**Outline of a Theory of Truth**

Saul Kripke

*The Journal of Philosophy*, Vol. 72, No. 19, Seventy-Second Annual Meeting American Philosophical Association, Eastern Division. (Nov. 6, 1975), pp. 690-716.

Stable URL:
http://links.jstor.org/sici?sici=0022-362X%2819751106%2972%3A19%3C690%3AOOATOT%3E2.0.CO%3B2-7

*The Journal of Philosophy* is currently published by Journal of Philosophy, Inc..

ternal modifications grounding relations (iii c) are relational properties containing relata. On view (A) above, these internalized relata may be complete concepts. On both (A) and (B), substances, or their representations in substances, somehow include each other, thus violating something akin to the set-theoretical axiom of regularity (Harry Teichert). This complexity strengthens the value of the reduction in "Plato's *Phaedo* Theory of Relations" *(op. cit.)* which illuminates (ii). Perhaps (iii c) should be interpreted as (ii).

HECTOR-NERI CASTAÑEDA

Indiana University

## OUTLINE OF A THEORY OF TRUTH *

### I. THE PROBLEM

EVER since Pilate asked, "What is truth?" (*John* XVIII, 38), the subsequent search for a correct answer has been inhibited by another problem, which, as is well known, also arises in a New Testament context. If, as the author of the Epistle to Titus supposes (*Titus* I, 12), a Cretan prophet, "even a prophet of their own," asserted that "the Cretans are always liars," and if "this testimony is true" of all other Cretan utterances, then it seems that the Cretan prophet's words are true if and only if they are false. And any treatment of the concept of truth must somehow circumvent this paradox.

The Cretan example illustrates one way of achieving self-reference. Let $P(x)$ and $Q(x)$ be predicates of sentences. Then in some cases empirical evidence establishes that the sentence '$(x)(P(x) \supset Q(x))$' [or '$(\exists x)(P(x) \wedge Q(x))$', or the like] itself satisfies the predicate $P(x)$; sometimes the empirical evidence shows that it is the *only* object satisfying $P(x)$. In this latter case, the sentence in question "says

* To be presented in an APA symposium on Truth, December 28, 1975.

Originally it was understood that I would present this paper orally without submitting a prepared text. At a relatively late date, the editors of this JOURNAL requested that I submit at least an "outline" of my paper. I agreed that this would be useful. I received the request while already committed to something else, and had to prepare the present version in tremendous haste, without even the opportunity to revise the first draft. Had I had the opportunity to revise, I might have expanded the presentation of the basic model in sec. III so as to make it clearer. The text shows that a great deal of the formal and philosophical material, and the proofs of results, had to be omitted.

Abstracts of the present work were presented by title at the Spring, 1975, meeting of the Association for Symbolic Logic held in Chicago. A longer version was presented as three lectures at Princeton University, June, 1975. I hope to publish another more detailed version elsewhere. Such a longer version should contain technical claims made here without proof, and much technical and philosophical material unmentioned or condensed in this outline.

of itself" that it satisfies $Q(x)$. If $Q(x)$ is the predicate[1] 'is false', the Liar paradox results. As an example, let $P(x)$ abbreviate the predicate 'has tokens printed in copies of the *Journal of Philosophy*, November 6, 1975, p. 691, line 5'. Then the sentence:

$$(x)(P(x) \supset Q(x))$$

leads to paradox if $Q(x)$ is interpreted as falsehood.

The versions of the Liar paradox which use empirical predicates already point up one major aspect of the problem: *many, probably most, of our ordinary assertions about truth and falsity are liable, if the empirical facts are extremely unfavorable, to exhibit paradoxical features.* Consider the ordinary statement, made by Jones:

(1) Most (i.e., a majority) of Nixon's assertions about Watergate are false.

Clearly, nothing is intrinsically wrong with (1), nor is it ill-formed. Ordinarily the truth value of (1) will be ascertainable through an enumeration of Nixon's Watergate-related assertions, and an assessment of each for truth or falsity. Suppose, however, that Nixon's assertions about Watergate are evenly balanced between the true and the false, except for one problematic case,

(2) Everything Jones says about Watergate is true.

Suppose, in addition, that (1) is Jones's sole assertion about Watergate, or alternatively, that all his Watergate-related assertions except perhaps (1) are true. Then it requires little expertise to show

[1] I follow the usual convention of the "semantic" theory of truth in taking truth and falsity to be predicates true of sentences. If truth and falsity primarily apply to propositions or other nonlinguistic entities, read the predicate of sentences as "expresses a truth."

I have chosen to take sentences as the primary truth vehicles *not* because I think that the objection that truth is primarily a property of propositions (or "statements") is irrelevant to serious work on truth or to the semantic paradoxes. On the contrary, I think that ultimately a careful treatment of the problem may well need to separate the "expresses" aspect (relating sentences to propositions) from the "truth" aspect (putatively applying to propositions). I have not investigated whether the semantic paradoxes present problems when directly applied to propositions. The main reason I apply the truth predicate directly to linguistic objects is that for such objects a mathematical theory of self-reference has been developed. (See also footnote 32.)

Further, a more developed version of the theory would allow languages with demonstratives and ambiguities and would speak of utterances, sentences under a reading, and the like, as having truth value. In the informal exposition this paper does not attempt to be precise about such matters. Sentences are the official truth vehicles, but informally we occasionally talk about utterances, statements, assertions, and so on. Occasionally we may speak as if every utterance of a sentence in the language makes a statement, although below we suggest that a sentence may fail to make a statement if it is paradoxical or ungrounded. We are precise about such issues only when we think that imprecision may create confusion or misunderstanding. Like remarks apply to conventions about quotation.

that (1) and (2) are both paradoxical: they are true if and only if they are false.

The example of (1) points up an important lesson: it would be fruitless to look for an *intrinsic* criterion that will enable us to sieve out—as meaningless, or ill-formed—those sentences which lead to paradox. (1) is, indeed, the paradigm of an ordinary assertion involving the notion of falsity; just such assertions were characteristic of our recent political debate. Yet no syntactic or semantic feature of (1) guarantees that it is unparadoxical. Under the assumptions of the previous paragraph, (1) leads to paradox.[2] Whether such assumptions hold depends on the empirical facts about Nixon's (and other) utterances, not on anything intrinsic to the syntax and semantics of (1). (Even the subtlest experts may not be able to avoid utterances leading to paradox. It is said that Russell once asked Moore whether he always told the truth, and that he regarded Moore's negative reply as the sole falsehood Moore had ever produced. Surely no one had a keener nose for paradox than Russell. Yet he apparently failed to realize that if, as he thought, all Moore's *other* utterances were true, Moore's negative reply was not simply false but paradoxical.[3]) The moral: an adequate theory must allow our statements involving the notion of truth to be *risky*: they risk being paradoxical if the empirical facts are extremely (and unexpectedly) unfavorable. There can be no syntactic or semantic "sieve" that will winnow out the "bad" cases while preserving the "good" ones.

I have concentrated above on versions of the paradox using empirical properties of sentences, such as being uttered by particular people. Gödel showed essentially that such empirical properties are dispensable in favor of purely syntactic properties: he showed that, for each predicate $Q(x)$, a syntactic predicate $P(x)$ can be produced such that the sentence $(x)(P(x) \supset Q(x))$ is demonstrably the only object satisfying $P(x)$. Thus, in a sense, $(x)(P(x) \supset Q(x))$ "says of itself" that it satisfies $Q(x)$. He also showed that elementary syntax can be interpreted in number theory. In this way, Gödel put the issue of the legitimacy of self-referential sentences beyond doubt; he showed that they are as incontestably legitimate as arithmetic itself. But the examples using empirical predicates retain their importance: they point up the moral about riskiness.

[2] Both Nixon and Jones may have made their respective utterances without being aware that the empirical facts make them paradoxical.

[3] On an ordinary understanding (as opposed to the conventions of those who state Liar paradoxes), the question lay in the sincerity, not the truth, of Moore's utterances. Paradoxes could probably be derived on this interpretation also.

A simpler, and more direct, form of self-reference uses demonstratives or proper names: Let 'Jack' be a name of the sentence 'Jack is short', and we have a sentence that says of itself that it is short. I can see nothing wrong with "direct" self-reference of this type. If 'Jack' is not already a name in the language,[4] why can we not introduce it as a name of any entity we please? In particular, why can it not be a name of the (uninterpreted) finite sequence of marks 'Jack is short'? (Would it be permissible to call this sequence of marks "Harry," but not "Jack"? Surely prohibitions on naming are arbitrary here.) There is no vicious circle in our procedure, since we need not *interpret* the sequence of marks 'Jack is short' before we name it. Yet if we name it "Jack," it at once becomes meaningful and true. (Note that I am speaking of self-referential sentences, not self-referential propositions.[5])

In a longer version, I would buttress the conclusion of the preceding paragraph not only by a more detailed philosophical exposition, but also by a mathematical demonstration that the simple kind of self-reference exemplified by the "Jack is short" example could actually be used to prove the Gödel incompleteness theorem itself (and also, the Gödel-Tarski theorem on the undefinability of truth). Such a presentation of the proof of the Gödel theorem might be more perspicuous to the beginner than is the usual one. It also dispels the impression that Gödel was forced to replace direct self-reference by a more circumlocutory device. The argument must be omitted from this outline.[6]

It has long been recognized that some of the intuitive trouble with Liar sentences is shared with such sentences as

(3) (3) is true.

which, though not paradoxical, yield no determinate truth conditions. More complicated examples include a pair of sentences each one of which says that the other is true, and an infinite sequence of sentences $\{P_i\}$, where $P_i$ says that $P_{i+1}$ is true. In general, if a sentence such as (1) asserts that (all, some, most, etc.) of the sentences of a certain class $C$ are true, its truth value can be ascertained if the truth values of the sentences in the class $C$ are ascertained. If some of these sentences themselves involve the notion of truth, their truth value in turn must be ascertained by looking at *other* sentences,

[4] We assume that 'is short' *is* already in the language.

[5] It is *not* obviously possible to apply this technique to obtain "directly" self-referential *propositions*.

[6] There are several ways of doing it, using either a nonstandard Gödel numbering where statements can contain numerals designating their own Gödel numbers, or a standard Gödel numbering, plus added constants of the type of 'Jack'.

and so on. If ultimately this process terminates in sentences not mentioning the concept of truth, so that the truth value of the original statement can be ascertained, we call the original sentence *grounded*; otherwise, ungrounded.[7] As the example of (1) indicates, whether a sentence is grounded is not in general an intrinsic (syntactic or semantic) property of a sentence, but usually depends on the empirical facts. We make utterances which we hope will turn out to be grounded. Sentences such as (3), though not paradoxical, are ungrounded. The preceding is a rough sketch of the usual notion of groundedness and is not meant to provide a formal definition: the fact that a formal definition can be provided will be a principal virtue of the formal theory suggested below.[8]

## II. PREVIOUS PROPOSALS

Thus far the only approach to the semantic paradoxes that has been worked out in any detail is what I will call the "orthodox approach," which leads to the celebrated hierarchy of languages of Tarski.[9] Let $L_0$ be a formal language, built up by the usual operations of the first-order predicate calculus from a stock of (completely defined) primitive predicates, and adequate to discuss its own syntax (perhaps using arithmetization). (I omit an exact characterization.) Such a language cannot contain its own truth predicate, so a metalanguage $L_1$ contains a truth (really satisfaction) predicate $T_1(x)$ for $L_0$. (Indeed, Tarski shows how to define such a predicate in a higher-order language.) The process can be iterated, leading to a sequence $\{L_0, L_1, L_2, L_3, \ldots\}$ of languages, each with a truth predicate for the preceding.

Philosophers have been suspicious of the orthodox approach as an

[7] If a sentence asserts, e.g., that all sentences in class $C$ are true, we allow it to be false and grounded if one sentence in $C$ is false, irrespective of the groundedness of the other sentences in $C$.

[8] Under that name, groundedness seems to have been first explicitly introduced into the literature in Hans Hertzberger, "Paradoxes of Grounding in Semantics," this JOURNAL, XVII, 6 (March 26, 1970): 145–167. Hertzberger's paper is based on unpublished work on a "groundedness" approach to the semantic paradoxes undertaken jointly with Jerrold J. Katz. The intuitive notion of groundedness in semantics surely was part of the folklore of the subject much earlier. As far as I know, the present work gives the first rigorous definition.

[9] By an "orthodox approach", I mean any approach that works within classical quantification theory and requires all predicates to be totally defined on the range of the variables. Various writers speak as if the "hierarchy of languages" or Tarskian approach *prohibited* one from forming, for example, languages with certain kinds of self-reference, or languages containing their own truth predicates. On my interpretation, there are no *prohibitions*; there are only *theorems* on what can and cannot be done within the framework of ordinary classical quantification theory. Thus Gödel *showed* that a classical language can talk about its own syntax; using restricted truth definitions and other devices, such a language can say a great deal about its own semantics. On the other hand, Tarski *proved* that a classical language cannot contain its own truth predicate, and that a higher-order language can define a truth predicate for a language of lower order. None of this came from any a priori restrictions on self-reference other than those deriving from the restriction to a classical language, all of whose predicates are totally defined.

analysis of our intuitions. Surely our language contains just one word 'true', not a sequence of distinct phrases $\ulcorner \text{true}_n \urcorner$, applying to sentences of higher and higher levels. As against this objection, a defender of the orthodox view (if he does not dismiss natural language altogether, as Tarski inclined to do) may reply that the ordinary notion of truth is systematically ambiguous: its "level" in a particular occurrence is determined by the context of the utterance and the intentions of the speaker. The notion of differing truth predicates, each with its own level, seems to correspond to the following intuitive idea, implicit in the discussion of "groundedness" above. First, we make various utterances, such as 'snow is white', which do not involve the notion of truth. We then attribute truth values to these, using a predicate '$\text{true}_1$'. ('$\text{True}_1$' means—roughly—"is a true statement not itself involving truth or allied notions.") We can then form a predicate '$\text{true}_2$' applying to sentences involving '$\text{true}_1$', and so on. We may assume that, on each occasion of utterance, when a given speaker uses the word 'true', he attaches an implicit subscript to it, which increases as, by further and further reflection, he goes higher and higher in his own Tarski hierarchy.[10]

Unfortunately this picture seems unfaithful to the facts. If someone makes such an utterance as (1), he does *not* attach a subscript, explicit or implicit, to his utterance of 'false', which determines the "level of language" on which he speaks. An implicit subscript would cause no trouble if we were sure of the "level" of *Nixon's* utterances; we could then cover them all, in the utterance of (1) or even of the stronger

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

simply by choosing a subscript higher than the levels of any involved in Nixon's Watergate-related utterances. Ordinarily, however, a speaker *has no way of knowing the "levels" of Nixon's relevant utterances*. Thus Nixon may have said, "Dean is a liar," or "Haldeman told the truth when he said that Dean lied," etc., and the

[10] Charles Parsons, "The Liar Paradox," *Journal of Philosophical Logic*, III, 4 (October 1974): 380–412, may perhaps be taken as giving an argument like the one sketched in this paragraph. Much of his paper, however, may be regarded as confirmed rather than refuted by the present approach. See in particular his fn 19, which hopes for a theory that avoids explicit subscripts. The minimal fixed point (see sec. III below) avoids explicit subscripts but nevertheless has a notion of level; in this respect it can be compared with standard set theory as opposed to the theory of types. The fact that the levels are not intrinsic to the sentences is peculiar to the present theory and is additional to the absence of explicit subscripting.

The orthodox assignment of intrinsic levels guarantees freedom from "riskiness" in the sense explained in sec. I above. For (4) and (5) below, the very assignment of intrinsic levels which would eliminate their riskiness would also prevent them from "seeking their own levels" (see pp. 695–697). *If we wish to allow sentences to seek their own levels apparently we must also allow risky sentences.* Then we must regard sentences as *attempting* to express propositions, and allow truth-value gaps. See sec. 3 below.

"levels" of these may yet depend on the levels of Dean's utterances, and so on. If the speaker is forced to assign a "level" to (4) in advance [or to the word 'false' in (4)], he may be unsure how high a level to choose; if, in ignorance of the "level" of Nixon's utterances, he chooses too low, his utterance (4) will fail of its purpose. The idea that a statement such as (4) should, in its normal uses, have a "level" is intuitively convincing. It is, however, equally intuitively obvious that the "level" of (4) should not depend on the form of (4) alone (as would be the case if 'false'—or, perhaps, 'utterances'—were assigned explicit subscripts), nor should it be assigned in advance by the speaker, but rather its level should depend on the empirical facts about what Nixon has uttered. The higher the "levels" of Nixon's utterances happen to be, the higher the "level" of (4). This means that in some sense a statement should be allowed to seek its own level, high enough to say what it intends to say. It should not have an intrinsic level fixed in advance, as in the Tarski hierarchy.

Another situation is even harder to accommodate within the confines of the orthodox approach. Suppose Dean asserts (4), while Nixon in turn asserts

(5) Everything Dean says about Watergate is false.

Dean, in asserting the sweeping (4), wishes to include Nixon's assertion (5) within its scope (as one of the Nixonian assertions about Watergate which is said to be false); and Nixon, in asserting (5), wishes to do the same with Dean's (4). Now on any theory that assigns intrinsic "levels" to such statements, so that a statement of a given level can speak only of the truth or falsity of statements of lower levels, it is plainly impossible for both to succeed: if the two statements are on the same level, neither can talk about the truth or falsity of the other, while otherwise the higher can talk about the lower, but not conversely. Yet intuitively, we can often assign unambiguous truth values to (4) and (5). Suppose Dean has made at least one true statement about Watergate [other than (4)]. Then, independently of any assessment of (4), we can decide that Nixon's (5) is false. If all Nixon's other assertions about Watergate are false as well, Dean's (4) is true; if one of them is true, (4) is false. Note that in the latter case, we could have judged (4) to be false without assessing (5), but in the former case the assessment of (4) as true depended on a *prior* assessment of (5) as false. Under a different set of empirical assumptions about the veracity of Nixon and Dean, (5) would be true [and its assessment as true would depend on a prior

assessment of (4) as false]. It seems difficult to accommodate these intuitions within the confines of the orthodox approach.

Other defects of the orthodox approach are more difficult to explain within a brief outline, though they have formed a substantial part of my research. One problem is that of transfinite levels. It is easy, within the confines of the orthodox approach, to assert

(6) Snow is white.

to assert that (6) is true, that '(6) is true' is true, that ' "(6) is true" is true' is true, etc.; the various occurrences of 'is true' in the sequence are assigned increasing subscripts. It is much more difficult to assert that all the statements in the sequence just described are true. To do this, we need a metalanguage of transfinite level, above all the languages of finite level. To my surprise, I have found that the problem of defining the languages of transfinite level presents substantial technical difficulties which have never seriously been investigated.[11] (Hilary Putnam and his students essentially investigated—under the guise of a superficially completely different description and mathematical motivation—the problem for the special case where we start at the lowest level with the language of elementary number theory.) I have obtained various positive results on the problem, and there are also various negative results; they cannot be detailed here. But in the present state of the literature, it should be said that if the "theory of language levels" is meant to include an account of transfinite levels, then one of the principal defects of the theory is simply the *nonexistence* of the theory. The existing literature can be said to define "Tarski's hierarchy of languages" only for *finite* levels, which is hardly adequate. My own work includes an extension of the orthodox theory to transfinite levels, but it is as yet incomplete. Lack of space not only prevents me from describing the work; it prevents me from mentioning the mathematical difficulties that make the problem highly nontrivial.

Other problems can only be mentioned. One surprise to me was the fact that the orthodox approach by no means obviously guarantees groundedness in the intuitive sense mentioned above. The concept of truth for $\Sigma_1$ arithmetical statements is itself $\Sigma_1$, and this fact can be used to construct statements of the form of (3). Even if unrestricted truth definitions are in question, standard theorems easily allow us to construct a *descending* chain of first-order languages $L_0$, $L_1$, $L_2$, . . . , such that $L_i$ contains a truth predicate for $L_{i+1}$. I don't know whether such a chain can engender ungrounded sen-

[11] The problem of transfinite levels is perhaps not too difficult to solve in a canonical way at level $\omega$, but it becomes increasingly acute at higher ordinal levels.

tences, or even quite how to state the problem here; some substantial technical questions in this area are yet to be solved.

Almost all the extensive recent literature seeking alternatives to the orthodox approach—I would mention especially the writings of Bas van Fraassen and Robert L. Martin[12]—agrees on a single basic idea: there is to be only one truth predicate, applicable to sentences containing the predicate itself; but paradox is to be avoided by allowing truth-value gaps and by declaring that paradoxical sentences in particular suffer from such a gap. These writings seem to me to suffer sometimes from a minor defect and almost always from a major defect. The minor defect is that some of these writings criticize a strawmannish version of the orthodox approach, not the genuine article.[13] The major defect is that these writings almost invariably are mere suggestions, not genuine theories. Almost never is there any precise semantical formulation of a language, at least rich enough to speak of its own elementary syntax (either directly or via arithmetization) and containing its own truth predicate. Only if such a language were set up with formal precision could it be said that a theory of the semantic paradoxes has been presented. Ideally, a theory should show that the technique can be applied to arbitrarily rich languages, no matter what their "ordinary" predicates other than truth. And there is yet another sense in which the orthodox approach provides a theory while the alternative literature does not. Tarski shows how, for a classical first-order language whose quantifiers range over a set, he can give a *mathematical definition* of truth, using the predicates of the object language plus set theory (higher-order logic). The alternative literature abandons the attempt at a mathematical definition of truth, and is content to take it as an

[12] See Martin, ed., *The Paradox of the Liar* (New Haven: Yale, 1970) and the references given there.

[13] See fn 9 above. Martin, for example, in his papers "Toward a Solution to the Liar Paradox," *Philosophical Review*, LXXVI, 3 (July 1967): 279–311, and "On Grelling's Paradox," *ibid.*, LXXVII, 3 (July 1968): 325–331, attributes to "the theory of language levels" all kinds of restrictions on self-reference which must be regarded as simply refuted, even for classical languages, by Gödel's work. Perhaps there are or have been some theorists who believed that *all* talk of an object language must take place in a distinct metalanguage. This hardly matters; the main issue is: what constructions can be carried out within a classical language, and what require truth-value gaps? Almost all the cases of self-reference Martin mentions can be carried out by orthodox Gödelian methods without any need to invoke partially defined predicates or truth-value gaps. In fn 5 of his second paper Martin takes some notice of Gödel's demonstration that sufficiently rich languages contain their own syntax, but he seems not to realize that this work makes most of his polemics against "language levels" irrelevant.

At the other extreme, some writers still seem to think that some kind of general ban on self-reference is helpful in treating the semantic paradoxes. In the case of self-referential *sentences*, such a position seems to me to be hopeless.

intuitive primitive. Only one paper in the "truth-gap" genre that I have read—a recent paper by Martin and Peter Woodruff[14]—comes close even to beginning an attempt to satisfy any of these desiderata for a theory. Nevertheless the influence of this literature on my own proposal will be obvious.[15]

### III. THE PRESENT PROPOSAL

I do not regard any proposal, including the one to be advanced here, as definitive in the sense that it gives *the* interpretation of the ordinary use of 'true', or *the* solution to the semantic paradoxes. On the contrary, I have not at the moment thought through a careful philosophical justification of the proposal, nor am I sure of the exact areas and limitations of its applicability. I do hope that the model given here has two virtues: first, that it provides an area rich in formal structure and mathematical properties; second, that to a reasonable extent these properties capture important intuitions. The model, then, is to be tested by its technical fertility. It need not capture every intuition, but it is hoped that it will capture many.

Following the literature mentioned above, we propose to investigate languages allowing truth-value gaps. Under the influence of Strawson,[16] we can regard a sentence as an attempt to make a statement, express a proposition, or the like. The meaningfulness or well-formedness of the sentence lies in the fact that there are specifiable circumstances under which it has determinate truth conditions (expresses a proposition), not that it always does express a proposition. A sentence such as (1) is always *meaningful*, but under various circumstances it may not "make a statement" or "express a proposi-

[14] In the terminology of the present paper, the paper by Martin and Woodruff proves the existence of *maximal* fixed points (not the minimal fixed point) in the context of the weak three-valued approach. It does not develop the theory much further. I believe the paper is as yet unpublished, but is forthcoming in a volume dedicated to Yehoshua Bar-Hillel. Although it partially anticipates the present approach, it was unknown to me when I did the work.

[15] Actually I was familiar with relatively little of this literature when I began work on the approach given here. Even now I am unfamiliar with a great deal of it, so that tracing connections is difficult. Martin's work seems, in its formal consequences if not its philosophical basis, to be closest to the present approach.

There is also a considerable literature on three-valued or similar approaches to the set-theoretical paradoxes, with which I am not familiar in detail but which seems fairly closely related to the present approach. I should mention Gilmore, Fitch, Feferman.

[16] I am interpreting Strawson as holding that 'the present king of France is bald' fails to make a statement but is still meaningful, because it gives directions (conditions) for making a statement. I apply this to the paradoxical sentences, without committing myself on his original case of descriptions. It should be stated that Strawson's doctrine is somewhat ambiguous and that I have chosen a preferred interpretation, which I think Strawson also prefers today.

tion." (I am not attempting to be philosophically completely precise here.)

To carry out these ideas, we need a semantical scheme to handle predicates that may be only partially defined. Given a nonempty domain $D$, a monadic predicate $P(x)$ is interpreted by a pair $(S_1,S_2)$ of disjoint subsets of $D$. $S_1$ is the *extension* of $P(x)$ and $S_2$ is its *anti-extension*. $P(x)$ is to be true of the objects in $S_1$, false of those in $S_2$, undefined otherwise. The generalization to $n$-place predicates is obvious.

One appropriate scheme for handing connectives is Kleene's strong three-valued logic. Let us suppose that $\sim P$ is true (false) if $P$ is false (true), and undefined if $P$ is undefined. A disjunction is true if at least one disjunct is true regardless of whether the other disjunct is true, false, or undefined[17]; it is false if both disjuncts are false; undefined, otherwise. The other truth functions can be defined in terms of disjunction and negation in the usual way. (In particular, then, a conjunction will be true if both conjuncts are true, false if at least one conjunct is false, and undefined otherwise.) $(\exists x)A(x)$ is true if $A(x)$ is true for some assignment of an element of $D$ to $x$; false if $A(x)$ is false for all assignments to $x$, and undefined otherwise. $(x)A(x)$ can be defined as $\sim(\exists x) \sim A(x)$. It therefore is true if $A(x)$ is true for all assignments to $x$, false if $A(x)$ is false for at least one such assignment, and undefined otherwise. We could convert the preceding into a more precise formal definition of satisfaction, but we won't bother.[18]

[17] Thus the disjunction of 'snow is white' with a Liar sentence will be true. If we had regarded a Liar sentence as *meaningless*, presumably we would have had to regard any compound containing it as meaningless also. Since we don't regard such a sentence as meaningless, we can adopt the approach taken in the text.

[18] The valuation rules are those of S. C. Kleene, *Introduction to Metamathematics* (New York: Van Nostrand, 1952), sec. 64, pp. 332–340. Kleene's notion of regular tables is equivalent (for the class of valuations he considers) to our requirement of the monotonicity of $\phi$ below.

I have been amazed to hear my use of the Kleene valuation compared occasionally to the proposals of those who favor abandoning standard logic "for quantum mechanics," or positing extra truth values beyond truth and falsity, etc. Such a reaction surprised me as much as it would presumably surprise Kleene, who intended (as I do here) to write a work of standard mathematical results, provable in conventional mathematics. "Undefined" is not an *extra* truth value, any more than—in Kleene's book—u is an extra *number* in sec. 63. Nor should it be said that "classical logic" does not generally hold, any more than (in Kleene) the use of partially defined functions invalidates the commutative law of addition. *If* certain sentences express propostitions, any tautological truth function of them expresses a true proposition. Of course formulas, even with the forms of tautologies, which have components that do not express propositions may have truth functions that do not express propositions either. (This happens under the Kleene valuation, but not under the van Fraassen.) Mere conventions for handling terms that do not designate numbers should not be called changes in arithmetic; conventions for

We wish to capture an intuition of somewhat the following kind. Suppose we are explaining the word 'true' to someone who does not yet understand it. We may say that we are entitled to assert (or deny) of any sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself. Our interlocutor then can understand what it means, say, to attribute truth to (6) ('snow is white') but he will still be puzzled about attributions of truth to sentences containing the word 'true' itself. Since he did not understand these sentences initially, it will be equally nonexplanatory, initially, to explain to him that to call such a sentence "true" ("false") is tantamount to asserting (denying) the sentence itself.

Nevertheless, with more thought the notion of truth as applied even to various sentences themselves containing the word 'true' can gradually become clear. Suppose we consider the sentence,

(7) Some sentence printed in the *New York Daily News*, October 7, 1971, is true.

(7) is a typical example of a sentence involving the concept of truth itself. So if (7) is unclear, so still is

(8) (7) is true.

However, our subject, if he is willing to assert 'snow is white', will according to the rules be willing to assert '(6) is true'. But suppose that among the assertions printed in the *New York Daily News*, October 7, 1971, is (6) itself. Since our subject is willing to assert '(6) is true', and also to assert '(6) is printed in the *New York Daily News*, October 7, 1971', he will deduce (7) by existential generalization. Once he is willing to assert (7), he will also be willing to assert (8). In this manner, the subject will eventually be able to attribute truth to more and more statements involving the notion of truth itself. There is no reason to suppose that *all* statements involving 'true' will become decided in this way, but most will. Indeed, our suggestion is that the "grounded" sentences can be characterized as those which eventually get a truth value in this process.

A typically ungrounded sentence such as (3) will, of course, receive no truth value in the process just sketched. In particular, it will never be called "true." But the subject cannot express this fact by saying, "(3) is not true." Such an assertion would conflict directly with the stipulation that he should deny that a sentence is true

---

handling sentences that do not express propositions are not in any philosophically significant sense "changes in logic." The term 'three-valued logic', occasionally used here, should not mislead. All our considerations can be formalized in a classical metalanguage.

precisely under the circumstances under which he would deny the sentence itself. In imposing this stipulation, we have made a deliberate choice (see below).

Let us see how we can give these ideas formal expression. Let $L$ be an interpreted first-order language of the classical type, with a finite (or even denumerable) list of primitive predicates. It is assumed that the variables range over some nonempty domain $D$, and that the primitive $n$-ary predicates are interpreted by (totally defined) $n$-ary relations on $D$. The interpretation of the predicates of $L$ is kept fixed throughout the following discussion. Let us also assume that the language $L$ is rich enough so that the syntax of $L$ (say, via arithmetization) can be expressed in $L$, and that some coding scheme codes finite sequences of elements of $D$ into elements of $D$. We do not attempt to make these ideas rigorous; Y. N. Moschovakis's notion of an "acceptable" structure would do so.[19] I should emphasize that a great deal of what we do below goes through under much weaker hypotheses on $L$.[20]

Suppose we extend $L$ to a language $\mathfrak{L}$ by adding a monadic predicate $T(x)$ whose interpretation need only be partially defined. An interpretation of $T(x)$ is given by a "partial set" $(S_1,S_2)$, where $S_1$, as we said above, is the *extension* of $T(x)$, $S_2$ is the *antiextension* of $T(x)$, and $T(x)$ is undefined for entities outside $S_1 \cup S_2$. Let $\mathfrak{L}(S_1,S_2)$ be the interpretation of $\mathfrak{L}$ which results from interpreting $T(x)$ by the pair $(S_1,S_2)$, the interpretation of the other predicates of $L$ remaining as before.[21] Let $S_1'$ be the set of (codes of)[22] true sentences of $\mathfrak{L}(S_1,S_2)$, and let $S_2'$ be the set of all elements of $D$ which either are not (codes of) sentences of $\mathfrak{L}(S_1,S_2)$ or are (codes of) false sentences of $\mathfrak{L}(S_1,S_2)$. $S_1'$ and $S_2'$ are uniquely determined by the choice of $(S_1,S_2)$. Clearly, if $T(x)$ is to be interpreted as truth for the very language $L$ containing $T(x)$ itself, we must have $S_1 = S_1'$ and $S_2 = S_2'$. [This means that if $A$ is any sentence, $A$ satisfies (falsifies) $T(x)$ iff $A$ is true (false) by the evaluation rules.]

[19] *Elementary Induction on Abstract Structures* (Amsterdam: North-Holland, 1974). The notion of an acceptable structure is developed in chap. 5.

[20] It is unnecessary to suppose, as we have for simplicity, that all the predicates in $L$ are totally defined. The hypothesis that $L$ contain a device for coding finite sequences is needed only if we are adding satisfaction rather than truth to $L$. Other hypotheses can be made much weaker for most of the work.

[21] $\mathfrak{L}$ is thus a language with all predicates but the single predicate $T(x)$ interpreted, but $T(x)$ is uninterpreted. The languages $\mathfrak{L}(S_1,S_2)$ and the languages $\mathfrak{L}_\alpha$ defined below are languages obtained from $\mathfrak{L}$ by specifying an interpretation of $T(x)$.

[22] I parenthetically write "codes of" or "Gödel numbers of" in various places to remind the reader that syntax may be represented in $L$ by Gödel numbering or some other coding device. Sometimes I lazily drop the parenthetical qualification, identifying expressions with their codes.

A pair $(S_1,S_2)$ that satisfies this condition is called a *fixed point.* For a given choice of $(S_1,S_2)$ to interpret $T(x)$, set $\phi((S_1,S_2)) = (S_1',S_2')$. $\phi$ then is a unary function defined on all pairs $(S_1,S_2)$ of disjoint subsets of $D$, and the "fixed points" $(S_1,S_2)$ are literally the fixed points of $\phi$; i.e., they are those pairs $(S_1,S_2)$ such that $\phi((S_1,S_2)) = (S_1,S_2)$. If $(S_1,S_2)$ is a fixed point, we sometimes call $\mathfrak{L}(S_1,S_2)$ a fixed point also. Our basic task is to prove the existence of fixed points, and to investigate their properties.

Let us first construct a fixed point. We do so by considering a certain "hierarchy of languages." We start by defining the interpreted language $\mathfrak{L}_0$ as $\mathfrak{L}(\Lambda,\Lambda)$, where $\Lambda$ is the empty set; i.e., $\mathfrak{L}_0$ is the language where $T(x)$ is completely undefined. (It is never a fixed point.) For any integer $\alpha$, suppose we have defined $\mathfrak{L}_\alpha = \mathfrak{L}(S_1,S_2)$. Then set $\mathfrak{L}_{\alpha+1} = \mathfrak{L}(S_1',S_2')$, where as before $S_1'$ is the set of (codes of) true sentences of $\mathfrak{L}_\alpha$, and $S_2'$ is the set of all elements of $D$ which either are not (codes of) sentences of $\mathfrak{L}_\alpha$ or are (codes of) false sentences of $\mathfrak{L}_\alpha$.

The hierarchy of languages just given is analogous to the Tarski hierarchy for the orthodox approach. $T(x)$ is interpreted in $\mathfrak{L}_{\alpha+1}$ as the truth predicate for $\mathfrak{L}_\alpha$. But an interesting phenomenon, detailed in the following paragraphs, arises on the present approach.

Let us say that $(S_1^\dagger,S_2^\dagger)$ *extends* $(S_1,S_2)$ [symbolically, $(S_1^\dagger,S_2^\dagger) \geq (S_1,S_2)$ or $(S_1,S_2) \leq (S_1^\dagger,S_2^\dagger)$] iff $S_1 \subseteq S_1^\dagger$, $S_2 \subseteq S_2^\dagger$. Intuitively this means that if $T(x)$ is interpreted as $(S_1^\dagger,S_2^\dagger)$, the interpretation agrees with the interpretation by $(S_1,S_2)$ in all cases where the latter is defined; the only difference is that an interpretation by $(S_1^\dagger,S_2^\dagger)$ may lead $T(x)$ to be defined for some cases where it was undefined when interpreted by $(S_1,S_2)$. Now a basic property of our valuation rules is the following: $\phi$ is a monotone (order-preserving) operation on $\leq$: that is, if $(S_1,S_2) \leq (S_1^\dagger,S_2^\dagger)$, $\phi((S_1,S_2)) \leq \phi((S_1^\dagger,S_2^\dagger))$. In other words, *if* $(S_1,S_2) \leq (S_1^\dagger,S_2^\dagger)$, *then any sentence that is true (or false) in* $\mathfrak{L}(S_1,S_2)$ *retains its truth value in* $\mathfrak{L}(S_1^\dagger,S_2^\dagger)$. What this means is that *if the interpretation of* $T(x)$ *is extended by giving it a definite truth value for cases that were previously undefined, no truth value previously established changes or becomes undefined;* at most, certain previously undefined truth values become defined. This property—technically, the monotonicity of $\phi$—is crucial for all our constructions.

Given the monotonicity of $\phi$, we can deduce that for each $\alpha$, *the interpretation of* $T(x)$ in $\mathfrak{L}_{\alpha+1}$ *extends the interpretation* of $T(x)$ *in* $\mathfrak{L}_\alpha$. The fact is obvious for $\alpha = 0$: since, in $\mathfrak{L}_0$, $T(x)$ is undefined for all $x$, any interpretation of $T(x)$ automatically extends it. If the asser-

tion holds for $\mathfrak{L}_\beta$—that is, if the interpretation of $T(x)$ in $\mathfrak{L}_{\beta+1}$ extends that of $T(x)$ in $\mathfrak{L}_\beta$—then any sentence true or false in $\mathfrak{L}_\beta$ remains true or false in $\mathfrak{L}_{\beta+1}$. If we look at the definitions, *this says that the interpretation of* $T(x)$ *in* $\mathfrak{L}_{\beta+2}$ *extends the interpretation of* $T(x)$ *in* $\mathfrak{L}_{\beta+1}$. *We have thus proved by induction that the interpretation of* $T(x)$ *in* $\mathfrak{L}_{\alpha+1}$ *always extends the interpretation of* $T(x)$ *in* $\mathfrak{L}_\alpha$ *for all finite* $\alpha$. *It follows that the predicate* $T(x)$ *increases, in both its extension and its antiextension, as* $\alpha$ *increases. More and more sentences get declared true or false as* $\alpha$ *increases; but once a sentence is declared true or false, it retains its truth value at all higher levels.*

So far, we have defined only *finite* levels of our hierarchy. For finite $\alpha$, let $(S_{1,\alpha},S_{2,\alpha})$ be the interpretation of $T(x)$ in $\mathfrak{L}_\alpha$. Both $S_{1,\alpha}$ and $S_{2,\alpha}$ increase (as sets) as $\alpha$ increases. Then there is an obvious way of defining the first "transfinite" level—call it "$\mathfrak{L}_\omega$." Simply define $\mathfrak{L}_\omega = \mathfrak{L}(S_{1,\omega},S_{2,\omega})$, where $S_{1,\omega}$ is the union of all $S_{1,\alpha}$ for finite $\alpha$, and $S_{2,\omega}$ is similarly the union of $S_{2,\alpha}$ for finite $\alpha$. Given $\mathfrak{L}_\omega$, we can then define $\mathfrak{L}_{\omega+1}$, $\mathfrak{L}_{\omega+2}$, $\mathfrak{L}_{\omega+2}$, etc., just as we did for the finite levels. When we get again to a "limit" level, we take a union as before.

Formally, we define the languages $\mathfrak{L}_\alpha$ for each ordinal $\alpha$. If $\alpha$ is a successor ordinal ($\alpha = \beta + 1$), let $\mathfrak{L}_\alpha = \mathfrak{L}(S_{1,\alpha},S_{2,\alpha})$, where $S_{1,\alpha}$ is the set of (codes of) true sentences of $\mathfrak{L}_\beta$, and $S_{2,\alpha}$ is the set consisting of all elements of $D$ which either are (codes of) false sentences of $\mathfrak{L}_\beta$ or are not (codes of) sentences of $\mathfrak{L}_\beta$. If $\lambda$ is a limit ordinal, $\mathfrak{L}_\lambda = \mathfrak{L}(S_{1,\lambda},S_{2,\lambda})$, where $S_{1,\lambda} = \bigcup_{\beta<\lambda}S_{1,\beta}$, $S_{2,\lambda} = \bigcup_{\beta<\lambda}S_{2,\beta}$. So at "successor" levels we take the truth predicate over the previous level, and, at limit (transfinite) levels, we take the union of all sentences declared true or false at previous levels. *Even with the transfinite levels included, it remains true that the extension and the antiextension of* $T(x)$ *increase with increasing* $\alpha$.

It should be noted that 'increase' does not mean "strictly increase"; we have asserted that $S_{i,\alpha} \subseteq S_{i,\alpha+1}$ $(i = 1,2)$, which allows equality. Does the process go on forever with more and more statements being declared true or false, or does it eventually stop? That is to say, is there an ordinal level $\sigma$ for which $S_{1,\sigma} = S_{1,\sigma+1}$ and $S_{2,\sigma} = S_{2,\sigma+1}$, so that no "new" statements are declared true or false at the next level? The answer must be affirmative. The sentences of $\mathfrak{L}$ form a set. If new sentences of $\mathfrak{L}$ were being decided at each level, we would eventually exhaust $\mathfrak{L}$ at some level and be unable to decide any more. This can easily be converted to a formal proof (the technique is elementary and is well known to logicians) that there is an ordinal level $\sigma$ such that $(S_{1,\sigma},S_{2,\sigma}) = (S_{1,\sigma+1},S_{2,\sigma+1})$. But since $(S_{1,\sigma+1},S_{2,\sigma+1}) = \phi((S_{1,\sigma},S_{2,\sigma}))$, *this means that* $(S_{1,\sigma},S_{2,\sigma})$ *is a fixed*

*point.* It can also be proved that it is a "minimal" or "smallest" fixed point: *any* fixed point extends $(S_{1,\sigma},S_{2,\sigma})$. That is, if a sentence is valuated as true or false in $\mathfrak{L}_\sigma$, it has the same truth value in *any* fixed point.

Let us relate the construction of a fixed point just given to our previous intuitive ideas. At the initial stage ($\mathfrak{L}_0$), $T(x)$ is completely undefined. This corresponds to the initial stage at which the subject has no understanding of the notion of truth. Given a characterization of truth by the Kleene valuation rules, the subject can easily ascend to the level of $\mathfrak{L}_1$. That is, he can evaluate various statements as true or false without knowing anything about $T(x)$—in particular, he can evaluate all those sentences not containing $T(x)$. Once he has made the evaluation, he extends $T(x)$, as in $\mathfrak{L}_1$. Then he can use the new interpretation of $T(x)$ to evaluate more sentences as true or false and ascend to $\mathfrak{L}_2$, etc. Eventually, when the process becomes "saturated," the subject reaches the fixed point $\mathfrak{L}_\sigma$. (*Being a fixed point, $\mathfrak{L}_\sigma$ is a language that contains its own truth predicate.*) So the formal definition just given directly parallels the intuitive constructions stated previously.[23]

We have been talking of a language that contains its own truth predicate. Really, however, it would be more interesting to extend an arbitrary language to a language containing its own *satisfaction* predicate. If $L$ contains a name for each object in $D$, and a denotation relation is defined (if $D$ is nondenumerable, this means that $L$ contains nondenumerably many constants), the notion of satisfaction can (for most purposes) effectively be replaced by that of truth: e.g., instead of talking of $A(x)$ being satisfied by an object $a$, we can talk of $A(x)$ becoming true when the variable is replaced by a name of $a$. Then the previous construction suffices. Alternatively, if $L$ does not contain a name for each object, we can extend $L$ to $\mathfrak{L}$ by adding a binary satisfaction predicate $Sat(s,x)$ where $s$ ranges over finite sequences of elements of $D$ and $x$ ranges over formulas. We define a hierarchy of languages, parallel to the previous construction with truth, eventually reaching a fixed point—a language that contains its own satisfaction predicate. If $L$ is denumerable but $D$ is not, the

[23] A comparison with the Tarski hierarchy:

The Tarski hierarchy uses a new truth predicate at each level, always changing. The limit levels of the Tarski hierarchy, which have not been defined in the literature, but have been to some extent in my own work, are cumbersome to characterize.

The present hierarchy uses a single truth predicate, ever increasing with increasing levels until the level of the minimal fixed point is reached. The limit levels are easily defined. The languages in the hierarchy are not the primary object of interest, but are better and better approximations to the minimal language with its own truth predicate.

construction with truth alone closes off at a countable ordinal, but the construction with satisfaction may close off at an uncountable ordinal. Below we will continue, for simplicity of exposition, to concentrate on the construction with truth, but the construction with satisfaction is more basic.[24]

The construction could be generalized so as to allow more notation in $L$ than just first-order logic. For example, we could have a quantifier meaning "for uncountably many $x$," a "most" quantifier, a language with infinite conjunctions, etc. There is a fairly canonical way, in the Kleene style, to extend the semantics of such quantifiers and connectives so as to allow truth-value gaps, but we will not give details.

Let us check that our model satisfies some of the desiderata mentioned in the previous sections. It is clearly a theory in the required sense: any language, including those containing number theory or syntax, can be extended to a language with its own truth predicate, and the associated concept of truth is *mathematically* defined by set-theoretic techniques. There is no problem about the languages of transfinite level in the hierarchy.

Given a sentence $A$ of $\mathfrak{L}$, let us define $A$ to be *grounded* if it has a truth value in the smallest fixed point $\mathfrak{L}_\sigma$; otherwise, *ungrounded*. What hitherto has been, as far as I know, an intuitive concept with no formal definition, becomes a precisely defined concept in the present theory. If $A$ is grounded, define the *level* of $A$ to be the smallest ordinal $\alpha$ such that $A$ has a truth value in $\mathfrak{L}_\alpha$.

There is no problem, if $\mathfrak{L}$ contains number theory or syntax, of constructing Gödelian sentences that "say of themselves" that they are false (Liar sentences) or true [as in (3)]; all these are easily shown to be ungrounded in the sense of the formal definition. If the Gödelian form of the Liar paradox is used, for example, the Liar

[24] Consider the case where $L$ has a canonical name for every element of $D$. We can then consider pairs $(A,\mathbf{T})$, $(A,\mathbf{F})$, where $A$ is true, or false, respectively. The Kleene rules correspond to closure conditions on a set of such pairs: e.g., if $(A(a),\mathbf{F}) \in S$ for each name of $a$ element of $D$, put $((\exists x)A(x),\mathbf{F}))$ in $S$; if $((A(a),\mathbf{T}) \in S$, put $((\exists x)A(x),\mathbf{T})$ in $S$, etc. Consider the least set $S$ of pairs closed under the analogues of the Kleene rules, containing $(A,\mathbf{T})((A,\mathbf{F}))$ for each true (false) atomic $A$ of $L$, and closed under the two conditions: (i) if $(A,\mathbf{T}) \in S$, $(T(\mathbf{k}),\mathbf{T}) \in S$; (ii) if $(A,\mathbf{F}) \in S$, $(T(\mathbf{k}),\mathbf{F}) \in S$, where '$\mathbf{k}$' abbreviates a name of $A$. It is easily shown that the set $S$ corresponds (in the obvious sense) to the minimal fixed point [thus, it is closed under the converses of (i) and (ii).] I used this definition to show that the set of truths in the minimal fixed point (over an acceptable structure), is inductive in Moschovakis's sense. It is probably simpler than the definition given in the text. The definition given in the text has, among others, the advantages of giving a definition of 'level', facilitating a comparison with the Tarski hierarchy, and easy generalization to valuation schemes other than Kleene's.

sentence can get the form

(9) $$(x)(P(x) \supset \sim T(x))$$

where $P(x)$ is a syntactic (or arithmetical) predicate uniquely satisfied by (the Gödel number of) (9) itself. Similarly (3) gets the form

(10) $$(x)(Q(x) \supset T(x))$$

where $Q(x)$ is uniquely satisfied by (the Gödel number of) (10). It is easy to prove, under these hypotheses, by induction on $\alpha$, that neither (9) nor (10) will have a truth value in any $\mathfrak{L}_\alpha$, that is, that they are ungrounded. Other intuitive cases of ungroundedness come out similarly.

The feature I have stressed about ordinary statements, that there is no intrinsic guarantee of their safety (groundedness) and that their "level" depends on empirical facts, comes out clearly in the present model. Consider, for example, (9) again, except that now $P(x)$ is an empirical predicate whose extension depends on unknown empirical facts. If $P(x)$ turns out to be true only of (9) itself, (9) will be ungrounded as before. If the extension of $P(x)$ consists entirely of grounded sentences of levels, say, 2, 4, and 13, (9) will be grounded with level 14. If the extension of $P(x)$ consists of grounded sentences of arbitrary finite level, (9) will be grounded with level $\omega$. And so on.

Now let us consider the cases of (4) and (5). We can formalize (4) by (9), interpreting $P(x)$ as "$x$ is a sentence Nixon asserts about Watergate." [Forget for simplicity that 'about Watergate' introduces a semantic component into the interpretation of $P(x)$.] Formalize (5) as

(11) $$(x)(Q(x) \supset \sim T(x))$$

interpreting $Q(x)$ in the obvious way. To complete the parallel with (4) and (5), suppose that (9) is in the extension of $Q(x)$ and (11) is in the extension of $P(x)$. Now nothing guarantees that (9) and (11) will be grounded. Suppose, however, parallel to the intuitive discussion above, that some true grounded sentence satisfies $Q(x)$. If the lowest level of any such sentence is $\alpha$, then (11) will be false and grounded of level $\alpha + 1$. If in addition all the sentences other than (11) satisfying $P(x)$ are false, (9) will then be grounded and true. The level of (9) will be at least $\alpha + 2$, because of the level of (11). On the other hand, if some sentence satisfying $P(x)$ is grounded and true, then (9) will be grounded and false with level $\beta + 1$, where $\beta$ is the lowest level of any such sentence. It is crucial to the ability of the present model to assign levels to (4) and (5) [(9) and (11)] that the levels depend on empirical facts, rather than being assigned in advance.

We said that such statements as (3), though ungrounded, are not intuitively paradoxical either. Let us explore this in terms of the model. The smallest fixed point $\mathfrak{L}_\sigma$ is not the only fixed point. Let us formalize (3) by (10), where $Q(x)$ is a *syntactic* predicate (of $L$) true of (10) itself alone. Suppose that, instead of starting out our hierarchy of languages with $T(x)$ completely undefined, we had started out by letting $T(x)$ be true of (10), undefined otherwise. We then can continue the hierarchy of languages just as before. It is easy to see that if (10) is true at the language of a given level, it will remain true at the next level [using the fact that $Q(x)$ is true of (10) alone, false of everything else]. From this we can show as before that the interpretation of $T(x)$ at each level extends all previous levels, and that at some level the construction closes off to yield a fixed point. The difference is that (10), which lacked truth value in the smallest fixed point, is now *true*.

This suggests the following definition: a sentence is *paradoxical* if it has no truth value in *any* fixed point. That is, a paradoxical sentence $A$ is such that if $\phi((S_1,S_2)) = (S_1,S_2)$, then $A$ is neither an element of $S_1$ nor an element of $S_2$.

(3) [or its formal version (10)] is ungrounded, but not paradoxical. This means that we *could* consistently use the predicate 'true' so as to give (3) [or (10)] a truth value, though the minimal process for assigning truth values does not do so. Suppose, on the other hand, in (9), that $P(x)$ is true of (9) itself and false of everything else, so that (9) is a Liar sentence. Then the argument of the Liar paradox easily yields a proof that (9) cannot have a truth value in any fixed point. So (9) is paradoxical in our technical sense. Notice that, if it is merely an empirical fact that $P(x)$ is true of (9) and false of everything else, the fact that (9) is paradoxical will itself be empirical. (We could define notions of "intrinsically paradoxical", "intrinsically grounded", etc., but will not do so here.)

Intuitively, the situation seems to be as follows. Although the smallest fixed point is probably the most natural model for the intuitive concept of truth, and is the model *generated* by our instructions to the imaginary subject, the other fixed points never *conflict* with these instructions. We *could* consistently use the word 'true' so as to give a truth value to such a sentence as (3) without violating the idea that a sentence should be asserted to be true precisely when we would assert the sentence itself. The same does not hold for the paradoxical sentences.

Using Zorn's Lemma, we can prove that *every fixed point can be extended to a maximal fixed point*, where a maximal fixed point is a fixed point that has no proper extension that is also a fixed point. Maximal fixed points assign "as many truth values as possible"; one could not assign more consistently with the intuitive concept of

truth. Sentences like (3), though ungrounded, have a truth value in every maximal fixed point. Ungrounded sentences exist, however, which have truth values in some but not all maximal fixed points.

It is as easy to construct fixed points which make (3) false as it is to construct fixed points which make it true. So the assignment of a truth value to (3) is *arbitrary*. Indeed any fixed point which assigns no truth value to (3) can be extended to fixed points which make it true and to fixed points which make it false. Grounded sentences have the same truth value in all fixed points. There are ungrounded and unparadoxical sentences, however, which have the same truth value in all the fixed points where they have a truth value. An example is:

(12) Either (12) or its negation is true.

It is easy to show that there are fixed points which make (12) true and none which make (12) false. Yet (12) is ungrounded (has no truth value in the minimal fixed point).

Call a fixed point *intrinsic* iff it assigns no sentence a truth value conflicting with its truth value in any other fixed point. That is, a fixed point $(S_1,S_2)$ is intrinsic iff there is no other fixed point $(S_1^\dagger,S_2^\dagger)$ and sentence $A$ of $L'$ such that $A \in (S_1 \cap S_2^\dagger) \cup (S_2 \cap S_1^\dagger)$. We say that a sentence has *an intrinsic truth value* iff some intrinsic fixed point gives it a truth value; i.e., $A$ has an intrinsic truth value iff there is an intrinsic fixed point $(S_1,S_2)$ such that $A \in S_1 \cup S_2$. (12) is a good example.

There are unparadoxical sentences which have the same truth value in all fixed points where they have truth value but which nevertheless lack an intrinsic truth value. Consider $P \vee \sim P$, where $P$ is any ungrounded unparadoxical sentence. Then $P \vee \sim P$ is true in some fixed points (namely, those where $P$ has a truth value) and is false in none. Suppose, however, that there are fixed points that make $P$ true and fixed points that make $P$ false. [For example, say, $P$ is (3).] Then $P \vee \sim P$ cannot have a truth value in any intrinsic fixed point, since, by our valuation rules, it cannot have a truth value unless some disjunct does.[25]

There is no "largest" fixed point that extends every other; indeed, any two fixed points that give different truth values to the same formula have no common extension. However, it is not hard to show that there is a largest intrinsic fixed point (and indeed that the intrinsic fixed points form a complete lattice under $\leq$). The largest intrinsic fixed point is the unique "largest" interpretation of $T(x)$ which is consistent with our intuitive idea of truth and makes no

[25] If we use the supervaluation technique instead of the Kleene rules, $P \vee \sim P$ will always be grounded and true, and we must change the example. See p. 711 below.

arbitrary choices in truth assignments. It is thus an object of special theoretical interest as a model.

It is interesting to compare "Tarski's hierarchy of languages" with the present model. Unfortunately, this can hardly be done in full generality without introducing the transfinite levels, a task omitted from this sketch. But we can say something about the finite levels. Intuitively, it would seem that Tarski predicates $\ulcorner \text{true}_n \urcorner$ are all special cases of a single truth predicate. For example, we said above that '$\text{true}_1$' means "is a true sentence not involving truth." Let us carry this idea out formally. Let $A_1(x)$ be a syntactic (arithmetical) predicate true of exactly the formulas of $\mathfrak{L}$ not involving $T(x)$, i.e., of all formulas of $L$. $A_1(x)$, being syntactic, is itself a formula of $L$, as are all other syntactic formulas below. Define '$T_1(x)$' as '$T(x) \wedge A_1(x)$'. Let $A_2(x)$ be a syntactic predicate applying to all those formulas whose atomic predicates are those of $L$ plus '$T_1(x)$'. [More precisely the class of such formulas can be defined as the least class including all formulas of $L$ and $T(x_i) \wedge A_1(x_i)$, for any variable $x_i$, and closed under truth functions and quantification.] Then define $T_2(x)$ as $T(x) \wedge A_2(x)$. In general, we can define $A_{n+1}(x)$ as a syntactic predicate applying precisely to formulas built out of the predicates of $L$ and $T_n(x)$, and $T_{n+1}(x)$ as $T(x) \wedge A_{n+1}(x)$. Assume that $T(x)$ is interpreted by the smallest fixed point (or any other). Then it is easy to prove by induction that each predicate $T_n(x)$ is totally defined, that the extension of $T_0(x)$ consists precisely of the true formulas of $L$, while that of $T_{n+1}(x)$ consists of the true formulas of the language obtained by adjoining $T_n(x)$ to $L$. This means that *all the truth predicates of the finite Tarski hierarchy are definable within* $\mathfrak{L}_\sigma$, *and all the languages of that hierarchy are sublanguages of* $\mathfrak{L}_\sigma$.[26] This kind of result could be extended into the transfinite if we had defined the transfinite Tarski hierarchy.

There are converse results, harder to state in this sketch. It is characteristic of the sentences in the Tarski hierarchy that they are safe (intrinsically grounded) and that their level is intrinsic, given independently of the empirical facts. It is natural to conjecture that any grounded sentence with intrinsic level $n$ is in some sense "equivalent" to a sentence of level $n$ in the Tarski hierarchy. Given proper definitions of 'intrinsic level', 'equivalent', and the like, theorems of this kind can be stated and proved and even extended into the transfinite.

[26] We suppose that the Tarski hierarchy defines $L_0 = L$, $L_{n+1} = L + T_{n+1}(x)$ (truth, or satisfaction, for $L_n$). Alternatively, we might prefer the inductive construction $L_0 = L$, $L_{n+1} = L_n + T_{n+1}(x)$ where the language of each new level contains all the previous truth predicates. It is easy to modify the construction in the text so as to accord with the second definition. The two alternative hierarchies are equivalent in expressive power at each level.

So far we have assumed that truth gaps are to handled according to the methods of Kleene. It is by no means necessary to do so. Just about any scheme for handling truth-value gaps is usable, provided that the basic property of the monotonicity of $\phi$ is preserved; that is, provided that extending the interpretation of $T(x)$ never changes the truth value of any sentence of $\mathfrak{L}$, but at most gives truth values to previously undefined cases. Given any such scheme, we can use the previous arguments to construct the minimal fixed point and other fixed points, define the levels of sentences and the notions of 'grounded', 'paradoxical', etc.

One scheme usable in this way is van Fraassen's notion of *supervaluation*.[27] For the language $\mathfrak{L}$, the definition is easy. Given an interpretation $(S_1,S_2)$ of $T(x)$ in $\mathfrak{L}$, call a formula $A$ true (false) iff it comes out true (false) by the ordinary classical valuation under every interpretation $(S_1^{\dagger},S_2^{\dagger})$ which extends $(S_1,S_2)$ and is *totally defined*, i.e., is such that $S_1^{\dagger} \cup S_2^{\dagger} = D$. We can then define the hierarchy $\{\mathfrak{L}_\alpha\}$ and the minimal fixed point $\mathfrak{L}_\sigma$ as before. Under the supervaluation interpretation, all formulas provable in classical quantification theory become true in $\mathfrak{L}_\sigma$; under the Kleene valuation, one could say only that they were true whenever they were defined. Thanks to the fact that $\mathfrak{L}_\sigma$ contains its own truth predicate, we need not express this fact by a schema, or by a statement of a metalanguage. If $PQT(x)$ is a syntactic predicate true exactly of the sentences of $\mathfrak{L}$ provable in quantification theory, we can assert:

$$(13) \qquad (x)(PQT(x) \supset T(x))$$

and (13) will be true in the minimal fixed point.

Here we have used supervaluations in which *all* total extensions of the interpretation of $T(x)$ are taken into account. It is natural to consider restrictions on the family of total extensions, motivated by intuitive properties of truth. For example, we could consider only *consistent* interpretations $(S_1^{\dagger},S_2^{\dagger})$, where $(S_1^{\dagger},S_2^{\dagger})$ is consistent iff $S_1$ contains no sentence together with its negation. Then we could define $A$ to be true (false) with $T(x)$ interpreted by $(S_1,S_2)$ iff $A$ is true (false) classically when $A$ is interpreted by any *consistent* totally defined extension of $(S_1,S_2)$.

$$(14) \qquad (x) \sim (T(x) \wedge T(\mathrm{neg}(x))$$

will be true in the minimal fixed point. If we restricted the admissible total extensions to those defining *maximal* consistent sets of sentences, in the usual sense, not only (14) but even

$$(x)(Sent(x) \supset . T(x) \vee T(\mathrm{neg}(x))$$

[27] See his "Singular Terms, Truth-value Gaps, and Free Logic," this JOURNAL, LXIII, 17 (Sept. 15, 1966): 481–495.

will come out true in the minimal fixed point.[28] The last-mentioned formula, however, must be interpreted with caution, since it is still not the case, even on the supervaluation interpretation in question, that there is any fixed point that makes every formula or its negation true. (The paradoxical formulas still lack truth value in all fixed points.) The phenomenon is associated with the fact that, on the supervaluation interpretation, a disjunction can be true without it following that some disjunct is true.

It is not the purpose of the present work to make any particular recommendation among the Kleene strong three-valued approach, the van Fraassen supervaluation approaches, or any other scheme (such as the Fregean weak three-valued logic, preferred by Martin and Woodruff, though I am in fact tentatively inclined to consider the latter excessively cumbersome). Nor is it even my present purpose to make any firm recommendation between the minimal fixed point of a particular valuation scheme and the various other fixed points.[29] Indeed, without the nonminimal fixed points we could not have defined the intuitive difference between 'grounded' and 'paradoxical'. My purpose is rather to provide a family of flexible instruments which can be explored simultaneously and whose fertility and consonance with intuition can be checked.

I am somewhat uncertain whether there is a definite factual question as to whether natural language handles truth-value gaps—at least those arising in connection with the semantic paradoxes—by the schemes of Frege, Kleene, van Fraassen, or perhaps some other. Nor am I even *quite* sure that there is a definite question of fact as to whether natural language should be evaluated by the minimal fixed point or another, given the choice of a scheme for handling gaps.[30] We are not at the moment searching for *the* correct scheme.

The present approach can be applied to languages containing modal operators. In this case, we do not merely consider truth, but we are given, in the usual style of modal model theory, a system of possible worlds, and evaluate truth and $T(x)$ in each possible world. The inductive definition of the languages $\mathfrak{L}_\alpha$ approximating to the

[28] A version of the Liar paradox due to H. Friedman shows that there are limits to what can be done in this direction.

[29] Though the minimal fixed point certainly is singled out as natural in many respects.

[30] I do not mean to *assert* that there are no definite questions of fact in these areas, or even that I myself may not favor some valuation schemes over others. But my personal views are less important than the variety of tools that are available, so for the purposes of this sketch I take an agnostic position. (I remark that if the viewpoint is taken that logic applies primarily to propositions, and that we are merely formulating conventions for how to handle sentences that do not express propositions, the attractiveness of the supervaluation approach over the Kleene approach is somewhat decreased. See fn 18.)

minimal fixed point must be modified accordingly. We cannot give details here.[31]

Ironically, the application of the present approach to languages with modal operators may be of some interest to those who dislike intensional operators and possible worlds and prefer to take modalities and propositional attitudes as predicates true of sentences (or sentence tokens). Montague and Kaplan have pointed out, using elementary applications of Gödelian techniques, that such approaches are likely to lead to semantic paradoxes, analogous to the Liar.[32] Though the difficulty has been known for some time, the extensive literature advocating such treatments has usually simply ignored the problem rather than indicating how it is to be solved (say, by a hierarchy of languages?). Now, if a necessity operator and a truth predicate are allowed, we could define a necessity predicate $Nec(x)$ applied to sentences, either by $\Box T(x)$ or $T(\text{nec}(x))$ according to taste,[33] and treat it according to the possible-world scheme sketched in the preceding paragraph. (I do think that any necessity predicate of sentences should intuitively be regarded as derivative, defined in terms of an operator and a truth predicate. I also think the same holds for propositional attitudes.) We can even "kick away the ladder" and take $Nec(x)$ as primitive, treating it in a possible-world scheme *as if* it were defined by an operator plus a truth predi-

[31] Another application of the present techniques is to "impredicative" substitutional quantification, where the terms of the substitution class themselves contain substitutional quantifiers of the given type. (For example, a language containing substitutional quantifiers with arbitrary sentences of the language itself as substituends ) It is impossible in general to introduce such quantifiers into classical languages without truth-value gaps.

[32] Richard Montague, "Syntactical Treatments of Modality, with Corollaries on Reflection Principles and Finite Axiomatizability," *Acta Philosophica Fennica, Proceedings of a Colloquium on Model and Many Valued Logics*, 1963: 153–167; David Kaplan and Montague, "A Paradox Regained," *Notre Dame Journal of Formal Logic*, I, 3 (July 1960): 79–90.

At present the problems are *known* to arise only if modalities and attitudes are predicates applied to sentences or their tokens. The Montague-Kaplan arguments do not apply to standard formalizations taking modalities or propositional attitudes as intensional operators. Even if we wish to quantify over objects of belief, the arguments do not apply if the objects of belief are taken to be propositions and the latter are identified with sets of possible worlds.

However, if we quantify over propositions, paradoxes may arise in connection with propositional attitudes given appropriate empirical premises. [See, e.g., A. N. Prior, "On a Family of Paradoxes," *Notre Dame Journal of Formal Logic*, II 1 (January 1961): 16–32.] Also, we may wish (in connection with propositional attitudes but not modalities), to individuate propositions more finely than by sets of possible worlds, and it is possible that such a "fine structure" may permit the application of Gödelian arguments of the type used by Montague and Kaplan directly to propositions.

[33] As a formalization of the concept intended by those who speak of modalities and attitudes as predicates of sentences, the second version is generally better. This is true especially for the propositional attitudes.

cate. Like remarks apply to the propositional attitudes, if we are willing to treat them, using possible worlds, like modal operators. (I myself think that such a treatment involves considerable philosophical difficulties.) It is possible that the present approach can be applied to the supposed predicates of sentences in question without using either intensional operators or possible worlds, but at present I have no idea how to do so.

It seems likely that many who have worked on the truth-gap approach to the semantic paradoxes have hoped for a universal language, one in which everything that can be stated at all can be expressed. (The proof by Gödel and Tarski that a language cannot contain its own semantics applied only to languages without truth gaps.) Now the languages of the present approach contain their own truth predicates and even their own satisfaction predicates, and thus to this extent the hope has been realized. Nevertheless the present approach certainly does not claim to give a universal language, and I doubt that such a goal can be achieved. First, the induction defining the minimal fixed point is carried out in a set-theoretic metalanguage, not in the object language itself. Second, there are assertions we can make about the object language which we cannot make in the object language. For example, Liar sentences are *not true* in the object language, in the sense that the inductive process never makes them true; but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate. If we think of the minimal fixed point, say under the Kleene valuation, as giving a model of natural language, then the sense in which we can say, in natural language, that a Liar sentence is not true must be thought of as associated with some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point. It is not itself a part of that process. The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us.[34]

The approach adopted here has presupposed the following version of Tarski's "Convention T", adapted to the three-valued approach: If '**k**' abbreviates a name of the sentence $A$, $T(\mathbf{k})$ is to be true, or

[34] Note that the metalanguage in which we write this paper can be regarded as containing no truth gaps. A sentence either does or does not have a truth value in a given fixed point.

Such semantical notions as "grounded," "paradoxical," etc. belong to the metalanguage. This situation seems to me to be intuitively acceptable; in contrast to the notion of truth, none of these notions is to be found in natural language in its pristine purity, before philosophers reflect on its semantics (in particular, the semantic paradoxes). If we give up the goal of a universal language, models of the type presented in this paper are plausible as models of natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers.

false, respectively iff $A$ is true, or false. This captures the intuition that $T(\mathbf{k})$ is to have the same truth conditions as $A$ itself; it follows that $T(\mathbf{k})$ suffers a truth-value gap if $A$ does. An alternate intuition[35] would assert that, if $A$ is either false or undefined, then $A$ is *not true* and $T(\mathbf{k})$ should be *false*, and its negation *true*. On this view, $T(x)$ will be a totally defined predicate and there are no truth-value gaps. Presumably Tarski's Convention T must be restricted in some way.

It is not difficult to modify the present approach so as to accommodate such an alternate intuition. Take any fixed point $L'(S_1,S_2)$. Modify the interpretation of $T(x)$ so as to make it false of any sentence outside $S$. [We call this "closing off" $T(x)$.] A modified version of Tarski's Convention T holds in the sense of the conditional $T(\mathbf{k}) \vee T(\text{neg}(\mathbf{k})). \supset . A \equiv T(\mathbf{k})$. In particular, if $A$ is a paradoxical sentence, we can now assert $\sim T(\mathbf{k})$. Equivalently, if $A$ had a truth value before $T(x)$ was closed off, then $A \equiv T(\mathbf{k})$ is true.

Since the object language obtained by closing off $T(x)$ is a classical language with every predicate totally defined, it is possible to define a truth predicate for that language in the usual Tarskian manner. This predicate will *not* coincide in extension with the predicate $T(x)$ of the object language, and it is certainly reasonable to suppose that it is really the metalanguage predicate that expresses the "genuine" concept of truth for the closed-off object language; the $T(x)$ of the closed-off language defines truth for the fixed point *before* it was closed off. So we still cannot avoid the need for a metalanguage.

On the basis of the fact that the goal of a universal language seems elusive, some have concluded that truth-gap approaches, or any approaches that attempt to come closer to natural language than does the orthodox approach, are fruitless. I hope that the fertility of the present approach, and its agreement with intuitions about natural language in a large number of instances, cast doubt upon such negative attitudes.

There are mathematical applications and purely technical problems which I have not mentioned in this sketch; they would be beyond the scope of a paper for a philosophical journal. Thus there is the question, which can be answered in considerable generality, of characterizing the ordinal $\sigma$ at which the construction of the minimal fixed point closes off. If $L$ is the language of first-order arithmetic, it turns out that $\sigma$ is $\omega_1$, the first nonrecursive ordinal. A set is the extension of a formula with one free variable in $\mathfrak{L}_\sigma$ iff it is $\Pi^1_1$, and it is

[35] I think the primacy of the first intuition can be defended philosophically, and for this reason I have emphasized the approach based on this intuition. The alternate intuition arises only after we have reflected on the process embodying the first intuition. See above.

the extension of a totally defined formula iff it is hyperarithmetical. The languages $\mathcal{L}_\alpha$ approximating to the minimal fixed point give an interesting "notation-free" version of the hyperarithmetical hierarchy. More generally, if $L$ is the language of an acceptable structure in the sense of Moschovakis, and the Kleene valuation is used, a set is the extension of a monadic formula in the minimal fixed point iff it is inductive in the sense of Moschovakis.[36]

SAUL KRIPKE

Rockefeller University

## HOW TO RUSSELL A FREGE-CHURCH *

THE philosophies of language of Frege and Russell are the two great competing classical theories, and any exact comparison of them requires attention to their intensional logics, which represent the pure theoretical (in the sense of theoretical vs. observational) superstructures—or perhaps one should say deep structures—of their theories. My earlier work on the logic of demonstratives, which argued against what I take to be tenets of Frege's philosophy of language, had led me to a greater appreciation of Russell's views. I wanted to determine what essential features of Frege's doctrine could not be accommodated within a Russellian approach. This attempt led to a surprising result.

### I

I began by noting that, for a variety of puzzles, including Frege's puzzle about the meaning of identity statements and the three puzzles explicitly discussed by Russell in "On Denoting," one can directly compare the solutions of Frege and Russell and assess the theoretical apparatus each brings into play. (When I refer to Russell's logical doctrines, I have in mind the doctrines of "On Denoting" and the first edition of *Principia Mathematica*. Russell held several other doctrines throughout his career, and, of course, the doctrine of *Principia* was not his alone. In attributing doctrines to Frege, I take account not only of his own writings but of those of his great modern exponent and proponent, Alonzo Church.) De-

[36] Leo Harrington informs me that he has proved the conjecture that a set is the extension of a totally defined monadic formula iff it is hyperelementary. The special case of the $\Pi^1_1$ and hyperarithmetical sets if $L$ is number theory is independent of whether the Kleene or the van Fraassen formulation is used. Not so for the general case, where the van Fraassen formulation leads to the $\Pi^1_1$ sets rather than the inductive sets.

* To be presented in a joint APA/ASL symposium on Sets, Concepts, and Extensions, December 29, 1975. Charles Parsons will be co-symposiast; his paper is not available at this time.