Many philosophers find Tarski's solution to the Liar Paradox unsatisfactory on the ground that it does not enable us to define truth for natural languages. These philosophers are not dissuaded by Tarski's claims that: (1) it is impossible to define truth for natural languages, since being universal, such languages are inconsistent (Tarski 1933: 164–5), and (2) the hierarchical solution accounts for, and legitimizes, the use of 'true' in many segments of natural language, namely, all segments which are open and can be represented by artificial languages whose structure is precisely specified. In particular, truth can be defined for all segments used in the formulation of scientific theories (Tarski 1944: 347; 1969: 68). Soames (1999), for example, rejects the claim that natural languages are inconsistent. Others point out that Tarski's solution is too strict: it eliminates not only paradoxical uses of 'true' and related notions (e.g. 'false') in discourse, but also legitimate uses of these notions. One example, due to Kripke (1975), is the following: Consider two sentences, the one uttered by Dean and the other by Nixon during the Watergate crisis:

(4) All of Nixon's utterances about Watergate are false,

and

(5) Everything Dean says about Watergate is false.

This pair of sentences is perfectly consistent, yet there is no room for it in Tarski's hierarchy: According to Tarski's principles, (4) must belong to a language higher in the hierarchy than the language to which (5) belongs, and (5) must belong to a language higher in the hierarchy than the language to which (4) belongs. But this is impossible.

Triviality and relativity to language

It is common to interpret Tarski's theory as a *reductionist* theory or, more specifically, a theory whose goal is to *reduce* the notion of truth for a given language to the satisfaction conditions of the atomic formulas (the denotation conditions of the nonlogical constants) of this language. (To simplify the discussion I will ignore the case of atomic sentences containing logical constants, i.e. Identity). Given a language L, we determine the truth value of sentences of L by first listing the denotations of the primitive nonlogical constants of L, and then applying the recursive 'instructions' in the definition of truth for L to these lists. For example, if L is a language with two primitive nonlogical constants, an individual constant, '*a*,' and a 1-place predicate, 'P,' whose denotations are the number 1 and the set of all even natural numbers, respectively, we first prepare a denotation list for L, <'*a*,'1>, <'P,'{0,2,4,6, . . . }>, and then we calculate the truth value of sentences of L by applying the recursive rules in the definition of truth to this list: 'P*a*' is true (in L) iff $1 \in \{0,2,4,6, \ldots\}$, '~P*a*' is true (in L) iff 'P*a*' is false (in L), that is iff $1 \notin \{0,2,4,6, \ldots\}$, etc.

Two influential criticisms, based on this analysis, are: (1) Tarski's notion of truth is trivial; (2) Tarski's notion of truth is relative to language.

The triviality criticism

Tarski's definition of truth for a language L reduces the truth of sentences of L to the satisfaction of atomic formulas of L. But its treatment of atomic satisfaction is utterly uninformative. Instead of identifying a feature (or features) in virtue of which an object (an n-tuple of objects) satisfies a given atomic formula, it says that an object satisfies an atomic formula iff it belongs to a certain list. (In the above example, an object satisfies 'Px' iff it belongs to the list 0,2,4, ...) But a definition of this kind is a definition by *enumeration* ('x is a P iff x is 0 or is 2 or x is 4 or . . .'), and as such it lacks informative value.

This criticism is forcefully articulated in Field (1972). Field likens Tarski's definition of satisfaction to a definition by enumeration of a scientific concept. Consider, for example, a definition by enumeration of the concept *valence:*

$$(\forall x)\{\text{Valence}(x) = n \equiv [(x = \text{potassium} \,\&\, n = +1) \vee \ldots \vee (x = \text{sulfur} \,\&\, n = -2)]\}.$$

The valence of a chemical element is an integer which represents the sort of chemical combinations the element will enter into based on its physical properties. A definition associating valences with physical properties of elements would be highly informative; a definition by enumeration, on the other hand, would be utterly trivial. (Expanding the definition from chemical elements to configurations of chemical elements by using

recursive entries will not change the situation: if the 'base' is trivial, the definition as a whole is trivial.)

Although Field is particularly concerned with one aspect of the Tarskian project, namely its success in reducing semantic notions to nonsemantic (specifically, physicalistic) notions, his criticism is not restricted to this aspect. The standards used in philosophy, Field says, should not be lower than those used in other sciences, and a method for defining truth by enumeration "has no philosophical interest whatsoever" (Field 1972: 102).

The relativity criticism

Another criticism of Tarski's theory (based on the above interpretation) concerns its relativization of truth to language. The argument can be summed up as follows: Tarski's method generates definitions of truth for particular languages, where (as we have seen before) the notion of truth for a given language is based on a list of denotations specific to that language (i.e. a list which cannot serve as a basis of a definition of truth for any other language). For that reason, Tarski's notion of truth is *relative to language.* Blackburn (1984: 267) compares Tarski's definitions of 'true in L_1,' 'true in L_2,' . . . , to definitions of 'well-grounded verdict on Monday,' 'well-grounded verdict on Tuesday,' . . . In the same way that the latter would not amount to a definition of the *absolute* notion 'well-grounded verdict,' so Tarski's definitions do not amount to a definition of the *absolute* notion 'true'. Just as there is no philosophical interest in the *relative* jurisprudential notion 'well-grounded verdict on day X,' so there is no philosophical interest in the *relative* semantic notion 'true in L.'

While the criticisms of Tarski's hierarchical solution to the Liar Paradox have motivated philosophers to construct new, nonhierarchical solutions to that paradox, the triviality and relativity criticisms have led many philosophers to give up hope of an informative theory of truth. Below I will describe a nonhierarchical solution to the Liar Paradox, due to Kripke, and I will offer a new interpretation of Tarski's theory as an informative theory, immune to the relativity and triviality criticisms.

6 Kripke's Solution to the Liar Paradox

In a 1975 paper, "An outline of a Theory of Truth," Kripke offered a new, nonhierarchical solution to the Liar Paradox. The idea underlying Kripke's proposal is this: Instead of defining truth for an infinite hierarchy of languages that do not contain their own truth predicate, we can define truth for a single language that does contain its own truth predicate in an infinite number of stages. In Tarski's method we start with a language L_0 which does not contain its own truth predicate, and construct stronger and stronger languages, L_1, L_2, . . . , each containing a truth predicate, T_1, T_2, T_3, . . . , for the previous language in the hierarchy. In Kripke's method we have a single language, L, which contains its own unique truth predicate, T, and we define the extension of T (i.e. the set of all sentences of L satisfying 'Tx') in stages: S_0, S_1, S_2, S_3, . . .

The definition of T proceeds by constructing two sets: Σ_1 – the extension of T, and Σ_2 – the counter-extension of T. Σ_1 is the set of all true sentences of L in the domain D of L. Σ_2 is the set of all false sentences of L in D plus all objects in D which are not sen-

tences of L. (D may contain codes of sentences of L instead of sentences of L, but for the sake of simplicity I will assume it contains (only) the latter.) Let us think of L as a union of a Tarskian hierarchy, $\cup\{L_0, L_1, L_2, \ldots\}$, where '$T_1$', '$T_2$', '$T_3$', ... represent partial applications of T. Σ_1 and Σ_2 are constructed in stages as follows:

Stage 0: $\Sigma_1 = \varnothing$
$\Sigma_2 = \{a \in D: a$ is not a sentence of L$\}$

Stage 1: $\Sigma_1 = \{a \in D: a$ is a true sentence of L_0 **or** a is a true sentence of L whose truth value is logically determined based on the truth value of sentences of $L_0\}$
$\Sigma_2 = \{a \in D: a$ is a false sentence of L_0 **or** a is a false sentence of L whose truth-value is logically determined based on the truth-value of sentences of L_0 **or** a is not a sentence of L$\}$

Stage 2: $\Sigma_1 = \{a \in D: a$ is a true sentence of L_0 or L_1, **or** a is a true sentence of L whose truth-value is logically determined based on the truth-value of sentences of L_0 or $L_1\}$
$\Sigma_2 = \{a \in D: a$ is a false sentence of L_0 or L_1, **or** a is a false sentence of L whose truth-value is logically determined based on the truth-value of sentences of L_0 or L_1, **or** a is not a sentence of L$\}$

Thus, if 'Snow is white' and 'Snow is green' are sentences of L, then since both belong to the L_0 part of L, in stage 0 neither belongs to Σ_1 or Σ_2. In stage 1, 'Snow is white' and '~ Snow is green' are among the sentences added to Σ_1, and 'Snow is green' and '~ Snow is white' are among the sentences added to Σ_2. In stage 2, 'T "Snow is white"' and 'T "~ Snow is green"' are among the sentences added to Σ_1, and 'T "~ Snow is white"' and 'T "Snow is green"' are among the sentences added to Σ_2. And so on. The list of stages can be extended into the transfinite, using standard set theoretic methods. Thus we can have transfinite stages ω, $\omega + 1$, $\omega + 2$, ..., (where ω is the smallest infinite ordinal), including higher limit ordinals. (The details of the transfinite stages can be omitted.)

Throughout the finite stages, Σ_1 and Σ_2 are continuously extended and their extensions are *forced* by (1) the rules for the nonlogical, nonsemantic primitive constants of L (i.e. the rules determining the denotations of these constants and the truth/satisfaction of sentences/formulas composed of these constants (and, possibly, variables) – eventually, facts about what constant denotes what object, property or relation, what object has what nonlogical property and/or what objects stand in what nonlogical relation); (2) the rules for the logical constants of L; and (3) the rules for the semantic constants of L. (See Rules I–III below.) Thus, 'Snow is white' and 'NOT snow is green' must be added to Σ_1 in Stage 1 (due to facts concerning the denotations of 'snow,' 'white,' and 'green' and the color of snow, as well as the semantic rule for 'NOT'), 'True "Snow is white"' must be added to Σ_1 in Stage 2 (due to the semantic rule for 'true' and the fact that 'Snow is white' belongs to Σ_1 in stage 1), 'True "True 'Snow is white' "' must be added to Σ_1 in Stage 3 (due to the rule for 'true' and the fact that 'True "Snow is white"' belongs to Σ_1 in Stage 2), etc. And similarly for Σ_2. We say that all the sentences placed in Σ_1 and Σ_2 in the finite stages are *grounded*. However, since no sentence of L contains infinitely many occurrences of 'T,' and in particular, infinitely many embedded occurrences of 'T' (or other semantic predicates), eventually we arrive at a stage in

which neither Σ_1 nor Σ_2 is properly extended. We call such a stage a *fixed point*. It is important to note that not all sentences of L belong to either Σ_1 or Σ_2 in the *least fixed point*. For example, Liar sentences as well as sentences like

(10) T(10)

do not. How does Kripke deal with such sentences?

To deal with paradoxical sentences Kripke constructs T as a *partial* truth-predicate and L as a language with *truth-value gaps*: some sentences of L are either in the extension of T or in its anti-extension, but other sentences are in neither; some sentences of L have a truth value, others do not. All paradoxical sentences are truth-valueless in Kripke's semantics, but sentences like (10) can either be assigned a truth value (True or False) in later stages, or remain truth-valueless.

I will not formulate Kripke's semantics for L in detail here. But the following are its main principles:

I *Rules for determining the denotation, satisfaction and truth-value of expressions of L_0 (the L_0 part of L)*
Same as in Tarski's semantics.

II *Rules for determining the truth-value and satisfaction of sentences and formulas of L governed by logical constants*
Based on Kleene's strong 3-valued semantics. (Coincides with Tarski's semantics in the bivalent part of L, in particular, in the L_0 part of L.)

Let σ_1 and σ_2 be sentences of L. Then:

	true if	σ_1 is false
$\ulcorner\sim\sigma_1\urcorner$ is	false if	σ_1 is true
	undefined otherwise	

	true if	at least one of σ_1 and σ_2 is true
$\ulcorner\sigma_1 \vee \sigma_2\urcorner$ is	false if	both σ_1 and σ_2 are false
	undefined otherwise	

Let Φ be a formula of L, let g be an assignment function (as in Section 3), and let us use 'Φ is true under g' for 'g satisfies Φ'. Then:

	true	Φ is true under every g′ which differs from g at most in g_i
$\ulcorner\forall v_i\Phi\urcorner$ is	false under g if	Φ is false under some g′ which differs from g at most in g_i
	undefined otherwise	

III *Semantic rule for sentences governed by the truth predicate, T, of L (Kripke's version of Criterion (T)):*

Let σ be a sentence of L and s a name of σ in L. Then:

$$\ulcorner T(s) \urcorner \text{ is } \begin{cases} \text{true} & \text{iff} \quad \sigma \text{ is true} \\ \text{false} & \text{iff} \quad \sigma \text{ is false} \end{cases}$$

The definition of T can be viewed as completed in any of the fixed-points. If we view it as completed in the least fixed-point, then only grounded sentences are in the extension of T. If we see it as completed in later fixed-points, some ungrounded sentences (e.g. (10)) may also be in the extension of T. Paradoxical sentences are never in the extension of T.

Two noteworthy features of Kripke's method are: (1) it does not uniquely determine the truth predicate of a given closed language; and (2) it allows empirical circumstances to determine whether a sentence is paradoxical in a given language. The first point should be clear by now: the semantic status of some sentences (i.e. being true, false, or truth-valueless) is 'forced' by the semantic rules, that of others is a matter of choice or convention. Grounded and paradoxical sentences fall under the first category, ungrounded and unparadoxical sentences fall under the second.

The role of empirical circumstances

One important intuition captured by Kripke's proposal is that semantic properties of sentences (being true, false, ungrounded, paradoxical, etc.) are often determined by empirical circumstances. Consider, for example, the sentence

(11) $(\forall x)(Px \supset Tx)$

of a Kripkean language L. If P is an empirical predicate satisfied by exactly one object, *a*, then: if *a* = 'Snow is white,' (11) is true; if *a* = 'Snow is green,' (11) is false; if *a* = (11), (11) is ungrounded; and so on. And these semantic features hold or do not hold of (11) empirically. The same applies to

(12) $(\forall x)(Px \supset \sim Tx)$.

If the only object satisfying 'Px' is 'Snow is green,' (12) is true; if it is 'Snow is white,' (12) is false; if it is (12) itself, (12) is paradoxical. And the truth, falsity, or paradoxicality of (12) are due to empirical circumstances. In making statements, Kripke observes, we often take a risk. Under certain circumstances a sentence is grounded and true, under others – ungrounded and paradoxical.

This feature of Kripke's theory enables it to assign a truth value to sentences which (in the specific circumstances of their utterance) are not paradoxical, yet are regarded by Tarski as illegitimate. Let us go back to (4) and (5). If at least one statement made by Dean about Watergate is true and all Nixon's statements about Watergate other than (5) are false, then (4) is true and (5) is false.

The ghost of Tarski

While Kripke's method provides a semantics for languages containing their own truth predicate, the account itself is carried out in a Tarskian meta-language. Furthermore, some truths about sentences of a given Kripkean language L are, though expressible in L, true only in its meta-language, ML. Thus, if σ is a Liar sentence of L, the statements 'σ is not true,' 'σ is ungrounded' and 'σ is paradoxical' are true in ML but lack a truth value in L. In Kripke's words: "The ghost of the Tarski hierarchy is still with us" (Kripke 1975: 714).

Kripke's relegation of certain truths to the meta-language is not accidental. It is the means by which he avoids the so-called *strengthened Liar paradox*. The strengthened Liar paradox arises in languages with truth-value gaps as follows: Let

(13) ~T(13)

be a sentence of a 3-valued language L and let T be a truth predicate of L satisfying Kripke's version of Criterion T. Then: T((13)) *iff* (13) *iff* ~T(13).

Kripke avoids the strengthened Liar paradox by rendering (13) undefined but its meta-linguistic correlate, 'the sentence (13) of L is not true,' true. This means that Kripke's method falls short of providing a complete semantics for natural languages which, being universal, have no richer meta-languages.

Kripke's solution to the Liar Paradox is not the only alternative to Tarski's solution. For other alternatives see Martin (1984), Gupta and Belnap (1993), and others.

7 A Reinterpretation of Tarski's Theory

The deflationist approach to truth

The view that the base entries in Tarski's definitions render them uninformative has led some philosophers to search for an informative base for Tarski's definitions. Field (1972) suggested that instead of using lists of reference as a basis for a definition of truth, we use a general, informative theory of reference as such a basis, and pointed to Kripke's (1972) outline of a causal theory of reference as a promising starting point. But the slow progress and difficulties involved in the development of an informative and general theory of reference led Field (1986) and others to adopt a so-called *deflationist* or *minimalist* attitude towards truth.

The deflationist attitude is reflected by such statements as:

> [T]ruth is entirely captured by the initial triviality [that each proposition specifies its own condition for being true (e.g. the proposition *that snow is white* is true if and only if *snow is white*)]. (Horwich, 1990: xi)
>
> Unlike most other properties, *being true* is insusceptible to conceptual or scientific analysis. (*Ibid.*: 6)

> [The theory of truth] contains no more than what is expressed by the uncontroversial instances of the equivalence schema,
>
> (E) It is true *that p* if and only if *p*. (*Ibid.*: 6–7)

While deflationists differ on many issues, most agree that a theory of truth need not be more informative than Tarski's theory. Some would like to extend Tarski's definitions to a greater variety of linguistic structures: indexicals, adverbs, propositional attitudes, modal operators, etc., but none requires a more substantive analysis. According to deflationists, "the traditional attempt to discern the *essence* of truth – to analyze that special quality which truths supposedly have in common – is just a pseudo-problem". (Horwich, 1990: 6) There is no substantive common denominator of all truths, and therefore there is no substantive theory of truth. The task of a theory of truth is to generate a list of all instances of the Equivalence schema, and regardless of how this list is generated, the theory of truth is still a collection of trivialities.

Critique of the deflationist approach

The deflationist approach is based on a traditional conception of theories: A theory of a concept X is a theory of the common denominator of all objects falling under X. If the common denominator of all these objects is trivial, X is trivial and a theory of X is a collection of trivialities. This conception of a philosophical theory is, however, based on an unfounded assumption: namely, that the content of a given concept X is the common denominator of all instances of X. It is quite clear that the content of some concepts is not exhausted, or even close to being exhausted, by the common denominator of their instances. The concept of *game* is a case in point (Wittgenstein, 1958). Yet if 'game' is not a common-denominator concept, it is clearly not an empty or a trivial concept. And neither is a theory of games empty or trivial. A theory of games may not be able to condense all there is to say about games into a single principle, expressible by a single formula, but it could identify a number of significant principles governing games and describe their nature, workings, interrelations, and consequences in a general and informative manner.

The question arises as to whether Tarski's theory of truth is – or can be made to be – substantive in this (non-traditional) sense.

What does Tarski's theory actually accomplish?

One thing that both defenders and critics of Tarski's theory agree about is its substantial contribution to logic (see above). Now, it is striking that Tarski's theory does not make similar contributions to other disciplines. While Tarski's definition of truth for a language L yields, all by itself, a definition of *logical consequence* for L (assuming ML has a sufficiently rich set-theoretical apparatus), it does not yield (all by itself) definitions of *epistemic*, *modal*, *physical*, or *biological consequence* for L. (Examples of the latter kinds of consequence are: '*a* knows that P; therefore, *a* believes that P,' 'Necessarily P; therefore Possibly P,' 'The force exerted on body *a* at time *t* is zero; therefore the acceleration of *a* at *t* is zero,' '*a* is a human female; therefore *a* does not have a Y chromosome,' etc.)

Why does Tarski's theory yield an account of *logical* consequence, but not of other types of consequence? What features should a theory of truth have in order to yield a concept of consequence of type X?

The answer to this question is quite clear. A consequence relation is a relation of preservation (or transmission) of truth: If C stands in a consequence relation R to a set of sentences, Γ, and all the sentences of Γ are true, then their truth is preserved through R (or is transmitted to C through R). If R is a relation of consequence of type X, the preservation (or transmission) of truth is due to the *X-structure* of the sentences of Γ and X, that is due to the *content* and *organization* of constants of type X in these sentences (where for non-X constants, only their identities and differences, but not their content or interrelations, play a role). Thus, if C stands to Γ in a relation of *logical* consequence, this is due (except in the trivial case of $C \in \Gamma$) to the *logical* structure of the sentences involved; if C stands to Γ in the relation of modal, epistemic, physical, or biological consequence, this is due to the modal, epistemic, physical, or biological structure of those sentences. To yield a definition of consequence of type X for a language L, a definition of truth for L has to specify the contribution of X-structure to the truth value of sentences of L. Tarski's definition of truth for a language L is tuned to the *logical* structure of sentences of L: therefore, it gives rise to the notion of *logical* consequence for L. (Note that due to the generality of logic, it is common to conceive of non-logical consequences of type X as based not only on the content and interrelations of the X vocabulary, but also on the interrelations of the X vocabulary and the logical vocabulary. Yet what renders these consequences X-consequences is the role played by the X-vocabulary.)

These observations suggest that what Tarski's theory actually accomplishes is an account of the *contribution of logical structure to truth*. Tarski's theory tells us how the logical structure of a given sentence affects its truth value, not how other types of structure (modal, physical, . . .) do. Tarski's theory, on this interpretation, is a theory of a specific, albeit basic and general constituent of truth, namely, its logical constituent. Its goal is to describe, in an exhaustive, systematic and informative manner, that part of the truth-conditions of sentences which is due to their logical structure. This interpretation explains why Tarski's theory of *truth* is so important and fruitful in *logic*. Furthermore, it shields Tarski's theory from the relativity and triviality criticisms.

Relativity

While the role played by nonlogical constituents of sentences in determining their truth conditions is relative to language (in Tarski's theory), the role of the logical constituents is not. The denotation lists for the nonlogical constants vary from one Tarskian language to another, but the semantic rules for the logical constants are fixed across languages. The difference between Tarski's treatment of logically-structured and nonlogically-structured formulas of a given language is a difference between *rule* and *applications*. To calculate the truth value of a sentence – say, 'John loves Mary and John loves Jane' – of a Tarskian language L we take the *fixed* truth condition associated with 'and' in Tarski's method and apply it to the truth conditions of 'John loves Mary' and 'John loves Jane' in L. We may say that the *principles* governing the contribution of logical structure to truth are *absolute*: their *instances* (*applications*) – *relative* to language.

But this is the case with any theory: the rule of, say, addition, is the same in all applications of arithmetic, but in biology this rule operates on sets (quantities) of biological entities, while in theoretical physics it operates on sets (quantities) of abstract physical entities.

Triviality

The triviality criticism, like the relativity criticism, is directed at Tarski's treatment of the nonlogical constituents of truth. Considering Tarski's definition of truth for a given language L, the claim is that the satisfaction and denotation conditions for formulas and terms with no logical constants of L are given by enumeration (i.e. based on lists), and as such they trivialize the entire definition. While this criticism is warranted with respect to the first interpretation of Tarski's theory, it is unwarranted with respect to the second. On the first interpretation, Tarski's theory is a *reductionist* theory. Its task is to reduce the notion of truth for a given language to the satisfaction and denotation conditions of its nonlogically-structured formulas and its nonlogical constants. As such, the burden of informativeness falls on its *nonlogical* entries. Since these are trivial, the definition as a whole is trivial. But on the second, *logical* interpretation, the burden of informativeness falls on the *logical* entries. (The nonlogical entries play a merely auxiliary role.) So long, and to the extent that, the logical entries are informative, the definitions of truth are informative.

Are the logical entries in Tarski's definitions informative? To be informative, the logical entries have to describe the truth conditions associated with different logical structures based on principles, rather than by enumeration. Now, on a first reading, the logical entries in Tarski's definitions are not very informative. Take the logical connectives. The entries for Negation and Disjunction essentially say that ⌜**not** σ⌝ is true iff σ is **not** true, and that ⌜σ **or** ς⌝ is true iff σ is true **or** ς is true. These entries do not *explain* the satisfaction conditions of 'not' and 'or'; they take them as given. ('Not' in the definiens merely repeats 'not' in the definiendum.) But on a less literal and more charitable interpretation we may view the entries for the logical connectives as implicitly referring to the highly informative Boolean, or truth-functional, account of these connectives. The Boolean account provides (1) an informative a criterion of logicality for connectives, and (2) a systematic characterization of the satisfaction conditions of each logical connective based on this criterion. According to this characterization, Negation is characterized by a 1-place Boolean function, $f_{\neg}$, defined by: $f_{\neg}(T) = F$ and $f_{\neg}(F) = T$, Disjunction is characterized by 2-place function $f_{\vee}$, defined by: $f_{\vee}(T,T) = f_{\vee}(T,F) = f_{\vee}(F,T) = T$ and $f_{\vee}(F,F) = F$, and these definitions are precise and informative. In 1933 there did not exist an analogous criterion for logical predicates and quantifiers, but in later years such a criterion, and a systematic characterization of the satisfaction conditions of individual logical predicates and quantifiers based on it, have been developed. (See Mostowski 1957; Lindström 1966; Tarski 1966; Sher 1991 and others.) Today, therefore, it is possible to avoid the triviality criticism altogether by expanding Tarski's definitions to languages containing any logical constant satisfying this criterion and constructing (interpreting) the satisfaction entries for the logical constants as referring to the informative characterizations of these constants based on this criterion. (For further details and examples see Sher 1999b, Sections 6, 7, and 9).

8 Truth Beyond Logic

Aside from its direct contributions to pure logic, Tarski's work on truth has indirectly contributed to other fields as well. Kripke (1963) developed a semantics for modal logic which incorporates elements from Tarski's logical semantics; Hintikka (1962) and others developed a semantics for epistemic statements based on Tarski's semantics; Davidson (1980, 1984) has begun an influential project of developing a general theory of meaning for natural languages based on Tarski's method; etc. How far Tarski's theory can be extended beyond logic without losing its informativeness is an open question.

References

Aristotle (1941) *Metaphysics. The Basic Works of Aristotle*, ed. R. McKeon. New York: Random House.

Blackburn, S. (1984) *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Oxford University Press.

Davidson, D. (1980) *Actions and Events*. Oxford: Oxford University Press.

Davidson, D. (1984) *Truth and Interpretation*. Oxford: Oxford University Press.

Devitt, M. (1984) *Realism and Truth*. Oxford: Blackwell.

Enderton, H. B. (1972) *A Mathematical Introduction to Logic*. San Diego: Academic Press.

Field, H. (1972) Tarski's theory of truth. *Journal of Philosophy*, 69, 347–75.

Field, H. (1986) The deflationary conception of truth. *Fact, Science, and Modality*, eds. G. MacDonald and C. Wright. Oxford: Blackwell, 55–117.

Gupta, A. and Belnap, N. (1993) *The Revision Theory of Truth*. Cambridge, MA: MIT.

Hintikka, J. (1962) *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell.

Kripke, S. (1963) Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16, 83–94.

Kripke, S. (1972) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kripke, S. (1975) Outline of a theory of truth. *Journal of Philosophy*, 72, 690–716.

Lindström, P. (1966) First order predicate logic with generalized quantifiers. *Theoria*, 32, 186–95.

Martin, R. L. (ed.) (1984) *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford University Press.

Mostowski, A. (1957) On a generalization of quantifiers. *Fundamenta Mathematicae*, 44, 12–36.

Quine, W. V. (1951) *Mathematical Logic*, revised edn. Cambridge, MA: Harvard University Press.

Ramsey, F. (1927) Facts and propositions. *The Foundations of Mathematics*. Paterson, NJ: Littlefield, Adams, 1960, 138–55.

Sher, G. (1991) *The Bounds of Logic: A Generalized Viewpoint*. Cambridge, MA: MIT.

Sher, G. (1999a) On the possibility of a substantive theory of truth. *Synthese*, 117, 133–72.

Sher, G. (1999b) Is logic a theory of the obvious? *European Review of Philosophy*, 4, 207–38.

Soames, S. (1999) *Understanding Truth*. New York: Oxford University Press.

Tarski, A. (1933) The concept of truth in formalized languages. In Tarski (1983) 152–278.

Tarski, A. (1936a) The establishment of scientific semantics. In Tarski (1983) 401–8.

Tarski, A. (1936b) On the concept of logical consequence. In Tarski (1983) 409–20.

Tarski, A. (1944) The semantic conception of truth. *Philosophy and Phenomenological Research*, 4, 341–76.

Tarski, A. (1966) What are logical notions? *History and Philosophy of Logic*, 7, 143–54.
Tarski, A. (1969) Truth and proof. *Scientific American*, 220, 63–77.
Tarski, A. (1983) *Logic, Semantics, Metamathematics*. 2nd edn. Indianapolis, IN: Hackett.
Vaught, R. L. (1974) Model theory before 1945. *Proceedings of the Tarski Symposium*, eds. L. Henkin *et al.* Providence, RI: American Mathematical Society.
Wittgenstein, L. (1958) *Philosophical Investigations*. 2nd edn. Oxford: Basil Blackwell.