THE LIAR AND THEORIES OF TRUTH

John Hawthorn

Philosophy Department

McGill University, Montreal

July 1983

A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

ABSTRACT

I first discuss Chihara's claim that the presence of Liar-paradoxical sentences presents no problem for our understanding of natural languages, and argue that this cannot be held as easily as he suggests. I then consider the theories advanced by Martin, van Fraassen, Kripke and Burge which attempt to meet some of the problems involved. I argue that the claim in the first two theories that Liar sentences are ill-formed cannot be maintained, and that Burge's theory is methodologically unsound and seriously incomplete. Kripke's theory, though it gives implausible truth-values to some sentences, is more satisfactory. Finally, I discuss Gupta's revision theory, and provide my own alternative which has the following advantages. First, it facilitates our understanding of Kripke's and Gupta's theories, and of their relationship. Second, it provides a third theory in which we can understand how the presence of the Liar does not radically infect a language.

PRECIS

Je discute en premier la thèse de Chihara selon laquelle la présence du paradoxe du menteur ne présentent aucune difficulté pour la compréhension des langues naturelles, et soutiens que ceci n'est pas aussi facile à défendre qu'il le suggère.

Je considère ensuite les théories de Martin, van Fraassen, Kripke et Burge, qui tentent de faire face à certains des problèmes relatifs à ce sujet. Je démontre que la thèse des premières deux théories selon laquelle les phrases du menteur sont mal construites ne peut être soutenue,et que la théorie de Burge n'est pas méthodologiquement solide et presente de sérieuses lacunes. La théorie de Kripke, bien que donnant d'implausibles valeurs de vérité à certain phrases, est plus acceptable.

Finalement, je discute la théorie de révision de Gupta, et présente ma propre alternative, qui a les avantages suivants. Premièrement, elle facilite notre compréhension de Kripke et de Gupta a ce sujet, et de la relation entre leurs théories respectives. De plus, elle nous offre une troisième théorie selon laquelle nous pouvons comprendre comment la présence du paradoxe n'infecte pas totalement une language.

PREFACE

The Liar Paradox has frequently been taken to present problems, even insuperable problems, for our understanding of the role of 'true' in English. Chihara has recently argued that this is not so: the Liar presents few, if any, problems, and current formal theories of truth which attempt to deal with the problems are misdirected. I argue that although Chihara is in one respect right, his position is not as readily maintained as he suggests, and that theories of truth do have a role to play, namely in supporting the claim that the Liar does not nullify all truth-ascriptions.

I then discuss theories of truth presented by Martin, van Fraassen, Kripke and Burge. The first two have been widely criticised: I present some of this criticism, but argue that there are deeper grounds for rejecting these theories. With Kripke, I again present some well-known objections, and use them to explain a certain scepticism Kripke expresses about the possibility of a single theory of truth. Burge's theory has not been widely discussed so far; however, as I argue, it too has serious inadequacies.

Finally, I present a theory which I have developed, which has the following advantages. First, it allows a reformulation of Kripke's theory in which new results about that theory can be established, and which removes the grounds for scepticism discussed earlier. Second, it allows a similar reformulation of Gupta's theory, in which again certain new results are provable. Third, it allows us to compare these two theories, since they are seen to share a common form of construction, and differ only in the "closure rules" which they use. Last, it allows the formulation of an independent theory of truth, in which we can see

that Chihara's first claim can be supported, and hence that in one sense, at least, the Liar is disarmed.

I would like to thank the Philosophy Department for its patient support, and to express my deep gratitude to my supervisor, Professor Anil Gupta, and to Danielle Macbeth for their invaluable assistance.

Table of Contents

Chapter 1

Introduction

In the following dissertation, I shall discuss the Liar Paradox, paying attention to the question of what problem, or problems, it presents us with, and what, in consequence, may be attempted by way of a solution. I shall be concentrating, in particular, on the problems it presents for our understanding of natural languages, problems connected with reference, especially self-reference, truth ascriptions, sentences, negation, and the like, rather than on such technical questions as 'What degree of functional completeness can consistently be combined with truth predication?' Not that such technical questions are irrelevant to the former topics; rather, I shall only enter into them in any detail where the results do seem to have some such significance.

What, in the first place, is the Liar Paradox, and why should it worry us at all? Speaking generally, a paradox is a sentence, or set of sentences, which is or are apparently indubitable, yet from which a contradiction is derivable by apparently irrefutable means. In the case of the Liar, the earliest version is attributed to Epimenides, a Cretan, who said 'All Cretans are liars', but this version has been held by many not to be truly paradoxical. In the first place, the notion of lying involves various intentional aspects which complicate the issue, but even when the notion is interpreted as not telling the truth, the Epimenidean version fails to satisfy the definition of paradox just given. To see this, we can consider the following argument. Suppose no Cretan utterance is true; then Epimenides's claim is true. But it is

itself a Cretan utterance, and hence not true. Consequently some Cretan utterance is true, and Epimenides's is not. Thus, the Epimenidean example implies not a contradiction, but the truth of some other Cretan utterance, and is hence not a paradox. Mendelson, for one, has acknowledged that while the Epimenidean Liar is not strictly paradoxical, the fact that by uttering that sentence, Epimenides could imply the existence of another, true Cretan utterence is "rather unsettling".[1] To bring this home, we can alter the example slightly: if I were to write 'All the sentences I am now writing are false', it would seem that I would bring into existence, miraculously as it were, another, true inscription, though what it would be, heaven only knows.

Nevertheless, the version normally taken as _the_ Liar Paradox is the sentence 'This sentence is false', since it avoids the problems that the Epimenidean version is liable to. I shall refer to this as the Ordinary Liar, to distinguish it from a close relation, the Strengthened Liar, 'This sentence is not true'. Occasionally, when discussing particular theories, I shall refer to various formal expressions as the Ordinary or Strengthened Liar, if these expressions seem analogous. To show that the Ordinary Liar is a paradox in the strict sense, the following argument should suffice.

(1.1)

Let '_o_' name the Ordinary Liar sentence.
Either _o_ is true or it is false. If _o_ is true,
then what it says must be the case, so _o_ is false.
If, on the other hand, _o_ is false, then, since
what it says is what is the case, it must be true.
Hence _o_ is both true and false.
But no sentence is both true and false.

This, then, is a preliminary sketch of what the Liar Paradox is. I shall now make some similarily preliminary remarks on the question 'Why should it worry us?'

---

[1] Elliot Mendelson, _Introduction to Mathematical Logic_, p. 3, fn.

It is clear enough why the presence of a paradox in a formal system is a serious problem, at least in any formal system in which any sentence at all can be derived from a contradiction, because in such cases the distinction between theorems and non-theorems becomes nugatory. One might just as easily declare all sentences to be theorems. Consequently, it is of the utmost importance to choose axioms and rules of inference in such a way that the system is consistent.

When we turn to natural languages, the question of what the problem is becomes less clear. Some people seem inclined to hold the view that there is nothing really to worry about--the Liar does not introduce any serious practical problems in everyday life, in the first place, and in the second it itself is an aberration of some kind, being meaningless, or vague, or nonsense. While I sympathize with the attitude behind this view, I think it is clear that more must be said before we can happily dismiss the Liar. Even if the Liar is meaningless, vague, or nonsense, saying so involves us immediately in two problems. First, we have to explain why it is that the words involved in the Liar sentence seem perfectly meaningful (or specific, or sensible) and are put together in a perfectly grammatical way, without any obvious failures of reference, etc. and yet turn out meaningless. Charges of meaninglessness, in other words, must be substantiated by some account of meaning in which it can be seen that the charges stand. Second, and this is part of the notorious dialectic of the Liar, even if such an account is offered, it has to meet the further challenge of the following sentence:

Either this sentence is meaningless or it is false. (1.2)

For if the Ordinary Liar could be dismissed as meaningless, this cannot be so easily dismissed: if (1.2) is meaningless, then, since that entails that it is either meaningless or false, apparently it is true. So it seems that sentences can be meaningless and true, something which runs contrary to most attitudes to meaning. I am not saying that no

such theory can be constructed, but I am claiming that it is not a trivial matter to do so, and the Liar cannot be lightly dismissed this way.

A different, though possibly related reason for not finding the Liar so troublesome in natural language as it would be in a formal language is also quite widely felt, though not often argued for. This is the view that questions of consistency or inconsistency do not arise for natural languages. Thus Putnam says:[2]

> only theories (systems of assertions) are inconsistent, and natural languages ... are not theories. Someone speaking English may assert that ... all grammatical English declarative sentences are either true or false, but 'English' does not assert this!

Of course in a literal sense the last remark is true: English asserts nothing, only users of English do. I take Putnam's cryptic remark to mean, moreover, that English is simply a medium for expressing various claims, and making assertions, and that if a set of such assertions turns out to be inconsistent, the fault is to be attributed to the assertor, not to English.

However, it does not seem to be true that every inconsistency asserted in English is attributable to the assertor per se. For we are all bound, as users of English, to abide by the conventions of English usage. So I cannot, on pain of being accused of making a mistake, infer 'Some sentence is true and false' from 'Some sentence is true, and some sentence is false', but nobody can readily point out what I have done wrong in inferring 'Some assertion is true and false' from the example of the Liar. Thus it does not seem that the notions of consistency and

---

[2]Hilary Putnam, Mind, Language and Reality, p. 73. Other authors who have expressed similar views include Tyler Burge, in "Semantical Paradox", and Yehoshua Bar-Hillel, in "Do Natural Languages Contain Paradoxes?".

inconsistency are inapplicable to natural languages. In fact, to take an ad hominem turn in the argument, both Putnam and Bar-Hillel go on from the starting point that inconsistency is not predicable of natural language to the conclusion that consideration of the notions of sentence, statement, and the like enable us to see that English is not inconsistent!

So putatively, consistency and inconsistency are applicable to English, and other natural languages. Does the Liar show that English is in fact inconsistent? With some qualifications, Tarski said it did.[3] Insofar as speakers of English could be held to be committed to anything, he held that they were committed to the following principles:

[Tr] A sentence is true if, and only if, what is said to be the case by the sentence is in fact the case. (1.3)

[L] The principles of standard logic.

[R] 'a' can name 'a is not a true sentence' so long as 'a' is a referring expression.

Since [Tr] is not altogether clear or formally irreprehensible, we can consider instead the crucial Tarski schema [Ts], which is intended as an explication of [Tr].

[Ts] X is true if and only if p. (1.4)
(Where 'p' is replaced by any sentence of English, and 'X' by a name of that sentence.)

Since [Ts] seems just as plausible as [Tr], we seem committed to contradiction by the following version of the argument (1.1).

Let 'Jack' name 'Jack is not a true sentence'. (1.5)
1) Jack is true (Hypothesis)
2) Jack is true iff Jack is not a true sentence
(Instance of [Ts]: 'Jack' for 'X',
'Jack is not a true sentence' for 'p')

---

[3] See below, and the discussion in Bar-Hillel, op. cit.

3) Jack is not a true sentence
(Biconditional elimination)

The converse carries through in exactly the same way. Thus if Jack is true, it is not, and vice versa. But an additional well-established principle holds that every sentence is either true or not and hence contradiction follows. It is true that various writers have wanted to restrict this latter principle citing failure of reference, or of presupposition, or category error, but in none of these respects does the sentence Jack seem lacking. Admittedly it is not very common in English to call a sentence 'Jack', but there seems to be nothing particularily amiss in doing so.[4]

Quine similarly seems to hold that our ordinary intuitions about 'true' are inconsistent, but unlike Tarski, who seems to have held that changing our ideas about 'true' would represent a fairly radical shift in the nature of the language, he holds that such intuitions, being essentially arbitrary, can be changed without great upheaval if practical considerations should ever warrant such a change.[5] The case is the same, he thinks, as that in which we adopted new intuitions about sets once the deficiencies of our old ones were revealed and the pressure to establish foundations for mathematics forced reform on them.

This is not the place for a discussion of Quine's pragmatism, but I want to point out that he and Tarski may have been premature in claiming that our intuitions are incoherent: if it should turn out they are right, there may be more interest in the question of whether changing

---

[4]And of course other referential devices will do the job equally well. We are equally puzzled by the last sentence in this footnote. The last sentence in this footnote is not true.

[5]See, for instance, W. V. O. Quine, The Ways of Paradox, pp. 3ff, or Set Theory and its Logic, p.5.

our intuitions does or does not represent a major shift in our conceptual apparatus, but the question is not absolutely settled.

To give some initial plausibility to this point, I want to consider the Russell Barber paradox, in the following form: in Alcala, the barber shaves all and only those who do not shave themselves. The puzzle is, of course, who shaves the barber? This is frequently classified as a pseudo-paradox, because we can see that there can be no such barber, and once we realize this, the problem disappears. However, the example has, I think, served to inspire some attempts to solve the Liar. In such attempts, the project is to find some presupposition upon which the paradoxical argument is based, which, when exposed, will show that the problem no longer arises. It is not sufficient in such cases merely to find some restriction on some or all of [Tr], [L], [R] and [Ts] which will prevent derivation of the contradiction; it is necessary also to find some general grounds for accepting the restriction, relating it perhaps to some other intuitively plausible feature of language, or to ways we actually use the various expressions involved. To make this point clearer, let me take two examples.

Various people have held that what is going wrong in the case of the Liar is the self-reference which is permitted. If we banned self-reference, they go on to say, i.e. rejected [R], no problems would arise. Now there are numerous objections to this suggestion, not the least being that it not only throws away most of the baby, but it does not even get rid of all the bath water. However, the objection I want to raise here is that the suggestion is not well enough motivated: it seems implausible to suggest that there are any grounds independent of the Liar for not wishing to permit self-reference, nor do we have any other problems with usage, with the exception of other paradoxical cases, such as the Berry or Grelling. The point is the following. Unless we have worries about self-reference in situations where paradox

does not threaten, we cannot argue that self-reference is impermissible, and hence that there is no problem about the Liar and other paradoxes. Rather the argument seems to go by reductio: the Liar is a problem, so self-reference is impermissible. But why pick self-reference as the target? Why not ordinary logic, or Tarski biconditionals, or whatever? We must have some independent plausible grounds for rejecting the assumption before we can feel comfortable that the Liar has indeed been dismissed.

Another proposal which has some interest is the Prior-Peirce-Buridan suggestion, in which every statement is held to assert its own truth.[6] This disarms the Liar by making it a merely contradictory, and hence false, assertion, though the suggestion has certain other deficiencies. What is interesting about the proposal is that it seems to fit in very well with certain pragmatic maxims of assertion, which might well be strong enough to provide considerable motivation for the position.

Of course, the proof of the pudding is in the eating; I merely want to suggest at this point that it is an open question whether or not English is inconsistent. The strong plausibility of [Tr], [L] and [R] suggest that it may be, but the possibility remains that it is consistent, and Prior's attempt, for one, gives us hope. What I want to discuss now is an argument given by Charles Chihara[7] which might seem to deny that hope, leaving all those who have attempted to provide a solution along the lines just sketched in the position of merely having provided alternative reforms of our (inconsistent) intuitions.

---

[6] A. N. Prior, "Some Problems of Self-Reference in John Buridan".

[7] Charles Chihara, "The Semantic Paradoxes: A Diagnostic Investigation", and also "A Diagnosis of the Liar and Other Semantical Vicious-Circle Paradoxes". The former paper will be referred to as TSP.

Chihara starts with the claim that paradoxes present two problems, the diagnostic and the preventative. The former is the problem of explaining what has gone wrong, and how it happened, and the latter is that of constructing systems in which it cannot arise. To have some idea of a diagnosis of a paradox, Chihara gives a fairly simple example, called the Sec Lib paradox. Suppose that a community has several clubs whose secretaries are not eligible for membership, and that these secretaries, feeling disgruntled, decide to form a club, Sec Lib. The eligibility condition for Sec Lib, then, is:

[SEC LIB] For every person x, x is eligible to join Sec Lib iff there is a club of which x is a secretary and which x is not eligible to join. (1.6)

All goes well, and Sec Lib flourishes to the extent where it decides to hire a secretary, Ms. Fineline. Problems arise when Ms. Fineline applies for membership of Sec Lib. Even this might not be totally disastrous, except that there is no other club of which she is secretary. In that unfortunate case, she turns out to be eligible iff she is not eligible.

What is the diagnosis here? Well, the contradiction arises from the eligiblity condition, [SEC LIB], plus two empirically ascertainable facts: that Ms. Fineline is secretary of Sec Lib, and that she is secretary of no other club. Thus we are led to say that the trouble arises because [SEC LIB], though false, seems to be perfectly in order: first, because it seems to be just the kind of thing we can declare true, and second, because, Fineline apart, it actually settles questions of eligibility quite adequately.

Chihara goes on to examine the Grelling and Berry paradoxes, and then turns his attention to the Liar. Once again, we find that a contradiction can be derived from a principle [T], another restricted version of [Tr], together with an appropriate Liar sentence.

[T] If α is a sentence consisting of the referring expression β immediately followed by the words 'is not true', and if β denotes the sentence γ, then α is true if, and only if, γ is not true. (TSP, p. 605) (1.7)

As before, the Liar sentence could be the sentence Jack, namely 'Jack is not true'. From this we can infer that Jack both is and is not true. One part of the diagnosis remains the same as before: the responsible principle is [Tr]. In this case, however, [Tr] has not been laid down, as was [SEC LIB], so that there remains the question of how we came to believe it. One way, Chihara suggests, was that we came to believe it as an empirical hypothesis: we found that those sentences that were true were just the ones for which things were the way they said they were. However, Chihara thinks that this account cannot explain the intuitive attractiveness and resistance to rejection of [Tr]. He couples this picture of the basis for [Tr] with another view about theories of truth, which he calls the consistency view; according to the consistency view, a correct theory of truth must be "consistent with all known facts" (p. 607; his italics).

Chihara finds the consistency view implausible; given theories which have presumed it, it is extremely hard to see how we ever could have learned such a complex account, either as a species in general, or as individuals one from another. Instead, then, he proposes an inconsistency view. According to this, [Tr] expresses the accepted conventions which give the meaning of 'true', just as [SEC LIB] expresses the explicit decision on the eligibility condition for Sec Lib. Taking this view, we can readily account for the ease with which we learn the concept, and for the intuitive acceptability of [Tr]. Finally, Chihara says his diagnosis can pass "a formidable test ... which few, if any, earlier proposals could pass convincingly". The test is, in brief, 'Is Jack true or not?' Chihara's diagnosis tells us that

Jack both is and is not true. Other solutions, based on the consistency view, are bound to have to say either that it is or that it is not. But then they have difficulty saying why it is not in turn the other. That is, having reached the verdict that it is true, how can they avoid having to go on and say in turn that it is not, since it is true and says it is not true?

There is no doubt that this problem is a severe one for theories espousing the consistency view, possibly the most severe one they face, and we shall see some of the difficulties when we examine such theories later. However, I should like to express grave doubts that all the advantages lie with the inconsistency view. In discussing the consistency view, Chihara says that few people have ever bothered to defend it, except for a brief remark from Parsons, that supposing that our language was inconsistent would be to attribute to ourselves "an incoherent conceptual apparatus".[8] Chihara replies that it would be no such thing: we can see that the inconsistency involved in the Sec Lib paradox, for instance, though of course creating problems for the question 'Is Ms. Fineline eligible?', is perfectly adequate to answer the question 'Is Joe Schmoe eligible?'. The conventions, although inconsistent, are not unusable, or incoherent. Later, again, Chihara emphasizes that it is extremely implausible that the semantic paradoxes create any difficulties in, say, determining the trajectories of rockets, or any other aspect of everyday or scientific life.

Chihara's reply seems plausible: Mackie gives the analogy of a car which is likely to explode if driven at ninety miles per hour, which can nevertheless be driven with perfect confidence at lesser speeds,[9] and we

---

[8]Charles Parsons, "The Liar Paradox", p. 399.

[9]J. L. Mackie, Truth, Probability and Paradox, p. 251.

can similarly imagine a remote Aristotelian tribe, who every day chant 'To say of what is that it is, or of what is not that it is not, is true', and who have never had the misfortune of meeting any breast-beating, confessional Cretans or other formulators of paradox. We could examine their uses of 'true', and would find them admirably consistent, except, of course, for occasional mistakes.

Nevertheless, in an important sense their truth predicate would be incoherent, and ours would be too, if Chihara's account is correct. For suppose a malicious Buridan visited our tribe, and presented them with an inscribed tablet:

Grass is not green. All the sentences on this tablet are false.	(1.8)

Being no doubt curious, and remembering their Aristotelian dictum, they might argue as follows:

> Well, the second sentence on the tablet cannot be true, because that would be contradictory, so it must be false. So one of those two must be true, and it must be 'Grass is not green'. (1.9)

The point is that neither they nor we can rationally defend saying that <u>any</u> sentence is either true, or false, rather than the opposite, if all we have to rely on is something like [Tr] plus [L]. The proclaimed achievement of the inconsistency view, that it can justify calling Jack both true and not true, is its downfall, since it can equally well justify calling any sentence both true and not true. It might be tempting to suppose that we could justify asserting '"Grass is green" is true' rather than '"Grass is not green" is true' because we can get the former sentence directly by application of [Tr] on 'Grass is green', rather than by the piece of Buridan funny business above, but while in the end I think that 'No Funny Business' is an important part of our

intuitive notion of truth, introducing it here is to acknowledge that [Tr] is not the whole truth about truth.

Another point is connected with the previous one. In discussing the Sec Lib paradox, Chihara clearly said that [SEC LIB] is false. In other words, it is not true that x is eligible iff .... But if that is not true then what is? Who is eligible? Chihara emphasized his claim that nobody else's eligibility was affected by the fact that Ms. Fineline is a problem case, but we might ask any of the members: 'By what right do you claim membership in Sec Lib? The condition under which you were declared eligible is false.' In point of fact, of course, we have practical conventions which settle these matters, so that presumably all past members would be allowed to continue as members and to decide on a new eligibility condition, but the logical point is that no reference to the old eligibility condition alone could guarantee eligibility for anyone. Similarly, we must regard [Tr] as false, since it implies a contradiction when conjoined with some empirical facts, so that no truth ascription can in fact be justified by reference to [Tr]. Further we may ask: Since [Tr] is false, what are the correct conditions for something to be a true sentence?

Finally, I must dispute the significance of passing the "formidable test". I agree with Chihara that [Tr], or something exceedingly like it, is the principle we mentally apply when faced with the Liar (or any other sentence). However, it is clear that in practice, we are not led (or, perhaps better, misled) by [Tr] into saying that the Liar is genuinely true and not true. Of course, if we are inattentive, we may at one point argue that it is true, and at another that it is not, but if our attention is drawn to the contradiction, we do not proudly hold to it, saying that we are committed to it by the meaning of 'true'; rather, we admit puzzlement and do not know what to say.

It seems useful here to recapitulate points of agreement and

disagreement. I agree with Chihara in the first part of his diagnosis, that we are tempted towards paradox by [Tr], and that it is plausible to say that [Tr] is the principle of truth that we all believe, and learn, or at least part of it. I also agree that in practice our use of 'true' is untainted by paradox. It is clear, however, that we cannot say that [Tr] expresses "the conventions which give the meaning of 'true'" (TSP, p. 611), if meaning is to have any relevance at all to use, or truth conditions, or assertability conditions. Moreover, it seems an essential aspect of the diagnosis of what is happening with truth and the Liar that it should explain not only what tempts us towards paradox, but also how we avoid falling for the temptation, with all its dire consequences.

I would like to offer a distinction here which may help to clarify matters. If we are interested in semantical and logical notions, we can attempt to express various principles which seem intuitively to govern these notions, and we may produce such well-known principles such as the Law of the Excluded Middle, or the Principle of Mathematical Induction, or a principle like [Tr]. These I take to be based on reflections about the concepts involved, and, presumably, their fit with other such notions. Moreover, a suitably reflective person, once introduced to these principles, recognizes them as tied up closely with the meanings of the words involved, and finds it puzzling if they are questioned.

Simultaneously, however, the reflective user of English has to admit that his use of concepts involved cannot always be referred straightforwardly to these principles. Reference to Mathematical Induction does not immediately lead us to accept every slippery slope argument, even though that is the form of the argument. So we have two classes of intuitions: intuitions about the principles of truth, inference and so on, and intuitions about which sentences are true, and which arguments valid, and neither class seems readily reducible to the

other. The principles do not seem to generate the use, nor the use the principles.

The situation here is very similar to Putnam's distinction between stereotype and extension. For natural kind terms, for example, we want to be able to say that 'zebra' means something like 'black and white striped horse-like animal', without thereby making the claim that there are, or might be, unstriped zebras, contradictory. I shall not discuss the details of Putnam's theory here, as they become quite complicated, but the basic idea is that 'black and white striped horse-like animal' is the stereotype appropriate to 'zebra', but it does not fix the extension of the term nor vice versa. Thus it would be equally mistaken to try to decide whether there were any non-striped zebras by reflecting on the stereotype, as it would be to try to determine the meaning of 'zebra' by finding out everything about zebras.

One disanalogy between the case of semantic principles and that of natural kinds is that while in the latter case we can easily distinguish between whatever meaning we invest the word 'zebra' with and however zebras turn out to be, in the former it seems odd that reflection on semantic concepts should not match how we use them. I think the reason for this is that it is not necessarily apparent on reflection that there may be the hard cases, possibly involving conflicts between various principles; the principles are in fact hasty generalizations over central cases, and leave out the tricky ones. One way of representing the semanticist's job, then, is as looking for a coherent picture which retains the central intuitions, but also fills in where there were conflicts.

Specifically, then, with respect to truth, I accept [Tr] as a stereotype for 'true', but deny that that is all that semantics has to say on the matter. In particular, we have to find an explanation of why we are justified both in saying that 'Grass is green' is true, and in

withholding judgment on the Liar, when [Tr] gives us grounds for calling both of them true and false. Of course, in admitting that [Tr] captures our conception of truth fairly well, I am admitting that that conception is inconsistent, and perhaps incoherent. However, I want to deny that any serious consequences follow from that admission.

First, it cannot be used to show that English is inconsistent in the most important sense. For to show that, we would have to show that the conventions of usage of English lead to contradiction, and I have carefully divorced the intuition which [Tr] captures from any conventions of use. Second, it does not open the door to a Quinean response to the Liar Paradox. That is to say, I have not conceded that our natural conception can be replaced by any other useful theory without regard to our intuitions, chosen merely for its practical application. In fact, I would not concede that the notion of replacement is in the least appropriate at this point.

What we do need is an understanding of how we are able to use 'true' as safely as we do, given our incoherent stereotype of truth. However, this understanding need not be an analysis of whatever rules of the mind enable us to reach decisions without serious mistakes--that seems to be a psycho-linguistic investigation. It is rather, for the semanticist, a question of whether we can find a theory of truth and associated notions which is consistent, agrees with whatever coherent intuitions we may have about the concepts involved, and agrees with our intuitions about the truth of sentences. If we can find such a theory, we can reasonably be said to have shown that the Liar does not show that English is inconsistent. Instead, we can justify our claims about the truth of sentences, and show, what is so far a mystery, how we are able to use 'true' without falling into the traps that [Tr] sets.

Notice that a theory of 'true' in this sense is not trying to compete with [Tr] in any way: it makes no claim about its intuitive

obviousness, or ease of learning, or any such thing. Nor is it proposed as a replacement for [Tr], as a consistent theory in which logic, semantics, and metamathematics could safely be pursued; it would, instead of being a replacement, be evidence that, despite [Tr], we were justified in using our ordinary intuitions in studying these pursuits. The only reason a question of replacement might arise would be if in some less natural theory of truth, such studies were simpler. Such a situation seems, however, fairly remote.

Suppose, nevertheless, that despite the continued efforts of semanticists, no satisfactory theory could be found. Would we have to admit that maybe, after all, English was inconsistent? Of course, unless someone could produce a categorical proof that no satisfactory theory existed, we would only have inductive grounds for admitting defeat, but they might become quite powerful. In two short articles, however, Hans Herzberger has argued that the mere suggestion that a language is inconsistent is itself contradictory.

In the first, "The Logical Consistency of Language", Herzberger argues that no language can have an inconsistent set of analytic sentences, and the kernel of the argument is short and sweet: if A is the set of analytic sentences of the language L, the supposition that A is inconsistent entails the two claims that some member of A is false (because A is inconsistent) and that every member of A is true (because all analytic sentences are true). Consequently, the supposition is contradictory. In "The Truth-Conditional Consistency of Natural Languages", he extends his claim: the supposition that a language could be inconsistent in its truth conditions is also contradictory. I shall not go into the details of the argument here--briefly, he takes the supposition to be that there is a possible truth condition for which an inconsistent set of sentences all are true, and has little difficulty deriving a contradiction from that supposition.

However, the consequences of these arguments are not altogether heartening for the view that I have been adumbrating. Admittedly, they appear to guarantee that there must be some consistent theory which can be taken as giving the truth-conditions for every use of 'true'. Unfortunately, they carry no guarantee that the result would be at all recognisable as a theory of truth: it might have no characteristics at all that we had expected. For instance, it might require us to deny that the Liar is a sentence at all, or force a different interpretation of 'not', or rule that 'true' is ambiguous. Thus it remains an open question whether there are any theories which are consistent and yet still conform to plausible constraints; of course, I have left it extremely vague as to what counts as plausible constraints, but this should become clearer when we look at some actual proposals, and examine their advantages and disadvantages. Some of the constraints, however, are worth introducing immediately.

Although initially I suggested that a sentence like 'This sentence is false' was the crucial case of the Liar Paradox, it should not be thought that finding a theory which disarms this sentence alone ends the matter. I have already mentioned the Strengthened Liar, whose importance will soon become apparent, but there are numerous other versions which are also important. In fact, it has frequently been observed that the Liar is excessively protean, changing into something vicious again just when it seemed pinned down. Many of the variants are only of interest as problems for specific theories, but I shall give a short list here of some of the more generally important ones.[10]

The Indirect Versions: (1.10)
A card has written on one side only the

---

[10]For an amusing collection of other examples, see Mackie, pp. 296-301, and also Patrick Hughes and George Brecht, Vicious Circles and Infinity.

sentence 'The sentence on the other side of this card is false', and on the other side only the sentence 'The sentence on the other side of this card is true'.

This example alone is enough to show that no naive ban on self-reference will resolve all the paradoxes.

Quantified Versions: (1.11)

a) The Epimenides Paradox
b) Jones: "Everything Smith says is false".
Smith: "Sometimes Jones speaks the truth".

These two assertions, like Epimenides's, may get off without trouble if the facts are all right. Here the problem arises when Jones and Smith have always otherwise spoken falsely. I gave an argument above that suggested these examples were not strictly paradoxical. However, Kripke has emphasized the extreme importance of being able to provide a plausible account of such cases. Moreover, some of them are the worst possible cases, resisting analysis long after other versions have been successfully subdued.

Truth-Teller Versions: (1.12)

The Ordinary Truth-Teller:
This sentence is true.

Truth-Teller variants of all the ordinary Liar forms are easily constructed.[11] Again, Truth-Tellers are not properly paradoxes: they are true if they are true, and false if they are false, but no contradiction, or even surprising consequence, results from supposing them either true or false. Nevertheless, they are a problem, because their truth or falsity is completely undetermined by any straightforward combination of empirical facts and logico-semantic inference. In one

---

[11] One of my favorites is the supererogatory barber of Alcala who shaves all and only those who shave themselves.

sense they are "don't care" cases: it does not really matter to our intuition what account is given of them. But some account must be given.

Another methodological question concerns the notorious question of the bearer of truth. Already it will be clear that I am treating 'true' as a predicate applied to sentences, and, furthermore, all the theories I shall discuss do likewise. I do not want to be interpreted as holding that in fact, in English, 'true' is normally applied to sentences, nor as holding that in the best analysis of sentences, statements, propositions, 'true', etc., that 'true' must end up as a predicate of sentences. Indeed, I share Kripke's suspicion that a full resolution of the problems presented by the Liar can only finally be achieved by a more thorough analysis of these notions, and that the result might well be that 'true' turns out to be applied to whatever it is that sentences express, rather than sentences themselves. However, the presupposition I share with the authors I discuss can be regarded as a methodological expedient with several significant advantages.

First, statements and propositions are notorious for the problems attached to their identity conditions. This does not mean that I therefore take it that they do not exist, but given the problems which analysis of the Liar presents anyway, the fact that sentences are relatively trouble-free in this respect is a considerable advantage.

Second, even if the ultimate theory attaches truth to statements, say, it must also contain the predicate 'expresses a true statement' which attaches to sentences, for the theory must provide some account of how sentences can be used to make statements, and hence will need to contain an account of which ones make true statements. In that case, we can just regard current investigations of the sentence-predicate 'true' as doing duty for the predicate 'expresses a true statement'. That is, if investigation of the sentence-predicate can produce interesting

results, they must appear in the more comprehensive theory in some guise or another.

Third, it should not be thought that retreat to statements can immediately solve the Liar Paradox. It has commonly been held that 'true' properly applies to statements, say, and therefore we can happily dismiss the Liar by saying that 'This sentence is false' makes no statement, and hence is neither true nor false.[12] It may be that in the end this is what should be done. However, let me make it quite clear that merely doing this leaves several questions unanswered. What should the response be to 'This sentence makes no true statement'? It is tempting to say that it both makes no statement and is true, and this problem must be resolved. Again, since we can find ourselves in paradox by accident, as in the quantification examples, it can turn out that whether or not we succeed in making statement with a given utterance of a sentence is a matter of circumstance, something which has not generally been acknowledged in theories of the relationship between sentences and statements. These and many other problems are discussed by Thomason in an interesting paper, "Paradoxes of Intentionality?".

The point again is that merely making the distinction between statements and sentences does not solve any problems, because variants of all the simple forms of the paradoxes can be formulated in terms of statements, and to these problems are added the special problems about the nature of statements.

Thus it seems easier, at this point in the development of logic and the philosophy of language, to use the theoretical simplification of taking 'true' to be a predicate of sentences.

I shall conclude my introductory remarks with a brief discussion of

---

[12]See, for example, William Kneale, "Propositions and Truth in Natural Languages".

Tarski's work on truth, because although it is clear that as it stands his theory of truth cannot be happily applied to natural languages, all subsequent work on the Liar, and especially the work discussed here, has to be considered in response to his results.[13] Tarski was concerned to construct a semantical definition of truth which conformed to the conception embodied in [Tr], yet which also had the virtues of formal correctness, clarity, and freedom from ambiguity, and as a step in that direction suggested that instances of [Ts] could be seen as partial definitions of the expression 'X is a true sentence'. However, he thought that overwhelming difficulties present themselves if we try to convert this schema by generalizing it to give us a general definition of 'is a true sentence'. Some of these difficulties attach to problems about how to understand quotation names, especially in an extensional language, which I shall not go into here; others, of course, relate to the possibility of expressing a Liar paradox. These problems led him to think that no satisfactory definition of 'is a true sentence' could be constructed in a natural language, and he expressed a belief about the general features of natural languages which led to this. He thought that the significant feature was the 'universality' of natural languages: if we can speak about anything in any kind of language, we must be able to speak about it in a given natural language. In particular, we can talk about the syntactic and semantic properties of languages, so every language must contain names for its sentences and expressions, and predicates like 'true', 'refer to', and so on. But once we have such resources, we can formulate the paradoxes, and all is lost.

Reflecting on the properties of natural languages, Tarski made the

---

[13]The prime source is A. Tarski, "The Concept of Truth in Formalized Languages" (CTFL), but see also "The Semantic Conception of Truth" (SCT).

following claim. No language in which the following conditions hold can be consistent:[14]

[TC] I The language contains a name for each of its sentences.
II All instances of [Ts] are true sentences of the language.
III A Liar sentence can be constructed in the language.
IV Ordinary logic holds (i.e. [L]). (1.13)

This claim led Tarski to two conclusions. First, since natural languages seem to satisfy all four conditions, there seems no possibility of constructing a definition of truth for natural languages; only for formal languages are there still any prospects for success.

Second, since he saw no possibility of rejecting condition IV, the only formal languages for which a definition of truth was possible would be ones in which one or more of I to III were not satisfied. He called languages which satisfy conditions I and II semantically closed, and his valuable result was to show that for certain formal languages that are not semantically closed, a truth definition can be constructed, not, to be sure, in the language itself, but in a metalanguage which names the expressions of the object language.

What is slightly odd about Tarski's claim is that he appears to have held that any language which is semantically closed also satisfies condition III. However, in the second section of "Truth and Paradox", Anil Gupta has shown that if the syntactic resources of the language are weak enough, a semantically closed language can contain its own truth predicate. This result alone, however, holds out little hope of showing the consistency of natural languages, since they are considerably more

---

[14]This formulation of the conditions comes from CTFL. In SCT, conditions I, II, III, and IV are rearranged: I and II are combined, and III and IV are interchanged.

powerful in their syntactic resources.[15] In the foregoing I said that to Tarski there seemed no possibility of constructing a truth definition for natural language; this remark needs some qualification and explanation. At one point, Tarski says that his discussions of the Liar and the problems of quotation "emphatically prove" that the concept of truth in natural languages inevitably leads to confusions and contradictions (CTFL, p. 267). On other occasions, he only says that the possibility of a consistent use of 'true sentence', and hence of a definition, seems very questionable (CTFL, p. 165). In "The Semantic Conception of Truth", the suggestion comes that, since definitions of truth can only be given for languages of precise structure, and since natural languages do not have appropriately precise structure, the question whether natural languages can have a truth definition is "more or less vague" (SCT, p. 19). Actually, it seems that the questions of consistency and the definability of truth diverge with respect to languages of imprecise structure: if the structure of the language is not precise, a Tarskian-style truth definition cannot be constructed for formal reasons, even if the language is consistent, and conversely, even the most vaguely constructed language could prove definitely inconsistent if it were sufficiently clearly committed to some

---

[15]Exactly what condition is necessary to make a language inconsistent remains an open question--see "Truth and Paradox". On the other hand, it is clear that if a language contains the syntactic function *, which takes any sentence S which begins with 'Every sentence', replaces 'Every' by 'The', and then inserts 'S' after the words 'The sentence', then the language is inconsistent (assuming the other conditions hold). This function is taken from footnote 11 of SCT, p. 43, in the Linsky reprint. The contradiction arises when we suppose that we can define a predicate 'is self-applicable', such that a sentence S is self-applicable just in case S* is true. Then we can ask whether 'Every sentence is non-applicable' is (self-) applicable or not, and of course if it is, then it is not, and vice versa. This is obviously a fairly powerful syntactic device; what lesser functions would suffice is still unclear.

contradiction. Thus the only grounds Tarski can have for hedging bets about consistency would be doubts about the extent to which natural languages do clearly satisfy conditions I to IV.

The further suggestion which Tarski made with respect to natural language definitions of truth concerns what would be required to adapt his formal result to represent truth in natural languages. He candidly admits that it would be an awesome task, and that results would bear little resemblance to the original phenomenon. First, we would have to remove any vagueness in the non-semantic part of the language, and formalize that. Then we would have to construct a metalanguage in which the semantics for that language could be defined, and then another for the semantics of that language, and so on in an unending hierarchy (CTFL, p. 267).

I am not sure that anybody ever boldly held that this is how truth is, or that this is how to understand it. Many writers have taken Tarski to hold this, and attacked him for it, but it is clear that he makes no such claim. However, I shall briefly give some of the objections which have been raised against this "Tarskian" account, because although they seem well-founded, there have been attempts to revive more or less Tarskian theories, and I shall consider one below, given by Tyler Burge, so it is interesting to see whether these objections can be met.

Two of them are quite straightforward. The first is a bold denial that we use any hierarchy of languages: we do not have one predicate, 'true_1' which is applied to sentences not containing 'true_i' for any i, and then another which is applied to any sentences containing 'true_1', and so forth. The Tarskian account may be offered as a revision, but it cannot claim to represent anything we currently do. The second objection makes the same point more politely, perhaps. It might be that some of our utterances could be understood as if they did contain

predicates like '$true_i$'. Nevertheless, the claim goes, there are some uses of 'true' which are intended to be understood as having no restriction on what sentences they apply to: for instance, 'Every sentence is true or not', or 'God is omniscient' (i.e. knows all truths). Since every truth predicate in the hierarchy can be applied only to sentences occurring at the level or levels below, no truth predicate can apply to all sentences, on pain of admitting the Liar. Thus just as a ban on self-reference impugns 'This sentence is in English', with no apparent justification, so does a Tarskian ban on a so-called "global" truth predicate impugn the apparently satisfactory 'God is omniscient'.

Two more complicated objections have been put forth by Saul Kripke in his "Outline of a Theory of Truth", which I shall just mention here, and take up in more detail when I discuss that paper below. They are, first, that it is necessary for the hierarchy to be capable of handling some cases that extend it beyond finite levels, but there seem to be difficulties in so extending it. The second objection is that on many occasions we can find examples of sentences to which it might be plausible to attach levels, yet the Tarski account gives implausible truth values.

Faced with all these objections, it seems that we can agree with Tarski that it is improbable that the language hierarchy represents with any degree of accuracy the natural-language use of 'true'. The question then is whether any other theory works any better, and this is what we shall go on to find out.

Chapter 2

Martin and the Category Approach

In a series of articles, Robert L. Martin has expounded and defended what he calls a "category approach" to the Liar.[16] One of the points of interest of these articles is that he commits himself to certain methodological claims about what is to count as a satisfactory solution of the paradox. Unfortunately, I think it is ultimately clear that his own theory does not satisfy those requirements.

Basically, he sees the Liar as presenting us with a challenge to our understanding of various concepts of our language, and a solution to it as casting doubts on some assumption which, while plausible, leads to contradiction. However, these doubts must not be generated by reflecting on the fact that this assumption leads to a contradiction, and hence concluding by reductio that it cannot be true; that argument would be palpably ad hoc. Instead, he says:

> What is wanting, ideally, is the uncovering, the making explicit, of some rulelike features of our language which when considered carefully have the effect of blocking at least one of the assumptions of the argument; if not actually showing an assumption to be false, at least casting doubt upon it. (CSL, p. 91)

Despite this methodological claim, Martin's expressed objection to the

---

[16]Principally in "Towards a Solution of the Liar Paradox" (TSLP), "On Grelling's Paradox", "A Category Solution to the Liar" (CSL), "Sortal Ranges for Complex Predicates" (SRCP), and finally, with Peter W. Woodruff, "On Representing 'True-in-L' in L" (ORTL).

"Tarskian" hierarchy outlined in the Introduction is not that it is ad hoc, but that it rules out certain kinds of self-referential sentences, like 'This sentence is in English'. Kripke has complained that this charge is ill-founded, and that Gödel has shown that self-reference is permissible.[17] What Martin seems to have in mind, however, is that for the hierarchical approach not to be ad hoc, there has to be a feature of language which can justify it, and the only one available, he seems to think, is the use-mention distinction. If a hierarchical approach to truth depends on the use-mention distinction, however, consistency demands that a hierarchy be constructed even for 'is in English', and hence no self-reference is possible.[18] With respect to Tarski's four conditions, Martin is going to reject [L], and adopt a three-valued logic. Since [Ts] turns out not to be true in all cases, it is rejected also, and the well-formedness of Liar sentences is also denied. Thus only condition I remains unchallenged.

The 'rulelike feature' which Martin's theory takes up is that of a semantic category. When we are confronted with the task of developing a theory of the deviant sentences of a language, we can readily identify some as ungrammatical--'Hat top very is a virtuous', for example--while others seem more or less grammatical, i.e. have the items of the right syntactic categories in permissible combinations, yet are stll plainly deviant, e.g. 'A top hat is very virtuous'. To explain this second kind of deviance, it is plausible to introduce the notion of the semantic category of a word, somehow related to its meaning, so that well-formed sentences would not only be grammatical, but also have words of the right semantic category in the right places. Martin does not claim to

---

[17]"Outline of a Theory of Truth", fn 13, p. 698.

[18]This seems to be what is suggested in TSLP, p. 280.

offer a systematic account of semantic categories, since such a theory would have to be extraordinarily wide-ranging; what he does is develop various results which, he claims, "any satisfactory theory would yield" (TSLP, p. 286). In particular, he concentrates on category mistakes arising out of subject-predicate category mismatches, although other kinds are clearly possible: e.g. 'Two plus two equals four violently'. However, this restriction seems adequate to discuss the Liar and other such sentences, and no problems seem to arise because of it.

If we look at a typical example of a sentence of subject-predicate form which seems to make a category mistake, like 'Virtue is triangular', the natural response is to say that there are things which we can comprehensibly assert to be triangular, and virtue does not happen to be one of them. Thus we associate with a predicate a range of application (RA), which is a set of objects to which the predicate is properly applicable (whether it is truly or falsely applicable is another matter altogether; ideally, knowing the RA of a predicate is part of knowing its meaning, so whether a predicate is applicable to an object can be determined before questions of fact arise). So if the referent of the subject term of a sentence falls in the RA if the predicate, the sentence is semantically correct, otherwise it is semantically incorrect. We could say that the RA of 'is triangular' contains physical objects and certain abstract mathematical ones, and since virtue is neither a physical nor a mathematical object, 'Virtue is triangular' is semantically incorrect.

To get nearer to the Liar, various obstacles have to be overcome, or bypassed. The first is the problem that, though in 'Virtue is triangular' we only had to ask whether virtue fell in the RA of 'is triangular', we do not want to have to decide whether 'The present King of France is bald' is semantically correct or not by asking whether the present King of France falls in the RA of 'is bald', since that seems

rather hard to say. What seems appropriate in this case is not to examine the referent, if any, of the referring expression, but just to examine its sense. On the other hand, we do not always want to judge by the sense, Martin thinks, because then we could decide that 'This semantically correct sentence is false' is semantically correct, and paradox would follow. So sometimes it seems important to follow sense, sometimes reference, and the question arises which is correct on any given occasion. Martin thinks that the line is drawn at self-reference: for self-referential sentences, take the referent, and for others, take the sense. I am inclined to doubt that this is the right place to draw the line, but will agree that in the case of 'This sentence is false', and its variants, the appropriate choice is the referent.

A problem which compounds the previous one is that many sentences are paradoxical only in certain circumstances, as we have seen with the Epimenides case among others. If the project is to show that Liar sentences are semantically incorrect, and semantic correctness is essentially a function of the meanings of words, these examples clearly present problems. For why should 'All Cretans are liars' be semantically correct (but false) on some possible occasions of use, and semantically incorrect on others? It seems that it should just be always correct or always incorrect. Martin suggests that in these cases an associated idea, token oddity, will do the necessary work; I suspect that it cannot, but since I hope to show that even in the central example, Martin does not make good his case that semantic incorrectness is at the heart of the Liar, I shall not pursue this question here.

Having agreed that the appropriate object to consider when determining the semantic correctness of the Liar sentence is the Liar sentence itself, we have next to determine the RA of 'is false'. Martin suggests that since 'true' and 'false' are partners in the same way that 'yellow', 'blue', 'green', etc. are, and since in the second case it is

clear that the RA of any of them is the RA of any other, the RA's of 'true' and 'false' are the same. Now it is plausible to hold that a sentence which is semantically incorrect is neither true nor false, just as meaningless sentences are, so equally it is plausible that for a sentence to be true or false, it must be semantically correct. But this suggests that the RA of both 'true' and 'false' is just the semantically correct. In other words, Martin says, '"Saturday eats algebra" is false' is just as semantically incorrect as 'Saturday eats algebra'.[19] Given these resources, we can turn to the question of whether the Liar is semantically correct or not. This question turns on whether o (if this is the name of 'This sentence is false') falls in the RA of 'is false'. But the latter is just the semantically correct sentences, so it turns out that o is semantically correct just in case it is semantically correct. Since this decision process can clearly never yield the result that o is semantically correct, Martin argues that it is unreasonable to assume, as the premise of the derivation of the contradiction does for the Liar, that it is semantically correct and hence is either true or false. Thus we are able to block the derivation of the contradiction, and the paradox is unmasked, because we can show that one of the premises rests on a tacit, and illicit, assumption.

Furthermore, the solution clearly does not result in any ban on self-reference in general: we can perfectly well understand 'This sentence is in English', or 'This sentence is interesting', because the RA's of the predicates involved are clearly both much wider than just the semantically correct sentences of English, so that whether those sentences fall in the RA's of their predicates can be determined independently of the test for semantic correctness.

---

[19]There are other considerations here; I shall look at some of them later.

Although I have expressed doubts that Martin's approach can encompass all the variants of the Liar, it is interesting to note that it handles the Truth-Teller quite happily: since the RA of 'is true' is just the same as that of 'is false', we can see that we are under no pressure to say that the Truth-Teller is either true or false, and can relegate it to the ranks of the semantically incorrect if we wish. However, overall the account so far is clearly limited in its scope, so Martin expanded and revised it on several occasions, producing the final version in an important paper entitled "On Representing 'True-in-L' in L", written with Peter W. Woodruff. In this paper, they present a formal language which, they show, can contain its own truth predicate. Before I give the formal details of this language, which I shall call MW, I shall discuss the background assumptions, based on category considerations, which lead to the particular choice of valuation rules for MW.

Much of the motivation for Martin's choice of valuation rules depends on intuitions about complex predications, as in such examples as 'Virtue is triangular or unattainable' and 'Virtue is not triangular', which we can think of as predicating 'is-triangular-or-unattainable' and 'is-not-triangular', respectively, of virtue. I shall start with the issues concerning disjunction. There are strong intuitions that 'Virtue is triangular or unattainable' is just as bad, semantically, as 'Virtue is triangular', and that 'The cup is coloured or yellow' is just a redundant way of saying 'The cup is yellow'. These and similar intuitions suggest that we can treat the RA of a disjunctive predicate, which I shall follow Martin in abbreviating as $R \vee S$, as the intersection of the RA's of R and S. Since we can go on to think of any disjunction as a complex predication to an appropriate kind of object, however, this intuition leads us to the general rule for disjunction, that if either disjunct is semantically incorrect, and hence neither true nor false,

the disjunction is likewise; otherwise if one disjunct is true, the disjunction is true, and if both are false, it is false.[20]

There is, however, a strong counter-intuition about the truth of disjunctions, which is that a disjunction is true if one of its disjuncts is, whatever the other may be. Thus, '2+2=4 or virtue is triangular' is held to be true because of the truth of '2+2=4'. This intuition also preserves the inference from any sentence A to AvB, which the preceding account does not. The difference between these two is given in the Kleene 'Strong' and 'Weak' three-valued tables, though van Fraassen's supervaluation technique also preserves this aspect of the strong intuition.

Strong Disjunction (2.1)

	t	f	u
t	t	t	t
f	t	f	u
u	t	u	u

Weak Disjunction

	t	f	u
t	t	t	u
f	t	f	u
u	u	u	u

Unfortunately, there are further counter-intuitions. As Thomason has pointed out, a sentence like 'Today is Tuesday or today is less than ten' is, on the Strong view, true on Tuesdays and otherwise semantically incorrect. However, this makes semantic incorrectness partly a matter of fact, contrary to the intuition that it is an aspect of meaning. Thomason takes the moral of this story to be that semantic incorrectness is only clearly a property of atomic sentences, and cannot readily be extended to complex ones. But Martin regards this as unjustified: in particular, since we may apparently have, in natural language, a

---

[20]This discussion occurs in SRCP: at various different stages, Martin espoused different disjunction rules. See, in particular, the confusion in CSL, p. 99 and fn 3, et seq. For Thomason's objections, discussed below, see "A Semantic Theory of Sortal Incorrectness".

predicate which is synonymous with a disjunctive predicate, the question of which sentences are truly atomic or complex becomes rather arbitrary. Moreover, it is logically possible, if the RA of the atomic predicate is not actually the intersection of the RA's of the predicates in the disjunction with which it is synonymous, that two sentences supposedly synonymous could have different semantic evaluations, one true, and the other semantically incorrect. As an example, suppose we had a word, in English, 'greentresting', synonymous with 'either green or interesting', and suppose that the RA of 'greentresting' were a proper subset of that of 'interesting'. Then the sentence 'Model theory is greentresting' would turn out to be true because of the strong inference from 'Model theory is interesting'.[21] Persuaded by such considerations, Martin finally opts for a thoroughgoing weak disjunction rule.

Martin follows the customary link between disjunction and existential quantification in his quantifier rule: an existentially quantified sentence is semantically incorrect if any instance is; otherwise, if there is a true instance it is true, and if not it is false. Left like this, however, the rule diverges considerably from our intuitions. A sentence like 'Some basketballs are round' would turn out to be semantically incorrect if any abstract objects, say, were in the domain, since they would give semantically incorrect instances, like 'Seven is round'. The technical remedy for this is to adopt a segmented domain and restricted quantification, so that the quantifier in a given sentence only ranges over a part of the domain. If the domain is appropriately divided, 'Some basketballs are round' can avoid being semantically incorrect by being interpreted as having a quantifier which only ranges over physical objects, say. An example Thomason gives presents a practical problem for this approach, namely how the domain

---

[21]This discussion follows Martin, SRCP, p. 160.

must be segmented: 'I finished doing one of the things we talked about yesterday'. Conditional on settling such issues, though, the segmented domain works well, and certainly saves the quantifier rule from utter implausibility. It may even be possible to justify the claim that we implicitly fix restrictions on the domain of quantification in ordinary speech, presumably depending on some context of assertion account.

The most problematic valuation rule is that for negation. Since his choice here opens Martin to powerful objections, I shall just briefly outline first some of the options which are apparently open, and then Martin's reason for the choice he makes, and reserve the topic for fuller discussion later. The problem is slightly reminiscent of the appropriate choice for disjunction, since we start by looking at 'Virtue is not triangular' as a predication of 'is-not-triangular' to virtue. Now one way of treating 'is-not-triangular' is as predicating a kind of shape to an object, non-triangularity. In that case, 'is-not-triangular' has as its RA that of all predicates of shape, in particular, that of 'is triangular'. Consequently, since 'Virtue is triangular' is semantically incorrect, so is 'Virtue is not triangular'.

On the other hand, if we think of negation as being more like the operator 'It is not the case that ...', we have a very strong intuition that 'Virtue is not triangular' is actually true, and even that the truth of that claim is a consequence of saying that 'Virtue is triangular' is semantically incorrect. These two intuitions, when combined with the classical notion, represent what is known as choice negation and exclusion negation (or complementation). Their representations in three-valued logic are:

(2.2)

	Choice	Exclusion
A	~A	-A
t	f	f
f	t	t
u	u	t

Martin chooses to reject exclusion negation entirely, saying that it represents a breaking-down of category distinctions. That is, it introduces, for every predicate P, an associated predicate -P whose RA is the whole domain, no matter what the RA of P was. Thus, if we are to maintain the framework in which categories have a role to play, Martin says we must reject exclusion negation.

Thus MW uses a standard first order quantification language L, with a one-place predicate constant T, and the usual formation rules, taking negation , conjunction, and universal quantification as primitive. Its semantics has two distinctive features:[22]

(2.3)

1. To each variable x_i is assigned a sort, s(i), where s is a function from the integers into the integers 1 through k, for some finite k. In addition, the domain D is segmented into k sorts, so that a given variable ranges over only the objects in the sub-domain which its sort assigns to it.

2. The valuation rules follow Kleene's weak three-valued approach.

Given these general remarks, we can give the semantics of MW:

Definition 1: M $(=\langle U_1 \ldots U_k, v\rangle)$ is a model for L iff

1. for every i, $1 \leq i \leq k$, U_i is a non-empty set. Let D (the domain) be $U_1 \cup U_2 \ldots \cup U_k$.

---

[22] My exposition follows Martin and Woodruff closely but not exactly.

2. $\underline{v}$ is a function (the valuation function) such that

   a. if $\underline{a}$ is an individual constant $\underline{v}(\underline{a})\in\underline{D}$.

   b. if $\underline{F}$ is an n-place predicate, $\underline{v}(\underline{F})=\langle\underline{S}_1,\underline{S}_2\rangle$, where $\underline{S}_1$, $\underline{S}_2\subseteq\underline{D}^n$ and $\underline{S}_1\cap\underline{S}_2=\Lambda$.
   $\underline{S}_1$ is called the extension and $\underline{S}_2$ the anti-extension of $\underline{F}$. Let $\underline{S}_1$ be represented by $\underline{vF+}$, $\underline{S}_2$ by $\underline{vF-}$.

Definition 2: α is an assignment iff α is a function such that $\alpha(\underline{x}_i)\in\underline{U}_{s(i)}$.

Definition 3: The value of an expression for a model $\underline{M}$ and assignment α, $\underline{val}_{\alpha,\underline{M}}(\underline{e})$, is the function such that:

1. $\underline{val}_{\alpha,\underline{M}}(\underline{t}_i)$
   $=\alpha(\underline{t}_i)$ if $\underline{t}$ is a variable
   $=\underline{v}(\underline{t})$ if $\underline{t}$ is a name

2. $\underline{val}_{\alpha,\underline{M}}(\underline{Ft}_1\ldots\underline{t}_n)$
   $=t$ iff $\langle\underline{val}_{\alpha,\underline{M}}(\underline{t}_1)\ldots\underline{val}_{\alpha,\underline{M}}(\underline{t}_n)\rangle\in\underline{vF+}$
   $=f$ iff $\langle\underline{val}_{\alpha,\underline{M}}(\underline{t}_1)\ldots\underline{val}_{\alpha,\underline{M}}(\underline{t}_n)\rangle\in\underline{vF-}$
   $=u$ otherwise.

3. $\underline{val}_{\alpha,\underline{M}}(\tilde{}\underline{A})$
   $=t$ iff $\underline{val}_{\alpha,\underline{M}}(\underline{A})=f$
   $=f$ iff $\underline{val}_{\alpha,\underline{M}}(\underline{A})=t$
   $=u$ otherwise

4. $\underline{val}_{\alpha,\underline{M}}(\underline{A}\&\underline{B})$
   $=t$ iff $\underline{val}_{\alpha,\underline{M}}(\underline{A})=\underline{val}_{\alpha,\underline{M}}(\underline{B})=t$
   $=u$ iff $\underline{val}_{\alpha,\underline{M}}(\underline{A})=u$ or $\underline{val}_{\alpha,\underline{M}}(\underline{B})=u$
   $=f$ otherwise

5. $\underline{val}_{\alpha,\underline{M}}((\underline{x}_i)\underline{A})$
   $=t$ iff for all α', $\underline{val}_{\alpha',\underline{M}}(\underline{A})=t$
   where α' is an assignment just like α, except perhaps at $\underline{x}_i$, and $\alpha'(\underline{x}_i)\in\underline{U}_{s(i)}$ $(\alpha'\underset{\underline{x}_i}{=}\alpha)$
   $=u$ iff for some $\alpha'\underset{\underline{x}_i}{=}\alpha$, $\underline{val}_{\alpha',\underline{M}}(\underline{A})=u$
   $=f$ otherwise

Definition 4: A sentence A is true (false) in model M ($V_M(A)=t(f)$) iff $val_{\alpha,M}(A)=t(f)$ for all α, and neither true nor false in model M ($V_M(A)=u$) otherwise.

An important relation in the proof which follows is the notion of a model M' extending a model M:

Definition 5: Model M' ($=\langle U_1'\ldots U_k', v'\rangle$) extends model M ($=\langle U_1\ldots U_k, v\rangle$) iff:

1. $k'=k$
2. $U_i'=U_i$ for all $i\leq k$
3. $v'(a)=v(a)$ for all individual constants a
4. $vF+\subseteq v'F+$ and $vF-\subseteq v'F-$, for all predicates F.

In particular, we are interested in the notion of a T-extension:

Definition 6: M' is a T-extension of M ($M\leq M'$) iff M' extends M and for all predicates F other than T, $vF+=v'F+$, $vF-=v'F-$.

In MW, a model which extends another has a crucial property: every sentence which is true in the "smaller" model is true in the "larger", and similarly for falsity. Thus if we consider a series of models each of which extends its predecessor, the set of sentences which is true grows conservatively: no sentence is true in earlier models and drops out later. We shall use this result in the following form:

Lemma 7: If $M\leq M'$, then for all A, if $V_M(A)=t$ then $V_{M'}(A)=t$, and if $V_M(A)=f$, $V_{M'}(A)=f$.

Proof: By induction on A. The proof is given below, in the chapter on Kripke.

To show the truth-representation result, we need to consider sets of models which share a common feature.

Definition 8: A set of models, S, is a common base set iff

1. any two models M, M'∈S agree everywhere except at T. Let S be complete in that respect.

2. for some j, L=U_j, i.e. one of the segments of the domain comprises the sentences of L.

In a common base set S we can define a subset PR(S):

Definition 9: M∈PR(S) iff for all A,

1. if A∈vT+ then $V_M(A)$=t and

2. if A∈vT- then $V_M(A)$=f, where v∈M.

For models in PR(S) we can say that T *partially represents truth* because every sentence assigned to the extension of T is true in the model, and every sentence assigned to its anti-extension is false. What we want to show is that in some member of S, T is (wholly) representative, that is, that the converse holds as well. The strategy is to apply Zorn's Lemma. To do so we need two lemmas.

Lemma 10: $\lessdot$ partially orders PR(S).

Proof: By definition of $\lessdot$, PR(S).

Lemma 11: Every $\lessdot$-chain in PR(S) has an upper bound in PR(S).

Proof: Let C be a $\lessdot$-chain in PR(S), and let Top be the model defined as follows. Top belongs to S, and the extension of T in Top is the union of the extensions of T in all members of C, and its anti-extension is the union of all the corresponding anti-extensions. Top is clearly an upper bound on C; now we have to show that Top∈PR(S). Suppose A∈vT+, where v is in Top. Then, for some member M of C, A∈v'T+, where v' is in M. But then $V_M(A)$=t, because C$\subseteq$PR(S). However, M$\lessdot$Top, so by 7, $V_{Top}(A)$=t. Similarly for vT-. So Top∈PR(S).

Lemma 12: For any common base set S, and every M in PR(S), there is a maximal model in PR(S).

Proof: By Lemmas 10 and 11, and Zorn's Lemma.

It only remains to be shown that a maximal partially representing interpretation is a wholly representing interpretation.

Lemma 13: If $\underline{M}$ is maximal in $\underline{PR}(\underline{S})$, then $\underline{v}(\underline{T})$ in $\underline{M}$ represents truth in $\underline{M}$.

Proof: Suppose, for *reductio*, $\underline{M}$ is maximal in $\underline{PR}(\underline{S})$, but $\underline{v}(\underline{T})$ in $\underline{M}$ is not truth-representing.

1. Suppose, for some $\underline{A}$, $V_{\underline{M}}(\underline{A})=t$, but $\underline{A}\notin\underline{vT}+$.

   Then we can construct a model $\underline{M}'$, whose valuation is $\underline{v}'$, which coincides everywhere with $\underline{M}$ except that $\underline{A}\in\underline{v}'\underline{T}+$. So $\underline{M}\leqslant\underline{M}'$. Now we show that $\underline{M}'\in\underline{PR}(\underline{S})$:

   a. Suppose $\underline{B}\in\underline{v}'\underline{T}+$. Then either $\underline{B}=\underline{A}$ or $\underline{B}\neq\underline{A}$.
      If $\underline{B}=\underline{A}$, $V_{\underline{M}}(\underline{B})=t$, by the initial hypothesis.
      If $\underline{B}\neq\underline{A}$, $\underline{B}\in\underline{vT}+$, by construction of $\underline{v}'$, hence $V_{\underline{M}}(\underline{B})=t$.
      Either way, $V_{\underline{M}}(\underline{B})=t$, so by Lemma 7 $V_{\underline{M}'}(\underline{B})=t$.

   b. Similarly for $\underline{v}'\underline{T}-$.

   But $\underline{M}$ was supposed to be maximal under $\leqslant$. Consequently there can be no $\underline{A}$ such that $V_{\underline{M}}(\underline{A})=t$, but $\underline{A}\notin\underline{vT}+$.

2. Suppose instead $V_{\underline{M}}(\underline{A})=f$, but $\underline{A}\notin\underline{vT}-$.

   By a similar argument, this is impossible.

So $\underline{v}(\underline{T})$ in $\underline{M}$ represents truth in $\underline{M}$.

The concluding theorem is:

Theorem 14: If $\underline{M}$ is a model, then there is a model $\underline{M}'$ which coincides with $\underline{M}$ on all sentences not containing $\underline{T}$, and whose truth predicate represents truth.

Proof: Let $\underline{M}''$ be the model which agrees with $\underline{M}$ everywhere but at $\underline{T}$, where $\underline{M}''$ has $\langle\mathbb{A},\mathbb{A}\rangle$. $\underline{M}$ and $\underline{M}''$ belong to the same common base set $\underline{S}$, and $\underline{M}''$ is in $\underline{PR}(\underline{S})$. Consequently there is a model $\underline{M}'$ which is maximal in $\underline{PR}(\underline{S})$ and has a representative truth-predicate. Since $\underline{M}$ and $\underline{M}'$ are both in $\underline{S}$, they agree on all sentences not containing $\underline{T}$, by the standard local determination lemma.

Given this result, we can see that a language can contain its own

truth predicate even when there are paradoxical sentences present. Since many people thought Tarski had shown that no language could contain its own truth predicate, the result was a surprising one. More interesting is the fact that it can be shown that a language using weak three-valued tables can include representative falsity and neither-truth-nor-falsity predicates.[23] The truth-representation theorem can be established for stronger connectives, but not this latter theorem. However, I shall discuss these details when looking at Kripke's theory.

It is interesting to see how MW, extended to include 'neither true nor false', treats Liar sentences and their cohorts. Since MW uses only choice negation, there is no difference between the treatments of 'is false' and 'is not true'. However, some of the difference between 'This sentence is false' and 'This sentence is not true' might seem reproduced by considering 'This sentence is either false or neither true nor false' as our Strengthened Liar. Suppose, then, that the valuation function specifies the following denotations for the sentences $\underline{s}_1$, $\underline{s}_2$ and $\underline{s}_3$:

$$\begin{aligned} &\underline{v}(\underline{s}_1)=\underline{Ts}_1 \\ &\underline{v}(\underline{s}_2)=\tilde{}\underline{Ts}_2 \\ &\underline{v}(\underline{s}_3)=\tilde{}\underline{Ts}_3 \vee \underline{Ns}_3 \end{aligned} \qquad (2.4)$$

Since all the truth-representing models whose existence we have proved are maximal, $\underline{s}_1$ must always be in either the extension or anti-extension of $\underline{T}$: if in some model it were not, that model would not be maximal. So $\underline{s}_1$ is either true or false; Martin points out that this models the Truth-Teller's easy-going nature.

With $\underline{s}_2$, however, we soon see that it cannot be put in either the extension or anti-extension of $\underline{T}$: suppose $\underline{M}'$ is the representing model,

---

[23] This was established independently by Martin and Gupta, who have jointly written a paper "A Fixed Point Theorem for the Weak Kleene Valuation Scheme".

and $\underline{v}'$ the corresponding valuation. Then if $\underline{s}_2 \in \underline{v}'\underline{T}^+$, $V_{\underline{M}'}(\underline{Ts}_2)=t$ by ordinary valuation rules, but also $V_{\underline{M}'}(\tilde{}\underline{Ts}_2)=t$, by the truth-representation claim, and similarly if $\underline{s}_2 \in \underline{v}'\underline{T}^-$. However, $\underline{s}_2$ can happily belong to the extension of $\underline{N}$. So $\underline{s}_2$ is neither true nor false in any truth-representing model.

The same result holds for $\underline{s}_3$, though the argument is longer. Since $\underline{s}_3$ is a sentence, either $\underline{s}_3 \in \underline{v}'\underline{N}^+$, or $\underline{s}_3 \in \underline{v}'\underline{N}^-$. Suppose the latter. Then $V_{\underline{M}'}(\underline{Ns}_3)=f$, and either $\underline{s}_3 \in \underline{v}'\underline{T}^+$ or $\underline{s}_3 \in \underline{v}'\underline{T}^-$ (if it is not neither true nor false, it must be either in the extension of $\underline{T}$ or its anti-extension). If the former, $V_{\underline{M}'}(\tilde{}\underline{Ts}_3 \vee \underline{Ns}_3)=t$, by the representation fact, and hence $\underline{s}_3 \in \underline{v}'\underline{T}^-$, by the valuation rules for $\underline{v}$. But similarly, if $\underline{s}_3 \in \underline{v}'\underline{T}^-$, $V_{\underline{M}'}(\tilde{}\underline{Ts}_3)=t$, and since $V_{\underline{M}'}(\underline{Ns}_3)=f$, $V_{\underline{M}'}(\tilde{}\underline{Ts}_3 \vee \underline{Ns}_3)=t$, so $\underline{s}_3 \in \underline{v}'\underline{T}^+$. However, no contradictions result if $\underline{s}_3 \in \underline{v}'\underline{N}^+$.

Thus in many respects MW handles the paradoxes quite neatly. Given Martin's defense of his category approach, do we not have sufficient grounds to regard this as a satisfactory solution of the paradoxes? Several strong objections have been lodged against the account which must be considered. Actually, these objections are strongly interconnected, so that picking up one tends to lead on to another: they focus on problems with negation, the RA of 'is true' and 'is false', the treatment of the Strengthened Liar, and general considerations about categories.

It is perhaps easiest to start with negation and the Strengthened Liar. As I mentioned above, there are strong intuitions that 'Virtue is not triangular' is ambiguous: in one sense, it attributes a shape property to virtue, and is semantically incorrect, and in the other it denies that virtue is triangular, and seems to assert a truth. These two senses can be expressed just by introducing a distinction between two senses of 'not', as I described above. However, another method is also available, which does not impute lexical ambiguity to 'not', but

rather a structural ambiguity to 'Virtue is not triangular'. In this method, negation is pure choice negation, and the exclusion sense is capture by treating 'Virtue is not triangular' as '"Virtue is triangular" is not true'.

Martin takes neither of these options. If we accept exclusion negation, we can no longer establish the existence of a truth-representing model, as can be quickly seen. Let $\underline{v}(\underline{s}_4)=-\underline{Ts}_4$. In any model $V_{\underline{M}}(-\underline{Ts}_4)=f$ or $V_{\underline{M}}(-\underline{Ts}_4)=t$, since $-\underline{A}$ only has t or f in its column. But then $\underline{s}_4$ must either belong to $\underline{vT}^+$ or $\underline{Vt}^-$ if $\underline{T}$ is to represent truth. But by familiar reasoning, if it is in one, it must be in the other. So $\underline{T}$ cannot represent truth. Intuitively, it seems that exclusion negation is what we need to express the Strengthened Liar, so the rejection of exclusion negation creates a concomitant objection that the Strengthened Liar is not properly expressible in MW.

If, on the other hand, we look at the paraphrase technique, we see that '"Virtue is triangular"is not true' can only represent the sense in which 'Virtue is not triangular' is true if 'is not true' can truly be predicable of a semantically incorrect sentence. Since in this account we are only using choice negation, this amounts to the requirement that the RA of 'is true' and 'is false' be taken to include sentences which are semantically incorrect. Permitting that, in this context, leads straight back into paradox, for if even semantically incorrect sentences can be true or false, we cannot ward off the contradiction by claiming that it illicitly presupposes that the Liar sentence is semantically correct.

Thus at the very least, the expense to our intuitions about numerous sentences like 'Seven is not red', 'The Elgin Marbles are not speaking tonight' and so forth is heavy, if we adopt MW. However, Martin's response to this objection is that our belief that these sentences are true is based on ignoring category considerations. It is

important to understand one consequence of Martin's methodological approach. This is that although the assumptions which give rise to paradox are plausible, recognizing the appropriate rulelike feature of language casts doubt on these assumptions, even shows them to be false. It follows that once we recognize that rulelike feature, we may be led to reject numerous intuitions we formerly held. Thus in a way, Martin is offering a proposal for a reform of language. The difference between his approach and the radical Quinean one is that the reform is motivated not by the fact that our intuitions lead us to contradiction and are dispensable, but by establishing a part of those intuitions as paramount in formation of the revised language. That the revision has the consequence of eliminating the paradox is, so to speak, a happy coincidence.

It follows from this that the mere fact that the theory diverges from intuition is not a decisive objection against it. Thus Martin's own objection against the hierarchical approach, that it led to a ban on self-reference, was followed by the claim that a category theory could provide at least as good an account at less cost. Moreover, it does not seem entirely implausible that we could account for our belief in the truth of these banned sentences by treating them as metaphorical, i.e. category-distorting, uses of 'not' and 'true'. Thus though strictly speaking they would be category mistakes, there would be an extended sense in which they could be understood.

A more serious objection is brought by Donnellan, who draws attention to a fact which he finds as puzzling as the paradoxes themselves.[24] This is that in Martin's early accounts, but also in MW, there are three mutually exclusive categories, namely truth, falsity,

---

[24]Keith S. Donnellan, "Categories, Negation, and the Liar Paradox". My discussion of these topics rests partly on Donnellan's clear and useful analysis.

and neither. Sentences which are true or false can be said not to be of the third category, but those which are neither cannot be said not to be true or false, or at least cannot be said truly. Of course, this point is connected to the previous ones because a different treatment of negation or the RA of 'is true' would permit us to do that very thing.

Martin's response to this is to say that it ignores the category difference between 'true' and 'false' on the one hand, and 'neither' on the other, though he has some difficulty in reconciling this claim with the semantics of MW. The point is this. We want to be able to say, what is true, that any sentence s is either true, false or neither. However, we cannot express that by

$$Ts \vee Fs \vee Ns \qquad (2.5)$$

because if s is neither, both Ts and Fs are neither too, so the disjunction is neither, by the valuation rule for $\vee$. However, Martin has a device to account for this, at least informally: he suggests that if we have a disjunction in which all the predicates which exhaust a category are applied to one and the same objects, we can treat that as a predication to that object of the covering category predicate. That is, if blue, yellow and red exhausted the category of colour, saying the moon is blue, yellow or red is just like saying that it is coloured. Since the predicates in the first two disjuncts in (2.5) exhaust the category of the semantically correct, the true sentence which corresponds to it is actually:

$$\sim Ns \vee Ns \qquad (2.6)$$

Thus Martin thinks that he can account for the intuition behind (2.6) while blocking any argument from Ns to $\sim$Ts. Of course this cannot be embodied directly in the language MW, so the situation is something of a standoff. The point to be made against Donnellan's objection is that the inference could be made if true, false, and neither were joint

category mates, but since in Martin's theory they are not, the air of paradox is lessened. In fact, the situation has an exact analogy with colour predicates: given the simplifying assumption about blue, yellow and red, and the supposition that 'is coloured' is applicable to anything, it is true that anything is blue, yellow, red, or not coloured, but it does not follow from the fact that something is not coloured that it is not red.

Thus Martin evades the whole force of all these objections by reiterating that the category approach demands certain sacrifices, which we must accept as long as no better alternative presents itself. However, I want to argue that these and similar objections constitute an overwhelming argument for rejecting Martin's account if they are advanced in the right way. If Martin is going to hold that we can regard the Liar as disarmed by his theory, despite its counter-intuitive consequences, because it is based on a principled discussion of categories and semantic incorrectness, that discussion cannot be allowed to settle questions about categories by any reference to the Liar, without opening itself up to charges of being ad hoc. I want to argue that Martin's approach to categories is distorted by the attempt to handle the Liar, and that conversely some of the arguments about how 'true' works are distorted by category considerations. Thus Martin cannot claim to have an independently motivated theory of the categories which happily solves the problem of the Liar, nor can he claim to have a useful theory of truth.

I shall start by returning to the debate about negation and the category of 'is true'. Martin's position is that the intuition that 'Virtue is not triangular' is true is mistaken, through ignoring category distinctions. I shall argue that this position cannot be defended by pure category considerations, and that its justification only comes from the desire to avoid paradox.

The starting point for any theory of categories must be, as usual, some collection of data about what sentences are accepted or rejected, plus various intuitions about the notion, such as the intuition that in hierarchical category structures, the RA of a predicate at one level is the extension of the predicate at the next more general level. Thus 'is yellow' is only correctly predicable of those things that 'is coloured' is true of, for example. These data and intuitions are then put to work in the usual way of theory construction, one being rejected here, another there.

What then, are the data we are interested in for negation and truth? I suggest that 'Virtue is triangular' and 'Virtue is not triangular' and '"Virtue is triangular" is not true' are respectively rejected; sometimes rejected, sometimes accepted; and accepted, and that we have to explain these facts. Two alternative explanations have been offered in criticism of Martin's position, namely that 'not' should be treated ambiguously, or that 'a is not ◆' should be: together with the accepted semantic incorrectness of 'Virtue is triangular,' these explain the ambivalence over 'Virtue is not triangular', though both also predict ambivalence over '"Virtue is triangular" is true', unless the RA of 'is true' is taken to include semantically incorrect sentences. If we are going to do that, however, we have no need of exclusion negation as well, as explained before. So one explanation of the data is that negation is choice, and the RA of 'is true' is all sentences, or perhaps all grammatical sentences.

Martin rejects this position because it leads to a breakdown of semantic categories, but this explanation alone is not satisfactory. There is nothing incoherent in the notion of a language which employs semantic categories, but also has the resources for cross-category assertions. Nothing that Martin has said should be taken as an argument that a language must be rigorously categorical. So there has to be an

argument, based on intuitions about categories, that at least 'is true' should not be the device that allows us to cross categories, and Martin in fact offered an argument for taking the RA of 'is true' to be just the semantically correct: first, we regard 'is true' and 'is false' to be category mates, having the same RA, and second, sentences which are semantically incoherent are neither true nor false. From this, Martin argued that the RA of either 'is true' or 'is false' is just the semantically correct.

However, this does not follow directly, without various other assumptions. We can accept both the claims just given, and deny the conclusion. The crucial assumption is that 'semantically correct' hierarchically dominates both 'is true' and 'is false'. If this were clearly so, then the yellow/coloured example would be strongly suggestive here: 'true' and 'false' would be properly predicable of just those things of which 'semantically correct' is truly predicable. But this assumption is not self-evident. For instance, there are sentences which seem to be semantically correct, but neither true nor false, such as 'The present King of France is bald'. This may just suggest that 'semantically correct' dominates a three-way split 'true', 'false' and 'neither true nor false', each only predicable of the semantically correct, but this would be altogether inappropriate, since we could no longer say that semantically incorrect sentences are neither true nor false. In conclusion, it appears that although there is a very close connection between the fact that 'Virtue is triangular' is semantically incorrect and the fact that it is not true, it cannot confidently be claimed that 'is semantically correct' dominates 'is true'.

Furthermore, the intuition that '"Virtue is triangular" is not true' is true is so strong, and indeed seems to be a consequence of adopting a category approach, that it is hard to see why a defender of

the category approach would wish to deny it: in Martin's case, the reason is that doing so gives an account of the Liar.

This situation is rather different for disjunction. Here, we have good reasons on sortal grounds for adopting Weak disjunction as the rule. The problem is that it is extremely implausible that this is the right rule for truth. Semantic incorrectness, like meaninglessness, is heavily infectious: if a part of a sentence has it, the whole has it. On the other hand, typically disjunctive inference goes by the Strong rule.

In one interesting attempt, Martin suggested combining the two rules, in "A Category Solution to the Liar". Essentially the idea was that there is a difference between a complex predication to a single object, where the Weak rule is appropriate, and a regular disjunction, where the Strong one is. Technically this was achieved by supervaluations over a restricted range of classical valuations, and the effect was to prevent the inference above from $\underline{Ns}_3$ to $\underline{\sim Ts}_3 \vee \underline{Ns}_3$, where $\underline{s}_3$ is the Strengthened Liar of (2.4). Since, given the interpretation of $\underline{\sim A}$ and the RA's of $\underline{N}$ and $\underline{T}$, $\underline{\sim Ts}_3$ is a category mistake if $\underline{Ns}_3$ is true, permitting such an inference would be to permit the inference of 'Virtue is triangular or unattainable' from 'Virtue is unattainable', which Martin holds is impermissible. The theory does permit the inference from 'Virtue is unattainable' to 'Virtue is unattainable or seven is green', however. Unfortunately, a more complicated paradox escapes the restriction. Suppose we have sentences $\underline{s}_5$ and $\underline{s}_6$:

$$\underline{v}(\underline{s}_5) = \underline{Fs}_5 \vee \underline{Ns}_6 \qquad (2.7)$$
$$\underline{v}(\underline{s}_6) = \underline{Ts}_5$$

Suppose $\underline{s}_6$ is neither. Then $\underline{s}_6$ is true, and $\underline{Fs}_5 \vee \underline{Ns}_6$ is permitted by the strong disjunction rule: trouble will clearly result. If $\underline{s}_6$ is either true or false, then in the first case $\underline{s}_5$ must also be true, which contradicts $\underline{Fs}_5 \vee \underline{Ns}_6$, and in the second case $\underline{s}_5$ must be false, and another permissible strong inference gives $\underline{Fs}_5 \vee \underline{Ns}_6$, and trouble again.

So in the end, the only option, apart from abandoning the project, is to adopt weak disjunction. However, as I said, weak disjunction does have some category arguments in its favour: it just cannot handle truth correctly.

These problems are exaggerated when we turn to quantification. First, it seems, unsurprisingly by now, that truth is not governed by the Weak quantification rule. Second, and more surprisingly, it seems that semantic correctness is not either.

The first point is easily established. Nobody can now truly say either 'Something a Cretan once said is false' or 'Something a Cretan once said is true' because Epimenides made his fateful remark. Yet obviously we might truly say both, and other such examples multiply beyond end.

As for the second point, consider the sentence 'Some creatures can see colours'. My intuition is that that is true and semantically correct, but that 'The bacteria in my drains can see colours' is bizarre. But any segmentation of the domain which allows the quantifier in the first sentence its full range will result in its being ruled semantically incorrect because of the bacteria in the drain. This is like Thomason's example 'I finished doing one of the things we talked about yesterday', but avoids the problem of the questionable sortal 'thing'. More immediately to our present interests, suppose we accept all of Martin's theory about atomic sentences and simple truth predications. It still seems that 'Something a Cretan once said is true' is semantically correct, whatever a Cretan may have said. The point is that what Cretans have said may make problems for evaluating the sentence as true or false, but we should not have to investigate these questions to decide whether our own assertions are proper or not. This would threaten once more to make semantic correctness too much a matter of fact, and not enough a matter of meaning.

It is interesting again to compare semantic correctness with meaningfulness. For meaningfulness, as I suggested, it seems plausible that a meaningless part of a sentence gives meaninglessness to the whole. However, it is extremely unlikely that meaninglessness of an instance infects a quantified sentence. 'Fred said something true' does not become meaningless if Fred happened to say 'Doo-be-doo-be-doo', for example!

Consequently it seems sensible to reject the normally useful parallel between disjunction and existential quantification. Semantic correctness may be governed by a Weak disjunction rule, but not by the corresponding quantifier rule. However, if we do this, paradox looms again, as Kripke's quantified examples show. If Smith and Jones say

Jones: Smith sometimes tells the truth about me. (2.8)
Smith: Jones always lies about me.

and otherwise each has always lied about the other, contradiction readily follows.

So once again, Martin only avoids paradox by using a rule which has little or no justification on category grounds. Thus finally we must reject Martin's theory as giving either of two kinds of solution to the paradoxes. First, he does not give us an independently justified theory of sortal correctness in which the paradoxes can be resolved, and second, he does not give us a plausible theory about 'is true' which could illuminate the problem either.

Chapter 3

Van Fraassen: Supervaluations and Presupposition

In his discussion of the Liar[25] , van Fraassen seems to apply a similar methodology to Martin's; I say 'seems to' here because van Fraassen's methodological remarks are rather less explicit than Martin's. In "Truth and Paradoxical Consequences", he suggests that a solution to the paradoxes comprises two parts: first, "an analysis of the logically relevant features of the paradoxes as stated in natural language", and second, "a formal construction in which corresponding sentences play roles roughly similar to those which our analysis ascribes to the paradoxical statements" (TPC, p. 13). He later goes on to say, after a discussion of the Epimenides version, "This is a solution, in the sense of an analysis which credits the Epimenidean statement with all the logical features that lead to the subordinate conclusions in the paradox, but blocks the final derivation of absurdity on general grounds concerning the structure of language" (TPC, p. 16).

I take this to be a contrastive characterization of a solution, contrasting, that is, with any proposal for a solution in which the paradoxical sentence does not have "all the logical features that lead

---

[25]The principal writings on the Liar are Bas van Fraassen, "Presupposition, Implication, and Self-Reference" (hereafter PIS-R), and "Truth and Paradoxical Consequences" (TPC). Some relevant details are also discussed in "Presuppositions, Supervaluations and Free Logic", "Inference and Self-Reference", and "Singular Terms, Truth-Value Gaps, and Free Logic".

to the subordinate conclusions in the paradox". I take it, for example, that a 'solution' based on banning self-reference would not satisfy this characterization because we would not credit the Liar sentence with generating either of the sentences 'If it is true, then it is false' and 'If it is false, then it is true'. As I understand van Fraassen, he is saying that his analysis allows these claims, but blocks the final absurdity by pointing to "general grounds concerning the structure of language" which enable us to say that it is in fact neither true nor false. In van Fraassen's case, the relevant feature of language is the semantic relation of presupposition between sentences, and I am going to suppose, just as with Martin, that it is crucial to the success of van Fraassen's solution of the Liar that his analysis of presupposition is both plausible and independent of the Liar.

Van Fraassen follows Strawson[26], among others, in taking presupposition to be a semantic relation between sentences such that, if one of the presuppositions of a given sentence fails to hold, then the sentence is neither true nor false. Thus, if 'There is a present King of France' is a presupposition of 'The present King of France is bald', then the latter sentence is neither true nor false because the former is false. So:

A presupposes B iff A is true or false only if B is true. (3.1)

If we introduce two pieces of notation, we can simplify this characterization. The first is that we are going to use choice negation, the second that we are going to express the relation between A and B that holds if, whenever A is true, B is true, by $A \mid\!= B$; this will be called semantic entailment or necessitation. Using these devices, (3.1) can be written as:

---

[26]Peter F. Strawson, Introduction to Logical Theory and "On Referring".

Definition 1: A presupposes B iff

1. $A \models B$ and
2. $\sim A \models B$

In a bivalent language, either A or $\sim A$ is always true, so that the only sentences which can be presupposed are ones that are always true, i.e. the logically true. Furthermore, presupposition becomes just a species of implication. If we follow Strawson, though, in allowing sentences with false presuppositions to be neither true nor false, we clearly need a three-valued language, and can then distinguish implication and presupposition. Consider:

$$\text{a)}\ \frac{A > B \quad A}{\therefore B} \qquad \text{b)}\ \frac{A > B \quad \sim B}{\therefore \sim A} \qquad \text{c)}\ \frac{A > B \quad \sim A}{\therefore B} \qquad (3.2)$$

If we take the corner as representing implication, a) and b) are familiarly valid inference forms, and c) is not, whereas if we take it to be presupposition, a) and c) are valid, but b) is not: if A presupposes B and B is false, A is neither true nor false, not false.

In order to construct a language in which presupposition can be represented, van Fraassen uses the notion of a supervaluation. In a classical valuation, every sentence is awarded one of the values t or f. But in a language in which we want to express presupposition, classical valuations go too far in assigning t or f to everything: a valuation which assigns f to 'There is a present King of France' and either t or f to 'The present King of France is bald' is not an appropriate valuation. So a supervaluation is generated by a set of classical valuations: it agrees with them when they all agree, and gives u (for undecided) where they differ. In the case just given, the two valuations which assign f to 'There is a present King of France' and variously t or f to 'The present King of France is bald' generate a supervaluation in which the former still gets f, and the latter gets u.

More precisely, in a propositional language L with just negation and disjunction;

Definition 2: v is a *classical valuation* of L iff v is a function such that

1. $v(A) \in \{t,f\}$.
2. $v(\sim A) = f$ iff $v(A) = t$.
3. $v(A \vee B) = f$ iff $v(A) = v(B) = f$.

Definition 3: s_k is the *supervaluation induced by* k iff

1. k is a set of classical valuations of L.
2. $s_k(A) = t$ iff for all $v \in k$, $v(A) = t$,
   $= f$ iff for all $v \in k$, $v(A) = f$;
   $= u$ otherwise.

The question now is which supervaluations are the ones we are interested in when trying to represent presupposition. Since presupposition is a semantic relation not captured by classical valuations, it has to be specified separately. Thus part of the semantics of a language is a relation N, the (non-classical) *necessitation* relation. N is just a set of ordered pairs of sentences: if we want it to represent presupposition, we must include in N the pair $\langle A,B\rangle$ and $\langle \sim A,B\rangle$, whenever A presupposes B. If we think of the set of sentences which are all true in a given situation, we can see that it must include all the sentences which are classically entailed by any true sentences, and those that are non-classically necessitated by them, for if 'The present King of France is bald' is true, we know that 'There is a present King of France' is too.

To define this set, then, we need the following:

Definition 4: An ordered pair $\langle V_0, N\rangle$ is a *presuppositional semantics for a language* L iff

1. $\underline{V_0}$ is a set of classical valuations of $\underline{L}$.

2. $\underline{N}$ is a relation whose field is a subset of sentences of $\underline{L}$.

Definition 5:

1. Valuation $\underline{v}$ *satisfies* sentence $\underline{A}$ iff $\underline{v}(\underline{A}) = t$.

2. $\underline{v}$ satisfies a set of sentences $\underline{X}$ iff $\underline{v}$ satisfies every member of $\underline{X}$.

3. $\underline{X}$ *classically entails* $\underline{A}$ ($\underline{X} \models_C \underline{A}$) iff every valuation which satisfies $\underline{X}$ satisfies $\underline{A}$.

Definition 6: A set of sentences $\underline{G}$ is *saturated with respect to* $\langle \underline{V_0}, \underline{N} \rangle$ iff:

1. there is a valuation $\underline{v} \in \underline{V_0}$ which satisfies $\underline{G}$.

2. $\underline{G}$ is closed under classical entailment.

3. $\underline{G}$ is closed under $\underline{N}$.

Definition 7: A supervaluation $\underline{s_k}$ is an *admissible valuation for a presuppositional semantics* $\langle \underline{V_0}, \underline{N} \rangle$ iff for some saturated set $\underline{G}$, $\underline{v} \in \underline{k}$ iff $\underline{v}$ satisfies $\underline{G}$.

We can now define supervaluational properties, parallel to Definition 5:

Definition 8:

1. Admissible valuation $\underline{s_k}$ *supervaluationally satisfies sentence* $\underline{A}$ *for presuppositional semantics* $\langle \underline{V_0}, \underline{N} \rangle$ iff $\underline{s_k}(\underline{A}) = t$.

2. $\underline{s_k}$ *supervaluationally satisfies set* $\underline{X}$ *for* $\langle \underline{V_0}, \underline{N} \rangle$ iff it supervaluationally satisfies every member of $\underline{X}$.

3. $\underline{X}$ *supervaluationally entails* $\underline{A}$ ($\underline{X} \models_{sup} \underline{A}$) iff every admissible valuation that supervaluationally satisfies $\underline{X}$ supervaluationally satisfies $\underline{A}$.

This use of $\langle \underline{V_0}, \underline{N} \rangle$ as a presuppositional semantics is an odd one in

some respects. There are two alternative ways to introduce supervaluations from scratch, both of which are theoretically simpler than this method. The first is merely to specify a set of necessitations, $\underline{N}$, and define a consistent saturated set under $\underline{N}$ alone, and then use the saturated set to define a supervaluation directly: a sentence is true if it is in the set, false if its negation is, and neither otherwise.[27] The other method is to specify a subset of the powerset of the classical valuations of a language: each member will define a supervaluation by Definition 3.

The method van Fraassen uses is neither flesh nor fowl, but it has certain advantages. First, presupposition, being a relation between sentences, is most readily represented using the relation $\underline{N}$, rather than as a direct constraint on sets of valuations. On the other hand, the supervaluations are partial valuations, and it is useful to be able to compare the initial set of valuations with the resultant set of supervaluations, whence the mixed mode of Definition 6, where both satisfaction and closure under $\underline{N}$ are used.

This definition of admissible valuation corresponds to the so-called "radical policy" for presuppositional languages. For some purposes, this admits some valuations which seem excessively "gappy": that is, these valuations give u to some sentences which do not suffer from failure of presupposition. For instance, suppose the language $\underline{L}_o$ contains only the sentences $\underline{A}$,$\underline{B}$, and $\underline{C}$:

$\underline{A}$: There is a present King of France. (3.3)

$\underline{B}$: The present King of France is bald.

$\underline{C}$: Roses are red.

Suppose further that $\underline{N}$ is the plausible $\underline{N}_0 = \{\langle \underline{B},\underline{A}\rangle,\langle \tilde{\ } \underline{B},\underline{A}\rangle\}$. Then one

---

[27] See also Herzberger's "Canonical Superlanguages."

saturated set is the logical closure of $\{A,B\}$. This generates a supervaluation in which C is given u, even though C does not have any presuppositions. If we want a theory which gives u to all and only those sentences whose presuppositions fail, this is of course an unsatisfactory situation, and there is some discussion in the literature on strategies for dealing with it.[28] However, I shall not go into this problem, as its consequences for treatment of the Liar are not immediately serious, and shall continue to use the above definition of admissible valuation.

Obviously from the definition of admissible valuation, there is a close connection between a supervaluation and the saturated set of sentences which engender it. Since it is often easier to discuss the membership of sentences in the saturated set, rather than the supervaluation, the following theorem is convenient:

Theorem 9: When s_k, k, and G are as defined above,

1. $s_k(A) = t$ iff $A \in G$.
2. $s_k(A) = f$ iff $\sim A \in G$.
3. $s_k(A) = u$ iff neither $A \in G$ nor $\sim A \in G$.

Proof:

1. Suppose $s_k(A) = t$: then for all $v \in k$, $v(A) = t$. But $v \in k$ only if v satisfies G. So every v that satisfies G satisfies A, i.e. $G \models_C A$. But G is closed under classical entailment. So A is in G.
   Suppose $A \in G$. $v(A) = t$ for all $v \in k$. Hence $s_k(A) = t$.
2. Suppose $s_k(A) = f$. Then, for all $v \in k$, $v(A) = f$ and $v(\sim A) = t$. Hence, as above, $\sim A \in G$.

---

[28] See van Fraassen "Presuppositions, Supervaluations and Free Logic", Hans Herzberger, "Presuppositional Policies", and T. P. Lightbody, "An Examination of Two Recent Approaches to the Liar Paradox", ch. 1.

Suppose $\sim A \in G$. Then $v(A) = t$ for all $v \in k$, and $v(A) = f$. So $s_k(A) = f$.

3. Suppose $s_k(A) = u$. Then $s_k(A) \neq t$ and $s_k(A) \neq f$. By a) and b), $A \notin G$ and $\sim A \notin G$. Similarly for the contrapositive.

Before going on to consider how to add truth to a presuppositional language, some other features are worth pointing out. First, since a supervaluation represents what various classical valuations have in common, any sentence which is either true, or false, in every classical valuation will be so in any supervaluation. Hence the laws of classical propositional logic remain true in a supervaluational language. In particular, the law of the excluded middle holds; for any sentence A, $A \vee \sim A$ is true in every admissible valuation, even when A itself is undefined in some of those valuations. This curious feature of supervaluations can create some problems of interpretation, accustomed as we are to inferring that one or other disjunct of a true disjunction is true. Supervaluations also retain the characteristic of strong three-valued tables, that a disjunction with a true disjunction is true, whatever the value of the other disjunct. To see this, consider a supervaluation s_k for which $s_k(A) = t$. Then in every $v \in k$, $v(A) = t$. By classical rules, then, $v(A \vee B) = t$, and hence $s_k(A \vee B) = t$ also.

So much for the relationship of supervaluational languages to other logics. In addition, there are some properties characteristic of presuppositions, which will become important when considering the Liar. In particular, the following results are readily obtained.

Theorem 10:

1. If $\langle A, B \rangle \in N$, then $\{A\} \models_{sup} B$.

Proof: If A is true in s_k, then $A \in G$. But if $A \in G$, so does B, by closure under N. Hence B is true in s_k. So $\{A\} \models_{sup} B$.

2. If $\{\underline{A}\} \models_c \underline{B}$, $\{\underline{A}\} \models_{sup} \underline{B}$.

Proof: Similarly, by closure under classical entailment.

3. If $\underline{A}$ presupposes $\underline{A}$, $\underline{A}$ cannot be false.

Proof: Suppose $\underline{A}$ is false: since $\underline{A}$ presupposes $\underline{A}$, $\underline{A}$ must be true. So $\underline{A}$ cannot be false.

4. If $\underline{A}$ presupposes $\sim\underline{A}$, $\underline{A}$ cannot be true. Similarly.

5. If $\underline{A}$ presupposes a contradiction, $\underline{A}$ is neither true nor false. Similarly.

Henceforth, all entailments will be supervaluational, and I shall omit the subscripts on $\models$. I shall also informally treat entailment as a relation between sentences.

The next question to be faced is how to introduce truth ascriptions into a propositional language. Doing this involves making decisions about a familiar nexus of problems concerning negation, bivalence, Tarski biconditionals and so forth. For simplicity's sake I shall follow van Fraassen in starting with only a truth operator, rather than a truth predicate, and will look at his theory which uses a predicate later. Thus, we have the additional monadic operator $\underline{T}$, whose intended interpretation is "It is true that...". The next question is how to express "It is false that...". The answer to this depends on two preliminary features of truth and choice negation. First, independent of any further decisions about the truth operator, if $\underline{A}$ is true, so is $\underline{TA}$, and vice versa. Second, $\sim\underline{A}$ is true just in case $\underline{A}$ is false. Consequently $\underline{T}\sim\underline{A}$ is true when $\underline{A}$ is false, and vice versa. So "It is false that $\underline{A}$" can be rendered by $\underline{T}\sim\underline{A}$. This apparently means that the Principle of Bivalence can be expressed by $\underline{TA} \vee \underline{T}\sim\underline{A}$, which is not, unlike $\underline{A} \vee \sim\underline{A}$, a logical truth. It is, however, a consequence of the Tarski biconditional:

1. $A \supset TA$	Tarski	(3.4)
2. $\sim A \supset T \sim A$	Tarski	
3. $\sim TA \supset \sim A$	Contraposition 1	
4. $\sim TA \supset T \sim A$	Transitivity, 2 and 3	
5. $\sim\sim TA \vee T \sim A$	Def $\supset$, 4	
6. $TA \vee T \sim A$	Double negation elim. 5	

Van Fraassen considers three courses of action: we interpret $TA \vee T \sim A$ differently; we reject the logic embodied in 1-6; or we reject the conditionals 1) and 2). He elects the third course: the grounds for retaining standard logic if at all possible do not need rehearsing, but the grounds van Fraassen gives for rejecting the first course are interesting, since they reflect a certain attitude towards truth and the interpretation of sentences in a supervaluation language.

The possibility of reinterpreting $TA \vee T \sim A$ in such a way that it is true even when A is neither true not false rests on the fact mentioned above, that in supervaluational languages, disjunction of sentences which are neither true nor false can turn out to be true, as for example $A \vee \sim A$. In order to take advantage of this opportunity to reinterpret $TA \vee T \sim A$, we might adopt the policy that if A is undefined, so is TA, and vice versa. This would certainly do what was required here, but it would also leave various necessary things undone. The problem is that we could no longer truly say of a sentence which is neither true nor false, that it is not true. For if we tried to do so by asserting $\sim TA$, we would only succeed in saying something which is itself undefined. Van Fraassen elects instead to reject the Tarski biconditionals and adopt as semantics for the truth operator: if A is true, so is TA; otherwise TA is false. This has the advantage of making truth ascriptions bivalent, and gives us a global truth operator, but eliminating the biconditionals leaves us with the problem of how to introduce the operator into L.

Van Fraassen's solution is to expand $\underline{N}$ in such a way that it handles not only presupposition relations but also the relation between $\underline{A}$ and $\underline{TA}$. He suggests that this can be done by including, for every sentence $\underline{A}$, $\langle \underline{A}, \underline{TA} \rangle$ and $\langle \underline{TA}, \underline{A} \rangle$ in $\underline{N}$. However, this alone will not accommodate all the characteristics of truth that van Fraassen has desiderated. To see this, we can consider all the classical valuations associated with a sentence $\underline{A}$; since no feature of $\underline{T}$ is defined at the level of classical valuation, both $\underline{TA}$ and $\underline{T}\tilde{}\underline{A}$ count as atomic sentences. Hence there are the familiar eight possibilities:

	$\underline{A}$	$\underline{TA}$	$\underline{T}\tilde{}\underline{A}$	(3.5)
$\underline{v}_1$	t	t	t	
$\underline{v}_2$	t	t	f	
$\underline{v}_3$	t	f	t	
$\underline{v}_4$	t	f	f	
$\underline{v}_5$	f	t	t	
$\underline{v}_6$	f	t	f	
$\underline{v}_7$	f	f	t	
$\underline{v}_8$	f	f	f	

Let us first consider the situation when $\underline{A}$ is true in a supervaluation, and thus belongs to $\underline{G}$. Since $\langle \underline{A}, \underline{TA} \rangle \in \underline{N}$, $\underline{TA} \in \underline{G}$. Hence the only members of $\underline{k}$ are $\underline{v}_1$ and $\underline{v}_2$. But in these two valuations $\underline{T}\tilde{}\underline{A}$ has different values, so in the supervaluation $\underline{T}\tilde{}\underline{A}$ has no value. However, it is hardly plausible that $\underline{A}$ be true, as well as 'It is true that $\underline{A}$', yet 'It is false that $\underline{A}$' have no value.

Similarly, if $\tilde{}\underline{A}$ is true, $\underline{T}\tilde{}\underline{A}$ is also, but $\underline{TA}$ has no value. Two possible solutions present themselves. First, we introduce some restrictions on the classical valuations, and second, we introduce some further necessitations between $\underline{A}$, $\underline{TA}$ and $\underline{T}\tilde{}\underline{A}$. The former can be done as follows:

In the definition of classical valuation, 2, add (3.6)

If $\underline{v}(\underline{TA}) = t$, then $\underline{v}(\underline{T}\tilde{}\underline{A}) = f$.

This will result in both $\underline{v}_1$ and $\underline{v}_5$ being eliminated as classical valuations, so that all classical valuations in which $\underline{A}$ is true, and $\underline{TA}$ is true will also be ones in which $\underline{T\sim A}$ is false, giving the desired result at the level of supervaluation. The same happens for $\sim\underline{A}$, $\underline{T\sim A}$ and $\underline{TA}$.

On the other hand, we could retain the original definition of classical valuation and add $\langle\underline{TA},\underline{T\sim A}\rangle$ and $\langle\underline{T\sim A},\underline{TA}\rangle$ to $\underline{N}$. Then if $\underline{A}$ is true, so are both $\underline{TA}$ and $\underline{T\sim A}$ by necessitation, and similarly if $\underline{A}$ is false, *mutatis mutandis*.

However, a further problem presents itself when we suppose that $\underline{A}$ is undefined in the supervaluation $\underline{s}_k$, and this time neither restrictions on classical valuations nor additional necessitations can be used to solve it. If $\underline{A}$ is neither true nor false, then it is not in $\underline{G}$ and neither is $\sim\underline{A}$. By van Fraassen's reasoning about truth in this situation, $\underline{TA}$ should be false, and so should $\underline{T\sim A}$, so both $\sim\underline{TA}$ and $\sim\underline{T\sim A}$ should belong to $\underline{G}$. However, nothing said so far has the effect of bringing this about. Intuitively, necessitation can add other sentences to $\underline{G}$ if certain sentences are already in it, but it has no power to add anything to $\underline{G}$ in virtue of any sentences which are not in it.

More precisely, we can look at the admissible valuations which can be formed from the classical valuations in (3.5). Using either the restriction (3.6) on classical valuations, or the additional necessitations suggested above, we can see that the following saturated supervaluations are possible:

to truth- saturated sets rather than ordinary saturated sets. With this revised definition, if $\underline{A}$ is true in a supervaluation, so is $\underline{TA}$ (by closure under $\underline{N}$) and so is $\underline{\sim T \sim A}$ (by truth-saturation and consistency). Similarly, if $\underline{A}$ is false, so is $\underline{TA}$, and $\underline{T \sim A}$ is true. If $\underline{A}$ is neither true nor false, $\underline{TA}$ cannot be true, so $\underline{\sim TA}$ must be, and similarly $\underline{\sim T \sim A}$ must be true also. Notice that the condition of truth-saturation eliminates the need for the restriction on classical valuations imposed by (3.6), or the similar additional necessitations.[29] I shall collect here various useful results: those noted (a) depend essentially on truth-saturation, and those noted (b) could be obtained by restrictions on classical valuations. They will be referred to subsequently as TP1, TP2, etc.

$$\begin{aligned} &1.\ \underline{A} \models \underline{TA} && (3.8)\\ &2.\ \underline{A} \models \underline{\sim T \sim A} && \text{b)}\\ &3.\ \underline{\sim A} \models \underline{T \sim A}\\ &4.\ \underline{\sim A} \models \underline{\sim TA} && \text{b)}\\ &5.\ \underline{TA} \models \underline{A} \end{aligned}$$

---

[29] Van Fraassens's views on the foregoing discussion are unclear. All he ever says that we have to include in $\underline{N}$ to obtain a truth theory are $\langle \underline{A}, \underline{TA} \rangle$ and $\langle \underline{TA}, \underline{A} \rangle$ (in PIS-R, p. 145, and TPC, p. 16 and p. 21). He does say that by restricting the set of classical valuations, or by careful construction of $\underline{N}$, we could make "various harmless principles about $\underline{T}$ valid; e.g. $\underline{TA} \supset \underline{A}$" (TPC, p. 16, and similarly at TPC p. 20). On the other hand, the principle 'If $\underline{A}$ is true, so is $\underline{TA}$; otherwise $\underline{TA}$ is false', which is espoused in PIS-R, p. 145, cannot, as I have shown, be made true by any restrictions on $\underline{V}_o$ or by any changes in $\underline{N}$. I do not know whether van Fraassen thought that it could, but he certainly gave the impression that it could.

Moreover, the inference from $\underline{\sim A}$ to $\underline{\sim TA}$ is not merely a "harmless principle": it must be validated if van Fraassen's claims about the paradoxes are to justified, as in the proof that the Strengthened Liar presupposes a contradiction, (3.12). All in all, the account of truth and the paradoxes might have been clearer.

6. $\underline{TA} \models \sim\underline{T}\sim\underline{A}$ b)

7. $\underline{T}\sim\underline{A} \models \sim\underline{A}$

8. $\underline{T}\sim\underline{A} \models \sim\underline{TA}$ b)

9. If $\underline{A}$ is assigned u in $\underline{s}_k$, then $\sim\underline{TA}$ and $\sim\underline{T}\sim\underline{A}$ are both true. a)

10. If $\underline{A}$ is assigned u in $\underline{s}_k$, $\underline{TA} \vee \underline{T}\sim\underline{A}$ is false. a)
Hence the principle of bivalence fails.

However, although bivalence fails, what van Fraassen calls the "ultimate presupposition" holds: every sentence presupposes its own bivalence. This is a simple result, using only TP1 and 3, and the "strong" disjunctive inference, which supervaluations license:

(3.9)

1. $\underline{A} \models \underline{TA}$ by TP1
   $\underline{TA} \models \underline{TA} \vee \underline{T}\sim\underline{A}$ by $\vee$

2. $\sim\underline{A} \models \underline{T}\sim\underline{A}$ by TP3
   $\underline{T}\sim\underline{A} \models \underline{TA} \vee \underline{T}\sim\underline{A}$ by $\vee$

Hence $\underline{A}$ presupposes $\underline{TA} \vee \underline{T}\sim\underline{A}$.

I would like to emphasize at this point that the principles for handling $\underline{T}$ have been justified by general considerations, rather than by the desire to accommodate the Liar. Given that we accept the initial claim that failure of presupposition makes a sentence neither true nor false, we must reject the Principle of Bivalence, and since Tarski sentences entail that principle, under a plausible interpretation of $\underline{TA}$ $\underline{T}\sim\underline{A}$, we reject Tarski biconditionals as well. We retain, however, the intuitions that if $\underline{A}$ is true, so is $\underline{TA}$, and vice versa, and that if $\underline{A}$ is neither true nor false, $\sim\underline{TA}$ and $\sim\underline{T}\sim\underline{A}$ are both true.

Given all this, how does the theory handle the Liar? Of course the propositional syntax gives us no way of expressing Liar sentences, relying as they do on self-reference and truth-predication. What we can do is introduce, for a sentence $\underline{X}$, enough necessitations to enable us to

get the effect of the Liar, and study that. What are the semantic necessitations of 'This sentence is false', for instance? We are able to see that if this sentence is $\underline{X}$, both $\langle \underline{X}, \underline{T}\tilde{\underline{X}} \rangle$ and $\langle \underline{T}\tilde{\underline{X}}, \underline{X} \rangle$ should belong to $\underline{N}$. In effect, $\underline{N}$ now contains three different types of relation: necessitations due to presuppositions, truth relations, and Liar sentences.

The gist of van Fraassen's solution of the Liar is that he can show that it has contradictory presuppositions, and hence is neither true nor false. The proof is quite straightforward, using principles from (3.8).

1. $\underline{X} \models \underline{X} \;\&\; \tilde{\underline{X}}$: (3.10)

a. $\underline{X} \models \underline{X}$	trivially
b. $\underline{X} \models \underline{T}\tilde{\underline{X}}$	by $\underline{N}$ for $\underline{X}$
$\underline{T}\tilde{\underline{X}} \models \tilde{\underline{X}}$	TP7
$\underline{X} \models \tilde{\underline{X}}$	transitivity

2. $\tilde{\underline{X}} \models \underline{X} \;\&\; \tilde{\underline{X}}$:

a. $\tilde{\underline{X}} \models \underline{T}\tilde{\underline{X}}$	TP3
$\underline{T}\tilde{\underline{X}} \models \underline{X}$	$\underline{N}$ for $\underline{X}$
$\tilde{\underline{X}} \models \underline{X}$	transitivity
b. $\tilde{\underline{X}} \models \tilde{\underline{X}}$	trivially

Since both $\underline{X}$ and $\tilde{\underline{X}}$ entail a contradiction, $\underline{X}$ presupposes the same contradiction. But a sentence that presupposes a contradiction is neither true nor false, by Theorem 10, above. It is instructive to look also at the consequences of $\underline{TX}$:

$\underline{TX} \models \underline{X} \;\&\; \tilde{\underline{X}}$: (3.11)

1. $\underline{TX} \models \underline{X}$	TP5
2. $\underline{TX} \models \underline{X}$	TP5
$\underline{X} \models \underline{T}\tilde{\underline{X}}$	$\underline{N}$ for $\underline{X}$
$\underline{T}\tilde{\underline{X}} \models \tilde{\underline{X}}$	TP7
$\underline{TX} \models \tilde{\underline{X}}$	by transitivity

However, no corresponding result holds for $\sim\underline{TX}$: that is, we cannot establish that $\underline{TX}$ presupposes a contradiction, only that it entails one. Correspondingly, $\underline{T}\sim\underline{X}$ entails but does not presuppose a contradiction. Since neither $\underline{TX}$ nor $\underline{T}\sim\underline{X}$ can be true, $\sim\underline{TX}$ and $\sim\underline{T}\sim\underline{X}$ must be (by truth saturation), in accordance with TP9 and indeed with common intuitions. A Liar sentence says of itself that it is false, but it is neither true nor false: consequently the claims that it is not false and not true are both true.

Everything thus seems wonderful, until we turn to the Strengthened Liar. For van Fraassen, the Strengthened Liar is a sentence that says of itself that it is either false or undefined. Thus, parallel to the Ordinary Liar, it can be represented by a sentence $\underline{Y}$ for which both $\langle\underline{Y},\underline{T}\sim\underline{Y}\vee(\sim\underline{TY}\ \&\ \sim\underline{T}\sim\underline{Y})\rangle$ and $\langle\underline{T}\sim\underline{Y}\vee(\sim\underline{TY}\ \&\ \underline{T}\sim\underline{Y}),\underline{Y}\rangle$ belong to $\underline{N}$. And just as for the Liar, it can easily be shown that $\underline{Y}$ presupposes a contradiction:

1. $\underline{Y}\models\underline{TY}\&\sim\underline{TY}$ (3.12)

a. $\underline{Y}\models\underline{TY}$	TP1
b. $\underline{Y}\models\underline{T}\sim\underline{Y}\vee(\sim\underline{TY}\&\sim\underline{T}\sim\underline{Y})$	$\underline{N}$ for $\underline{Y}$
$\underline{T}\sim\underline{Y}\models\sim\underline{TY}$	TP8
$\sim\underline{TY}\&\sim\underline{T}\sim\underline{Y}\models\sim\underline{TY}$	classical logic
$\underline{Y}\models\sim\underline{TY}$	transitivity

2. $\sim\underline{Y}\models\underline{TY}\&\sim\underline{TY}$

a. $\sim\underline{Y}\models\underline{T}\sim\underline{Y}$	TP3
$\underline{T}\sim\underline{Y}\models\underline{T}\sim\underline{Y}\vee(\sim\underline{TY}\&\sim\underline{T}\sim\underline{Y})$	logic
$\underline{T}\sim\underline{Y}\vee(\sim\underline{TY}\&\sim\underline{T}\sim\underline{Y})\models\underline{Y}$	$\underline{N}$ for $\underline{Y}$
$\underline{Y}\models\underline{TY}$	TP1
$\sim\underline{Y}\models\sim\underline{TY}$	transitivity
b. $\sim\underline{Y}\models\underline{T}\sim\underline{Y}$	TP3
$\underline{T}\sim\underline{Y}\models\sim\underline{TY}$	TP8
$\sim\underline{Y}\models\sim\underline{TY}$	transitivity

So $\underline{Y}$, like $\underline{X}$, is neither true nor false. However, when we look at $\underline{TY}$, a difference emerges:

1. $\underline{TY} \models \underline{TY} \& \sim \underline{TY}$ (3.13)

a. $\underline{TY} \models \underline{TY}$	trivially
b. $\underline{TY} \models \underline{Y}$	TP5
$\underline{Y} \models \sim \underline{TY}$	(3.12)
$\underline{TY} \models \sim \underline{TY}$	transitivity

2. $\sim \underline{TY} \models \underline{TY} \& \sim \underline{TY}$

a. $\sim \underline{TY} \models \underline{T} \sim \underline{Y}$ $(\sim \underline{TY} \& \sim \underline{T} \sim \underline{Y})$	classical
$\underline{T} \sim \underline{Y}$ $(\sim \underline{TY} \& \sim \underline{T} \sim \underline{Y}) \models \underline{Y}$	N for $\underline{Y}$
$\underline{Y} \models \underline{TY}$	TP1
$\sim \underline{TY} \models \underline{TY}$	
b. $\sim \underline{TY} \models \sim \underline{TY}$	trivially

So not only does $\underline{Y}$ presuppose a contradiction, $\underline{TY}$ does also, and thus $\underline{TY}$ is neither true nor false. But the buck does stop somewhere:

1. $\underline{TTY} \models \underline{TY}$	TP5	(3.14)
$\underline{TY} \models \underline{TY} \& \sim \underline{TY}$	(3.13)	
$\underline{TTY} \models \underline{TY} \& \sim \underline{TY}$		
2. $\underline{T} \sim \underline{TY} \models \sim \underline{TY}$	TP7	
$\sim \underline{TY} \models \underline{TY} \& \sim \underline{TY}$	(3.13)	
$\underline{T} \sim \underline{TY} \models \underline{TY} \& \sim \underline{TY}$		

Hence both $\underline{TTY}$ and $\underline{T} \sim \underline{TY}$ entail a contradiction, but neither presupposes one, since neither $\sim \underline{TTY}$ nor $\sim \underline{T} \sim \underline{TY}$ entails one. Hence $\sim \underline{TTY}$ and $\sim \underline{T} \sim \underline{TY}$ are true.

Before going on to consider various criticisms of this proposal, I shall give an outline of van Fraassen's quantificational version of the theory, which he gives in TPC, p. 19ff. Suppose we have a quantificational language $\underline{L}_0$ which includes a distinguished monadic predicate $\underline{T}$, and quotation names for all its expressions. The question is how to use supervaluations in such a case so as to define the truth predicate. In many ways, the easiest method of applying supervaluations in quantification is the following. In an ordinary three-valued model, every predicate is assigned an extension and an anti-extension, as

described for MW, above. Then we can define a supervaluation as the intersection of all the classical valuations which arbitrarily fill in the "blanks" in the three-valued model. Thus in a three object universe, we might have a three-valued model which assigned to F the ordered pair $\langle\{d_1\},\{d_3\}\rangle$. Then the appropriate classical valuations are those that assign either $\{d_1,d_2\}$ or just $\{d_1\}$ to F. Since, for example, both (Ex)Fx and (Ex)~Fx are true in all such valuations, the supervaluation will assign t to both these sentences.

However, when trying to keep track of presuppositions, the situation is more complicated. Once again, we have to define a saturated set G, and then take a supervaluation over those valuations that satisfy G. The reason for this is that since presupposition remains a semantic relation between sentences, we have to see which sentences can be true together before taking the supervaluation. So once again, we have to consider the result of adding a presuppositional semantics $\langle V_0, N\rangle$ to L_0. In this case, V_0 is the set of customary classical valuations, subject to the condition that quotation names receive their intended interpretation. If N contains, for every sentence A, $\langle A,T'A'\rangle$ and $\langle T'A',A\rangle$, then we can establish a representation theorem (the theorem numbers that follow are van Fraassen's own):

Theorem 1 In any admissible valuation for $L_0 + \langle V_0,N\rangle$, $s_k(A)=t$ iff $s_k(T'A')=t$. (3.15)

Theorem 1 follows because of the construction of the saturated set, given the condition on N. It is worth noting the difference between this representation theorem and that of Martin and Woodruff. In the latter, roughly speaking we were able to show that, given any model, we could construct another model which agreed on non-truth predications with the original and also had a representative truth-predicate. Here we have no guarantee that, if we start with a collection of true

sentences and then add a truth necessitation set, the original sentences are still true. This difference leads to the following two theorems:

Theorem 2: $\underline{X} \models \underline{A}$ holds in $\underline{L}$ ($=\underline{L}_0 + \langle \underline{V}_0, \underline{N} \rangle$) (3.16)
iff $\underline{A}$ belongs to the smallest set of sentences of $\underline{L}$ containing $\underline{X}$ and closed under $\underline{N}$ and the relation of semantic entailment in $\underline{L}_0$ and $\underline{V}_0$.

Theorem 3: If $\underline{N}$ only contains $\langle \underline{A}, \underline{T}'\underline{A}' \rangle$ and $\langle \underline{T}'\underline{A}', \underline{A} \rangle$ for each $\underline{A}$ in $\underline{L}_0$ ($\underline{N}$ is minimal), and $\underline{X}$ is a set of sentences containing neither $\underline{T}$ nor quotation names which is satisfiable in $\underline{L}_0 + \underline{V}_0$,
then $\underline{X}$ is satisfiable in $\underline{L}_0 + \langle \underline{V}_0, \underline{N} \rangle$.

The third theorem goes a long way towards assuring us that sentences which are true in a classical model will turn out true in a truth-representing model too: that is, that incorporating truth-talk in the language does not endanger non-truth talk. There are some limitations however: it does not protect the veracity of 'Sarah said that "Snow is white" is true'. Nevertheless, it is an important result.

However, I shall not discuss this quantificational language any further. As van Fraassen says, "this is largely an application of previous techniques and results to a new case" (TPC, p. 19), and it is easier to discuss the examples in the propositional theory developed earlier. Since this seems to entail little loss of generality, I shall return to that theory.

I want to start my discussion of the success of van Fraassen's theory with two general objections which rest on the Martin-style methodology which I have attributed to van Fraassen. According to this methodology, a solution to the Liar requires an independently motivated theory of some feature of the language, plus an account of how that theory disarms the Liar. The two objections I shall consider first hold that van Fraassen's theory of presupposition is a bad one, and that he has not shown that the Liar is really a problem of presupposition.

There seem to be two points advanced against van Fraassen's theory

of presupposition: first, that it has the implausible consequence that every sentence presupposes any logically true sentence, and second, that presupposition is not a semantic relation anyway.[30] With respect to the first objection, it seems clear that our intuitions about presupposition do not accord with van Fraassen's theory here, but this does not seem insurmountable. We could, perhaps, modify the definition of presupposition:

$$\underline{A} \text{ presupposes } \underline{B} \text{ iff } \underline{A}^{!}=\underline{B} \text{ and } \sim\underline{A}^{!}=\underline{B} \text{ and } \not\perp=\underline{B}. \qquad (3.17)$$

This fits the bill quite adequately. The second objection is less easily disposed of. I am generally inclined to agree that treating presupposition as a pragmatic matter gives rise to a more interesting general theory, as Stalnaker has argued: in this account, presupposing is something we do in conversational contexts, and what the presuppositions of an assertion are may depend on what has gone before, and what the shared beliefs of speaker and audience may be. Nevertheless, while van Fraassen's theory may not be a complete account of presupposition, it seems to capture some aspects. It seems plausible to argue, for instance, that the relationship of 'The present King of France is bald' to 'There is a King of France' is not merely in the way we use these sentences in conversation, and that there is a semantic relation between them, so that the former simply cannot be true if the latter is false , rather than just not assertible. On this view, then, some pragmatic presuppositions would be explained by reference to semantic presupposition. At any rate as a theory of semantic presupposition, van Fraassen's account is not to be dismissed out of hand.

---

[30] See Robert Stalnaker, "Presuppositions" and "Pragmatic Presuppositions".

The next objection, that the Liar has nothing to do with presupposition, rests partly on the realization that most of the results obtained about X and Y would hold even if there were no reference to presupposition at all. We can see that this is so from several features: first, truth is handled not as a matter of presupposition, but just as a set of co-necessitations, and second, although we explained the non-truth of the Liar sentences by reference to their contradictory presuppositions, this was in fact inessential. Since X and Y and their negations both necessitate contradictions, there is no way either could be true or false, since no true sentence can necessitate a contradiction. Therefore, the reference to presupposition in proofs (3.10), (3.12), and (3.13) above is irrelevant. In fact, all that there is in this account, the objection might run, is a disguised use of reductio: the Liar cannot be true and cannot be false, so it is neither.

This objection can be strengthened because of the problem, mentioned earlier in connection with presuppositional policies, that truth-value gaps do not necessarily signify failure of presupposition. Now it does seem possible to restrict the admissible valuations, and impose other conditions in such a way as to arrive at languages in which all and only those sentences whose presuppositions fail are neither true nor false. However, one of the conditions is that, for every pair $\langle A,B\rangle$ included in N, $\langle \sim A,B\rangle$ also must be included. In effect, this means that the only necessitations allowed are those due to presuppositions. But we have also been including $\langle A,TA\rangle$, and would hardly want to have to put in $\langle \sim A,TA\rangle$ as well, since this seems most implausible. In other words, the truth relation is not a presuppositional one, and in a language in which truth is treated this way, there are bound to be truth-value gaps not attributable to failure of presupposition. Thus we can hold that the Liar is neither true nor false without referring to presupposition.

However, all this ignores the fact that according to the theory the Liar does have contradictory presuppositions, and to demand that we prove that fact without reference to the necessitations concerning it is to demand excessive purity. The argument, stripped of its justification, might look like a bald reductio, but it is not. What van Fraassen is offering is a general account of semantic presupposition, and a theory of how to handle truth based on that account. Furthermore, the theory of truth does not rely on any observations about the Liar, as I pointed out above, but only on methods of avoiding the Principle of Bivalence. These taken together lead to certain results about the Liar, and justify our saying that the Liar is neither true nor false because of these results.

Nevertheless, critics have advanced various other objections against van Fraassen's account. Most of these focus on the treatment of the Strengthened Liar, and of bivalence. As we saw earlier, the Strengthened Liar, $\underline{Y}$, is a sentence which is neither true nor false, and for which $\underline{TY}$ is neither true nor false. Similarly, $\underline{T\sim Y}$ cannot be true:

$\underline{T\sim Y} \models \underline{T\sim Y} \vee (\sim\underline{TY} \ \& \ \sim\underline{T\sim Y})$	Classical Logic	(3.18)
$\underline{T\sim Y} \vee (\sim\underline{TY} \ \& \ \sim\underline{T\sim Y}) \models \underline{Y}$	N for $\underline{Y}$	
$\underline{Y} \models \underline{TY}$	TP1	
$\underline{TY} \models \underline{TY} \ \& \ \sim\underline{TY}$	(3.13)	
$\underline{T\sim Y} \models \underline{TY} \ \& \ \sim\underline{TY}$	Transitivity	

Nor can $\sim\underline{TY} \ \& \ \sim\underline{T\sim Y}$ be true:

$\sim\underline{TY} \ \& \ \sim\underline{T\sim Y} \models \underline{T\sim Y} \vee (\sim\underline{TY} \ \& \ \sim\underline{T\sim Y})$	Classical logic	(3.19)
$\underline{T\sim Y} \vee (\sim\underline{TY} \ \& \ \sim\underline{T\sim Y}) \models \underline{Y}$	N for $\underline{Y}$	
$\underline{Y} \models \underline{TY}$	TP1	
$\underline{TY} \models \underline{TY} \ \& \ \sim\underline{TY}$	(3.13)	

Hence none of $\underline{TY}$, $\underline{T\sim Y}$, and $\sim\underline{TY} \ \& \ \sim\underline{T\sim Y}$ are true. On the other hand, (3.20) is a tautology:

$$\underline{TY} \vee \underline{T\sim Y} \vee (\sim\underline{TY} \ \& \ \sim\underline{T\sim Y}). \qquad (3.20)$$

Now as we saw before, it is a commonplace of supervaluations that

true disjunctions need not have true disjuncts, so the fact that (3.20) can be true even though none of its disjuncts is not, in one respect, a surprise. However, the three disjuncts are the expressions which we have been associating with '$\underline{Y}$ is true', '$\underline{Y}$ is false', and '$\underline{Y}$ is neither true nor false', and one of these is true. In fact, we would say (I have already said it) that $\underline{Y}$ is neither true nor false. Van Fraassen therefore faces a fork: either he admits that we have an ineffability problem, in that we cannot say what is true,[31] or he denies that ($\sim\underline{TY}$ & $\sim\underline{T}\sim\underline{Y}$) says that the Strengthened Liar is neither true nor false, in which case it seems that the effect of the Strengthened Liar is not captured by the sentence $\underline{Y}$.

In fact, van Fraassen attempts to evade the fork in an interesting way (TPC p. 17). The sentence $\sim\underline{TTY}$ & $\sim\underline{T}\sim\underline{TY}$ is true, i.e. the sentence asserting that $\underline{TY}$ is neither true nor false is true. Now suppose we write a sequence of n $\underline{T}$'s as $\underline{T}^n$, and let $\underline{A}$ be written as $\underline{T}^0\underline{A}$. Then we can call a sentence $\underline{A}$ *of value type* $\underline{n}$ if n is the least integer for which $\underline{T}^n\underline{A}$ is true or false. Then 'Snow is white' is of degree 0, $\underline{X}$ is of degree 1, because $\underline{X}$ is neither true nor false but $\underline{TX}$ is false, and $\underline{Y}$ is of degree 2 because $\underline{TTY}$ is false. Then there is a sentence which, in an extended sense, expresses the non-truth of $\underline{Y}$. However, the breathing space this manoeuvre affords is slim.

Suppose for some sentence $\underline{Z}$, $\underline{N}$ contains $\langle\underline{Z},\sim\underline{T}^n\underline{Z}\rangle$ and $\langle\sim\underline{T}^n\underline{Z},\underline{Z}\rangle$, for every finite n. This sentence could be expressed by

For no value type n is this sentence true. (3.21)

The result is that $\underline{Z}$ is of value type ω, and its non-truth is once again inexpressible.

---

[31] The ineffability problem is urged by Brian Skyrms in "Return of the Liar", p. 161.

The Strengthened Liar creates another problem for van Fraassen's account. Y, after all, forces a failure of bivalence at the level of truth ascriptions: neither TY nor ~TY, and neither T~Y nor ~T~Y are true. What can we say about the notion of a truth-saturated set, as given in Definition 11? Clearly, no set containing any of TY, ~TY, T~Y, or ~T~Y and closed under N and classical entailment is satisfiable, so long as N contains both the standard truth necessitations and the necessitations for Y. So there can be no admissible valuations in my revised sense. Thus in order to be able to say anything true at all, we must abandon the restriction of admissible valuations to those generated by truth-saturated sets, and return to van Fraassen's original definition. However, as we saw, something like truth-saturation was essential if any semantic facts about sentences which are neither true nor false are to be reportable. Thus the Strengthened Liar shows not only ineffability in the language about its own non-truth: it forces us to stop saying anything about any non-true, non-false sentence. In the end, van Fraassen, just like Martin, has three exclusive and exhaustive categories, one of which is such that things that fall in it cannot be said not to be in either of the others.

However, problems in handling the Strengthened Liar and 'neither true nor false' are not peculiar to van Fraassen: they are in fact a pervasive problem for three-valued solutions. I shall discuss this aspect further when talking generally about such solutions, but I want now to turn to other desiderata of solutions to the paradoxes.

The Truth-Teller brings out strongly the problem, discussed above, of sentences being neither true nor false without failure of presupposition. We can introduce the Truth-Teller as a sentence W for which <W,TW> and <TW,W> are included in N. Since, however, these are just the standard necessitations for truth, W can clearly never be shown to have contradictory presuppositions. On the other hand, for any

saturated set $\underline{G}$ which does not contain $\underline{W}$, there is another which contains it and its consequences, but is otherwise identical, and a third which contains its negation. Thus in some supervaluations, $\underline{W}$ is true, in others false, and in yet others neither true nor false, all without any failure of presupposition. Once again, we see the easy-going nature of the Truth-Teller. It is perhaps worth noting that one of the restrictions imposed to minimize non-presuppositional truth-gaps, that of insisting that admissible valuations be generated by maximal saturated sets, would have the effect of making the Truth-Teller always true or false, a less plausible result than that given by the radical policy.

A serious problem arises when we turn to contingent paradoxes, like the card example, where one side of the card bears the sentence 'The sentence on the other side of this card is true' and the other bears 'The sentence on the other side of this card is false'. It might be thought that this could be dealt with as follows. Introduce two sentences, $\underline{V}_1$ and $\underline{V}_2$, with the following necessitations: $\langle \underline{V}_1, \underline{TV}_2 \rangle, \langle \underline{V}_2, \underline{\sim TV}_1 \rangle$. If we do this, however, we find it insufficient to derive any contradictory necessitations; perhaps surprisingly, we need to add $\langle \underline{\sim V}_1, \underline{TV}_2 \rangle$ and $\langle \underline{V}_2, \underline{TV}_1 \rangle$. Once we have done this, however, it is easy to show that $\underline{V}_1$ and $\underline{V}_2$ have contradictory presuppositions, and hence are neither true nor false. But the situation is not really this simple. The necessitation relation is a relation based on meaning; that is, it is part of our understanding the meaning of 'The present King of France is bald' that we recognize 'There is a present King of France' as a presupposition, and it is part of the meaning of 'It is true that $\underline{A}$' that $\underline{A}$ and $\underline{TA}$ co-necessitate one another. It is not, however, part of the meaning of 'The sentence on the other side of this card is true' that it presupposes the contradiction 'The sentence on the other side of this card is true and not true'. The contradiction is only an unhappy

consequence of the fact that the sentence on the other side of the card is 'The sentence on the other side of the card is false'. The card might have been a flash-card for arithmetic and said '3 x 9 = 27'.

This problem becomes worse when we look at van Fraassen's quantificational theory. Suppose we look at the sentence 'Sentence b) is true'. We are meant to enter in N the appropriate ordered pair. But what that ordered pair is can be revealed only by the model, the facts of the situation, and not by the meaning of the sentence. As Kripke has emphasized,[32] no theory concentrating on the semantic properties of individual sentences can hope to account properly for the "riskiness" of truth ascriptions, namely their liabiity to plunge us into paradox. Instead of taking 'This sentence is false' as the central case to be solved, we should worry about 'Fred said something false yesterday'.

In conclusion, it seems that van Fraassen's theory suffers a milder fate than Martin's, but a final one nonetheless. Unlike Martin's theory of semantic correctness, van Fraassen's theory of presupposition seems comparatively undisturbed by the attempt to handle truth and the Liar, since the motivations for its construction are independent of those problems. Even the strategy for introducing truth was independently motivated. However, the resulting theory fails to handle the Liar in a satisfactory way, and in particular, fails to explain how the Liar does not pollute ordinary truth ascriptions, for as soon as we introduce the Strengthened Liar, we lose the ability to assert the non-truth of non-true sentences. A desirable theory of truth would be one which would at least allow us to say that 'The present King of France is bald' is not true.

---

[32]Saul Kripke, "Outline of a Theory of Truth", pp. 690-694.

Chapter 4

Kripke and Fixed Points

In his "Outline of a Theory of Truth", Saul Kripke specifically denies any claim to be providing the theory of truth for natural languages. Instead, he proposes a collection of theories which have in common a certain mode of construction and he points out some of the virtues of certain aspects of the different theories, without committing himself to a preference. He even expresses a doubt as to whether there is altogether a matter of fact as to which is the best account. Rather, he claims to have provided a structure in which various comparisons can be made, and whose properties can be investigated.

Nevertheless, Kripke makes several methodological claims, both explicitly and implicitly, about what a theory of truth should do, and while all the theories that his construction generates satisfy these conditions, many others do not. Thus while he suggests that there may be no final answer to the question "What is the best theory of truth?", he indicates clearly a belief that some theories are better than others.

The central point about truth and paradox, for Kripke, is that we can find ourselves making paradoxical assertions through sentences which in many contexts would be harmless, but which, when things turn out badly, achieve some kind of vicious circularity of reference. Epimenides, for instance, was presumably trying to tell people about the Cretans, unless as well as being liars, they were logically more sophisticated than we think, and bloody-minded to boot. St. Paul, anyway, took him to have succeeded in saying something about them, even

though Epimenides could not truly say what St. Paul took him to have said. Given this feature of the relationship between truth ascriptions and how the world is, no formal theory of truth can hope to succeed in identifying, let alone disarming, paradoxical sentences merely by reference to some intrinsic semantic or syntactic property that a sentence may possess. There is nothing intrinsically wrong with the sentence "All Cretans are liars", it is just that it is a sentence which can be used in a situation in which paradox is the result.

Kripke uses this property, the "riskiness" of truth ascriptions, to point to a defect in the Tarski hierarchy. Even if we presume that in such sentences as 'Something Fred once said is not true', "true" has an implicit index i which fixes at what level in the hierarchy of languages the sentence is to be assessed, two problems present themselves. First, we are often in situations where we do not know at what level we should be making our claim: if I say 'Something Fred once said is not true', remembering that Fred once said 'Enid always lies to me', and that Enid once, in a helpful moment, told Fred that Paris was the capital of France, then we may have a problem. For if I do not know what the level is that Fred put on his use of 'true' in his remark about Enid, I cannot be sure that my choice of level for my remark may not cause it to misfire. Second, we can think of examples in which we can give plausible grounds for assigning truth-values to sentences which, whatever levels we choose, cannot be given values in Tarski's hierarchy.

Suppose, in addition to Fred's and Enid's utterances above, Enid, holding Fred in considerable esteem, said 'Fred always tells the truth'. We seem to be able to argue as follows: since Enid did once tell Fred a truth, Fred's 'Enid always lies to me' is false, and hence so is Enid's 'Fred always tells the truth'. On the other hand, in the Tarski hierarchy, no two sentences can assess one another's semantic properties: either Fred's or Enid's remark is construed as being at a

higher level than the other's, in which case the higher one can evaluate the lower one, but not vice versa, or both are at the same level, and neither can evaluate the other.

So in the Tarski hierarchy, riskiness comes out in the wrong way: we do not run the risk of saying anything paradoxical, but we do run the risk of not having succeeded in saying what we wanted, even in situations which are not paradoxical. Kripke's objections to van Fraassen's and Martin's work are different: in their work he complains that the theories are not rigorous enough, and that no definition of truth is offered (with the exception of Martin and Woodruff's paper). This last point has some significance when this work is construed as a reply to Tarski's work on truth-definitions. Tarski showed that some of the conditions on the relationship of ML to OL could be relaxed if all that was desired was a consistent theory of truth in ML, rather than a definition of truth. Thus finding a theory in which truth is taken as a primitive in the language, governed by axioms, is less of an achievement than defining the truth predicate in terms of the model, and it is the latter which Kripke's construction gives us.

However, although Kripke does not bring this objection against them, I want to argue that van Fraassen's and Martin's theories are also open to the riskiness objection, at least under the most natural interpretation of what they are doing. Both want to attribute to Liar-paradoxical sentences a certain kind of defect, semantic incorrectness or failure of presupposition, and use this defect to justify the claim that therefore those sentences are without truth-value. This may seem a justifiable approach with respect to 'This sentence is false', which, under one interpretation anyway, can never have a sensible, non-problematic use. Problems, arise, however, when we look at 'Fred always tells the truth', because of its riskiness. Do semantic correctness and failure of presupposition depend on the facts in the right kind of way

to account for the paradoxical sentences? To be sure, for 'Fred always tells the truth' to be true or false its presuppositions have to be satisfied: there has to be someone called Fred, who at least said something, and so forth. But once these facts have been determined, we should not have to go on and find out what he said before being satisfied about the propriety of the assertion. In fact, of course, even checking what Fred said cannot guarantee against vicious circularity, since Fred said things about Enid's remarks, and no doubt about other people's, so we should have to go and check those, too. All this makes presupposition too little a matter of the sense of a sentence, and too much a matter of fact.

It seems that this result arises from two aspects of the method Martin and van Fraassen used. The first is the concentration on 'This sentence is false' as the problem to be solved, and the second is the kind of solution that this example invites. It seems that it suffers from some semantic irregularity or other, so it is plausible to look for the culprit, and having found one, construct a theory which explains the mistake we made in accepting the Liar sentence. But if we accept Kripke's contention, what we should be looking at is our other uses of 'true', with their distressing tendency to fall into paradox, rather than examples whose paradox is flaunted too obviously. Paradox is just a natural danger of truth-ascriptions, not a result of semantic carelessness.

As a consequence of his approach, Kripke's only constraint, apart from purely formal ones, is agreement with our intuitions about the truth of sentences . But he does offer an inviting intuitive picture of the workings of his theory. Suppose we were teaching someone English, and wanted to specify how to use 'is true'. We could say: "If you can assert some sentence by virtue of what you know already, then you can also assert that that sentence is true, and if you can deny a sentence,

then you can deny that the latter sentence is true." By this means we could ascend from 'Roses are blooming in Picardy' to '"Roses are blooming in Picardy" is true' and then to 'Both the sentences just mentioned now are true' and 'I once uttered a true sentence', and so forth. Two questions arise, given this informal account. Can this process make every sentence containing 'true' assertible or deniable? If not, is it a complete account of the rules for 'true', or can it be supplemented to make it complete? As we shall see, the answer to the first question is 'No', and trying to answer the second gives rise to a host of new theories, and to new problems.

Kripke's formal construction models this intuitive picture very closely. Suppose we have a language $\underline{L}$, a standard first-order language with negation, conjunction, and universal quantification as primitives, plus a list of names, denumerable at most, and of n-place predicates $\underline{F}^n$. Let it further contain quotation names of its own sentences, and a truth predicate $\underline{T}$.

Definition 1: $\langle \underline{D}, \underline{v} \rangle$ is a *model for* $\underline{L}$ just in case:

1. $\underline{D}$ is a non-empty set, and $\underline{L} \subseteq \underline{D}$.
2. $\underline{v}$ is a function, the *valuation function*, such that

   a. $\underline{v}(\underline{a}) \in \underline{D}$, where $\underline{a}$ is an individual constant.

   b. $\underline{v}(\text{'}\underline{A}\text{'}) = \underline{A}$, where $\underline{A}$ is a sentence of $\underline{L}$.

   c. $\underline{v}(\underline{F}^n) = \langle \underline{S}_1, \underline{S}_2 \rangle$, where $\underline{F}^n$ is an n-place predicate. $\underline{S}_1, \underline{S}_2 \subseteq \underline{D}^n$ and $\underline{S}_1 \cap \underline{S}_2 = \Lambda$. $\underline{S}_1$ and $\underline{S}_2$ are the *extension* and *anti-extension* of $\underline{F}^n$, respectively, and will be written here as $\underline{F}^n_1$ and $\underline{F}^n_2$.

Definition 2: $\underline{I} = \langle \underline{M}, \langle \underline{S}_1, \underline{S}_2 \rangle \rangle$ (abbreviated $\underline{M}\langle \underline{S}_1, \underline{S}_2 \rangle$) is an *interpretation of* $\underline{L}$ just in case $\underline{M}$ is a model for $\underline{L}$ and $\underline{S}_1$, $\underline{S}_2 \subseteq \underline{L}$ and $\underline{S}_1 \cap \underline{S}_2 = \Lambda$.

Definition 3: α is an *assignment* for $\underline{L}$ just in case $\alpha(\underline{x}) \in \underline{D}$ when $\underline{x}$ is a variable.

Definition 4: $\underline{val}_{\underline{I},\alpha}(e)$ is the value of expression e in interpretation I for assignment α just in case:

1. $\underline{val}_{\underline{I},\alpha}(\underline{t})$
   $=\underline{v}(\underline{t})$ if $\underline{t}$ is a name.
   $=\alpha(\underline{t})$ if $\underline{t}$ is a variable.

2. $\underline{val}_{\underline{I},\alpha}(\underline{F}^n\underline{t}_1\ldots\underline{t}_n)$
   $=t$ iff $\langle\underline{val}_{\underline{I},\alpha}(\underline{t}_1)\ldots\underline{val}_{\underline{I},\alpha}(\underline{t}_n)\rangle\in\underline{F}_1^n$
   $=f$ iff $\langle\underline{val}_{\underline{I},\alpha}(\underline{t}_1)\ldots\underline{val}_{\underline{I},\alpha}(\underline{t}_n)\rangle\in\underline{F}_2^n$.
   $=u$ otherwise.

3. $\underline{val}_{\underline{I},\alpha}(\underline{T}t)$
   $=t$ iff $\underline{t}\in\underline{S}_1$,
   $=f$ iff $\underline{t}\in\underline{S}_2$,
   $=u$ otherwise.

4. $\underline{val}_{\underline{I},\alpha}(\sim\underline{A})$
   $=t$ iff $\underline{val}_{\underline{I},\alpha}(\underline{A})=f$.
   $=f$ iff $\underline{val}_{\underline{I},\alpha}(\underline{A})=t$.
   $=u$ otherwise.

5. $\underline{val}_{\underline{I},\alpha}(\underline{A}\&\underline{B})$
   $=t$ iff $\underline{val}_{\underline{I},\alpha}(\underline{A})=\underline{val}_{\underline{I},\alpha}(\underline{B})=t$.
   $=f$ iff $\underline{val}_{\underline{I},\alpha}(\underline{A})=f$ or $\underline{val}_{\underline{I},\alpha}(\underline{B})=f$.
   $=u$ otherwise.

6. $\underline{val}_{\underline{I},\alpha}((\underline{x})\underline{A})$
   $=t$ iff $\underline{val}_{\underline{I},\alpha'}(\underline{A})=t$,
   for all α' which agrees with α
   except perhaps at $\underline{x}$ ($\alpha'\overset{=}{\underline{x}}\alpha$).
   $=f$ iff for some $\alpha'\overset{=}{\underline{x}}\alpha$, $\underline{val}_{\underline{I},\alpha'}(\underline{A})=f$.
   $=u$ otherwise.

Then A is true (false) under an interpretation I ($\underline{VAL}_{\underline{I}}(\underline{A})=t$ (f)) iff $\underline{val}_{\underline{I},\alpha}(\underline{A})=t$ (f) for all assignments α, and neither true nor false under I otherwise ($\underline{VAL}_{\underline{I}}(\underline{A})=u$).

We can now define the important function $\blacklozenge$, which takes an ordered pair into an ordered pair. $\blacklozenge(\langle\underline{S}_1,\underline{S}_2\rangle)=\langle\underline{S}_1',\underline{S}_2'\rangle$, where $\underline{S}_1'$ is the set of sentences which are true in $\underline{M}\langle\underline{S}_1,\underline{S}_2\rangle$, and $\underline{S}_2'$ is the set of sentences

which are false in $\underline{M}\langle \underline{S}_1, \underline{S}_2 \rangle$. The crucial property of the function $\blacklozenge$ is that it has fixed points: that is, that there are values for which $\blacklozenge(\langle \underline{S}_1, \underline{S}_2 \rangle) = \langle \underline{S}_1, \underline{S}_2 \rangle$. What this signifies is that there are interpretations in which the sentences which are assigned to the extension of $\underline{T}$ are just those which are true in that interpretation, and the sentences which are assigned to the anti-extension of $\underline{T}$ are those which are false in the interpretation. In other words, at a fixed point, $\underline{T}$ is truth-representing.

If we define a precedence relation between the ordered pairs $\langle \underline{S}_1, \underline{S}_2 \rangle$, we can go on to prove a crucial result about $\blacklozenge$, its monotonicity.

Definition 5: $\langle \underline{S}_1', \underline{S}_2' \rangle$ extends $\langle \underline{S}_1, \underline{S}_2 \rangle$ ($\langle \underline{S}_1, \underline{S}_2 \rangle \leqslant \langle \underline{S}_1', \underline{S}_2' \rangle$) iff $\underline{S}_1 \subseteq \underline{S}_1'$ and $\underline{S}_2 \subseteq \underline{S}_2'$.

Theorem 6: If $\langle \underline{S}_1, \underline{S}_2 \rangle \leqslant \langle \underline{S}_1', \underline{S}_2' \rangle$, then $\blacklozenge(\langle \underline{S}_1, \underline{S}_2 \rangle) \leqslant \blacklozenge(\langle \underline{S}_1', \underline{S}_2' \rangle)$.

The proof relies on the following lemma:

Lemma 7: If $\underline{I} = \underline{M}\langle \underline{S}_1, \underline{S}_2 \rangle$ and $\underline{I}' = \underline{M}\langle \underline{S}_1', \underline{S}_2' \rangle$, and if $\langle \underline{S}_1, \underline{S}_2 \rangle \leqslant \langle \underline{S}_1', \underline{S}_2' \rangle$, then for all sentences $\underline{A}$, and all assignments α, if $\underline{val}_{\underline{I},\alpha}(\underline{A}) = t$, $\underline{val}_{\underline{I}',\alpha}(\underline{A}) = t$, and if $\underline{val}_{\underline{I},\alpha}(\underline{A}) = f$, $\underline{val}_{\underline{I}',\alpha}(\underline{A}) = f$.

Proof: By induction on the length of $\underline{A}$.
Induction Hypothesis (IH): for all $\underline{B}$, shorter than $\underline{A}$, and for all α, if $\underline{val}_{\underline{I},\alpha}(\underline{B}) = t$, $\underline{val}_{\underline{I}',\alpha}(\underline{B}) = t$, and if $\underline{val}_{\underline{I},\alpha}(\underline{B}) = f$, $\underline{val}_{\underline{I}',\alpha}(\underline{B}) = f$.

1. $\underline{A}$ is of the form $\underline{F}t_1 \ldots t_n$.
   If $\underline{val}_{\underline{I},\alpha}(\underline{A}) = t$,
   then $\langle \underline{val}_{\underline{I},\alpha}(t_1) \ldots \underline{val}_{\underline{I},\alpha}(t_n) \rangle \in \underline{F}_1$.
   Then $\underline{val}_{\underline{I}',\alpha}(\underline{A}) = t$.

   If $\underline{val}_{\underline{I},\alpha}(\underline{A}) = f$,
   then $\langle \underline{val}_{\underline{I},\alpha}(t_1) \ldots \underline{val}_{\underline{I},\alpha}(t_n) \rangle \in \underline{F}_2$.
   Then $\underline{val}_{\underline{I}',\alpha}(\underline{A}) = f$.

2. $\underline{A}$ is of the form $\sim\underline{B}$.
If $\underline{val}_{\underline{I},\alpha}(\underline{A})=t$, $\underline{val}_{\underline{I},\alpha}(\underline{B})=f$.
By IH, $\underline{val}_{\underline{I}',\alpha}(\underline{B})=f$,
so $\underline{val}_{\underline{I}',\alpha}(\underline{A})=t$.

Similarly if $\underline{val}_{\underline{I},\alpha}(\underline{A})=f$.

3. $\underline{A}$ is of the form $\underline{B}\&\underline{C}$.
If $\underline{val}_{\underline{I},\alpha}(\underline{A})=t$, $\underline{val}_{\underline{I},\alpha}(\underline{B})=\underline{val}_{\underline{I},\alpha}(\underline{C})=t$,
and by IH, $\underline{val}_{\underline{I}',\alpha}(\underline{B})=\underline{val}_{\underline{I}',\alpha}(\underline{C})=t$.
So $\underline{val}_{\underline{I}',\alpha}(\underline{A})=t$.

If $\underline{val}_{\underline{I},\alpha}(\underline{A})=f$,
either $\underline{val}_{\underline{I},\alpha}(\underline{B})=f$ or $\underline{val}_{\underline{I},\alpha}(\underline{C})=f$,
so by IH, either $\underline{val}_{\underline{I}',\alpha}(\underline{B})=f$ or $\underline{val}_{\underline{I}',\alpha}(\underline{C})=f$.
So $\underline{val}_{\underline{I}',\alpha}(\underline{A})=f$.

4. $\underline{A}$ is of the form $(\underline{x})\underline{B}$.
If $\underline{val}_{\underline{I},\alpha}(\underline{A})=t$,
then, for all $\alpha' \underset{\underline{x}}{\sim} \alpha$, $\underline{val}_{\underline{I},\alpha'}(\underline{B})=t$.
By IH, $\underline{val}_{\underline{I}',\alpha'}(\underline{B})=t$, for all $\alpha \underset{\underline{x}}{\sim} \alpha$.
So $\underline{val}_{\underline{I}',\alpha}(\underline{A})=t$.

Analogously for $\underline{val}_{\underline{I},\alpha}(\underline{A})=f$.

5. $\underline{A}$ is of the form $\underline{T}\underline{t}$.
If $\underline{val}_{\underline{I},\alpha}(\underline{A})=t$, $\underline{val}_{\underline{I},\alpha}(\underline{t})\in\underline{S}_1$.
Since $\underline{val}_{\underline{I},\alpha}(\underline{t})=\underline{val}_{\underline{I}',\alpha}(\underline{t})$ and $\underline{S}_1\subseteq\underline{S}_1'$,
$\underline{val}_{\underline{I}',\alpha}(\underline{t})\in\underline{S}_1'$.
So $\underline{val}_{\underline{I}',\alpha}(\underline{A})=t$.

Similarly for $\underline{val}_{\underline{I},\alpha}(\underline{A})=f$.

To complete the proof of Theorem 6, let $\blacklozenge(\langle\underline{S}_1,\underline{S}_2\rangle)=\langle\underline{R}_1,\underline{R}_2\rangle$ and $\blacklozenge(\langle\underline{S}_1',\underline{S}_2'\rangle)=\langle\underline{R}_1',\underline{R}_2'\rangle$. Suppose $\underline{A}\in\underline{R}_1$. Then $\underline{A}$ is true in $\underline{M}\langle\underline{S}_1,\underline{S}_2\rangle$, by definition of $\blacklozenge$, so $\underline{val}_{\underline{I},\alpha}(\underline{A})=t$ for all α. But by the lemma, $\underline{val}_{\underline{I}',\alpha}(\underline{A})=t$ for all α, so $\underline{A}$ is also true in $\underline{M}\langle\underline{S}_1',\underline{S}_2'\rangle$, i.e. $\underline{A}\in\underline{R}_1'$. Similarly if $\underline{A}\in\underline{R}_2$. So $\blacklozenge(\langle\underline{S}_1,\underline{S}_2\rangle)\leq\blacklozenge(\langle\underline{S}_1',\underline{S}_2'\rangle)$.

What the monotonicity of $\blacklozenge$ entails is a certain conservativeness: if a sentence $\underline{A}$ is true, or false, in a given interpretation then enlarging the extension or anti-extension of $\underline{T}$ cannot make that sentence

change its value. Instead, all that happens is that some sentences which were neither true nor false become one or the other . This fact, which is crucial to the proof of the existence of fixed points, can be generalized for other valuation schemes than the Strong Kleene valuations given. In particular, both the Weak Kleene and supervaluational schemes preserve the result. Clearly the significant cases in Lemma 7, given other standard relationships, are 2 and 3, negation and conjunction, and I shall look first at possible matrix schemes for these, and then at supervaluations.

For negation, the only other plausible scheme which preserves the classical relationship is exclusion negation. However, with exclusion negation, monotonicity fails. Take any interpretation $\underline{M}\langle\underline{S}_1,\underline{S}_2\rangle$ in which some sentence $\underline{A}$ belongs to neither $\underline{S}_1$ nor $\underline{S}_2$, and represent the exclusion negation of $\underline{A}$ by $-\underline{A}$. In such an interpretation, $-\underline{T}\text{'}\underline{A}\text{'}$ is true, since $\underline{T}\text{'}\underline{A}\text{'}$ is neither true nor false. But then there is an interpretation $\underline{M}\langle\underline{S}_1\cup\{\underline{A}\},\underline{S}_2\rangle$, which extends $\underline{M}\langle\underline{S}_1,\underline{S}_2\rangle$, and in which $-\underline{T}\text{'}\underline{A}\text{'}$ is false. So by extending the extension of $\underline{T}$, we have changed the value of the sentence $-\underline{T}\text{'}\underline{A}\text{'}$ from true to false, and monotonicity cannot be preserved; choice negation is the only possible scheme which preserves classical intuitions and monotonicity.

With conjunction, two options present themselves. I shall restrict myself to those valuation schemes that preserve the classical valuations when both conjuncts are true or false, and that preserve commutativity. These options are given in the following partial table:

A & B	t	f	u
t	t	f	u
f	f	f	f/u
u	u	f/u	u

(4.1)

To see that these are indeed the only options, we need only consider how to preserve the consequences required by Lemma 7

i) If $\underline{val}_{\underline{I}},\alpha(\underline{B}\&\underline{C})=t$, $\underline{val}_{\underline{I}'},\alpha(\underline{B}\&\underline{C})=t$ (4.2)

ii) If $\underline{val}_{\underline{I}},\alpha(\underline{B}\&\underline{C})=f$, $\underline{val}_{\underline{I}'},\alpha(\underline{B}\&\underline{C})=f$

No entry of t is possible in the bottom row or right hand column, if i) is to be preserved, for then $\underline{val}_{\underline{I}},\alpha(\underline{B}\&\underline{C})=t$ might be true and $\underline{val}_{\underline{I}'},\alpha(\underline{B}\&\underline{C})=t$ not, if one conjunct were neither true nor false in $\underline{I}$ but false in $\underline{I}'$, as might occur . Similarly, f is only possible in the places shown, otherwise it would be possible for $\underline{B}\&\underline{C}$ to be false in $\underline{I}$ and have another value in $\underline{I}'$.

Some general conditions in which ⧫ is monotonic have been given by Fine, cited by Martin and Woodruff[33] . Fixed points can be obtained for any set of truth functions ⧫ which include t and f in their field and have the following properties:

Stability: if ⧫ has the value t or f for given truth values as arguments, it retains that value when any argument not in {t,f} is replaced by either of them. (4.3)

Fidelity: when all arguments are in {t,f}, ⧫ behaves classically.

For supervaluations, the proof of Theorem 6 is somewhat different, and I shall not go into it in detail. The result follows quite easily from the notion of a supervaluation as what a certain class of classical valuations have in common. In this case, the class of classical valuations is determined by $\underline{M}\langle\underline{S}_1,\underline{S}_2\rangle$: it is the class of classical valuations which replace each partial interpretation $\langle\underline{F}_1^n, \underline{F}_2^n\rangle$ by a standard two-valued interpretation, in such a way that everything that is in $\underline{F}_1^n$ is in the classical interpretation, and everything that is in $\underline{F}_2^n$ is out of it, and the remainder are distributed in or out. From this we can see that the set of classical valuations generated by $\underline{M}\langle\underline{S}_1',\underline{S}_2'\rangle$ is

---

[33]Martin and Woodruff cite Kit Fine, "Vagueness, Truth and Logic."

included in that generated by $\underline{M}\langle \underline{S}_1, \underline{S}_2 \rangle$, if $\langle \underline{S}_1, \underline{S}_2 \rangle \leqslant \langle \underline{S}_1', \underline{S}_2' \rangle$. But then the appropriate supervaluation for $\underline{M}\langle \underline{S}_1', \underline{S}_2' \rangle$ must give the same value to anything given a value in the supervaluation for $\underline{M}\langle \underline{S}_1, \underline{S}_2 \rangle$, since it just takes what is common to a subset of the latter's set of classical valuations.

So we can see at least that monotonicity does hold for Weak Kleene and supervaluation schemes; to show that $\blacklozenge$ has fixed points, where $\blacklozenge(\langle \underline{S}_1, \underline{S}_2 \rangle) = \langle \underline{S}_1, \underline{S}_2 \rangle$, I shall revert to the Strong Kleene formulation given earlier. So far , all we know is that the sets of true and false sentences in an interpretation which extends another interpretation include the respective sets in the latter interpretation, but we do not know from that how to find a fixed point. Here our student of English gives us a helpful hint. This student knew for non-truth ascriptions which to assert and which to deny. Given that and the rule for 'true', he was able to generate a new collection of assertable and deniable sentences, and then to apply the truth rule again, and so forth. So we can start with $\underline{M}\langle \mathbb{A}, \mathbb{A} \rangle$, and see what sentences are true and false in that interpretation. That gives us an ordered pair, call it $\langle \underline{S}_1', \underline{S}_2' \rangle$. Now see what sentences are true and false in $\underline{M}\langle \underline{S}_1', \underline{S}_2' \rangle$, and use that to construct another interpretation. Since we know that $\langle \mathbb{A}, \mathbb{A} \rangle \leqslant \blacklozenge(\langle \mathbb{A}, \mathbb{A} \rangle)$, $\langle \mathbb{A}, \mathbb{A} \rangle \leqslant \langle \underline{S}_1', \underline{S}_2' \rangle$. So $\blacklozenge(\langle \mathbb{A}, \mathbb{A} \rangle) \leqslant \blacklozenge(\langle \underline{S}_1', \underline{S}_2' \rangle)$, by monotonicity. But $\langle \underline{S}_1', \underline{S}_2' \rangle = \blacklozenge(\langle \mathbb{A}, \mathbb{A} \rangle)$, so what we generate is a steadily growing chain, $\langle \mathbb{A}, \mathbb{A} \rangle$, $\blacklozenge(\langle \mathbb{A}, \mathbb{A} \rangle)$, $\blacklozenge(\blacklozenge(\langle \mathbb{A}, \mathbb{A} \rangle))$, and so on, and taking these as the interpretations gives us a steadily increasing truth set. However, this process alone will not normally produce fixed points. For consider the series:

Apples are green. (4.4)
The previous sentence is true.
The previous sentence is true.
.
.
.

The steadily increasing truth sets will, step by step, incorporate the members of this series. But suppose we also have the sentence

All the members of (4.4) are true (4.5)

This cannot be decided until all of them are, and that will never happen. To accommodate this kind of sentence, we have to carry on the construction of the truth sets at transfinite levels. The way we do this is by taking the union of all the previous extensions, and the union of the anti-extensions, at each limit ordinal, and then carrying on with reiterations of Φ. To see that we finally reach a fixed point, we must first define a series of interpretations, $\underline{I}(\alpha)=\underline{M}\langle\underline{S}_1,\alpha, \underline{S}_2,\alpha\rangle$, by transfinite recursion. In fact, since $\underline{M}$ is constant in all the following interpretations, mention of it will frequently be suppressed.

Definition 8:

For $\alpha=0$, $\underline{I}(\alpha)=\langle\Lambda,\Lambda\rangle$.

$\alpha=\beta+1$, $\underline{I}(\alpha)=\Phi(\underline{I}(\beta))$.

α a limit ordinal,

$$\underline{I}(\alpha)=\langle\bigcup_{\beta<\alpha}\underline{S}_{1,\beta},\bigcup_{\beta<\alpha}\underline{S}_{2,\beta}\rangle.$$

Theorem 9: If $\alpha\leqslant\beta$, then $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$.

Proof: The proof is by strong transfinite induction.
Induction Hypothesis:
(γ)(If $\gamma<\alpha$ then (β)(If $\gamma\leqslant\beta$, then $\underline{I}(\gamma)\leqslant\underline{I}(\beta)$)). We show that
(β)(If $\alpha\leqslant\beta$ then $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$)
Suppose, for arbitrary β, $\alpha\leqslant\beta$. There are three cases:

1. $\beta=0$
   Then $\alpha=0$
   So $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$

2. β is a limit ordinal
   If $\alpha=\beta$, $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$
   If $\alpha<\beta$, $\underline{I}(\alpha)=\underline{I}(\gamma)$, for some $\gamma<\beta$

But $\underline{I}(\beta)=\langle \bigcup_{\gamma<\beta}\underline{S}_1,\gamma, \bigcup_{\gamma<\beta}\underline{S}_2,\gamma\rangle$
So $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$

3. $\beta=\delta+1$

a. $\alpha=0$
$\underline{I}(\alpha)=\langle\Lambda,\Lambda\rangle$
Hence $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$

b. α is a limit ordinal.
$\underline{I}(\alpha)=\langle \bigcup_{\gamma<\alpha}\underline{S}_1,\gamma, \bigcup_{\gamma<\alpha}\underline{S}_2,\gamma\rangle$
By Induction Hypothesis,
(γ)(If $\gamma<\alpha$, then $\underline{I}(\gamma)\leqslant\underline{I}(\beta)$).
Hence $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$.

c. $\alpha=\varepsilon+1$
Instantiating the Induction Hypothesis,
If $\varepsilon<\alpha$ then if $\varepsilon\leqslant\delta$ then $\underline{I}(\varepsilon)\leqslant\underline{I}(\delta)$.
But $\varepsilon<\alpha$ and $\varepsilon\leqslant\delta$.
Hence $\underline{I}(\varepsilon)\leqslant\underline{I}(\delta)$.
By monotonicity, $\Phi(\underline{I}(\varepsilon))\leqslant\Phi(\underline{I}(\delta))$.
But $\underline{I}(\alpha)=\Phi(\underline{I}(\varepsilon))$ and $\underline{I}(\beta)=\Phi(\underline{I}(\delta))$.
Hence $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$.

Since β was arbitrary, (β)(If $\alpha<\beta$ then $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$).

Finally, we can demonstrate the existence of fixed points:

Theorem 10: Φ has a fixed point.

Proof: Suppose, for reductio, that Φ has no fixed points. Thus, for every α, $\underline{I}(\alpha)\neq\underline{I}(\alpha+1)$, and since by Theorem 9 if $\alpha\leqslant\beta$, $\underline{I}(\alpha)\leqslant\underline{I}(\beta)$, for every α,β, $\underline{I}(\alpha)\neq\underline{I}(\beta)$ $(\alpha\neq\beta)$. Since $\underline{I}(\alpha)\in(\mathcal{P}(\underline{D}))^2$, it would follow that there was a relation from $(\mathcal{P}(\underline{D}))^2$ to the ordinals which held between any pair $\langle\underline{S}_1,\underline{S}_2\rangle$ and at most one ordinal. But by the axiom schema of replacement, this entails that there is a set of all the ordinals. Hence for some distinct α,β, $\underline{I}(\alpha)=\underline{I}(\beta)$. But if $\alpha<\beta$, then by Theorem 9, for all γ, $\alpha\leqslant\gamma<\beta$, $\underline{I}(\alpha)=\underline{I}(\gamma)$. So $\underline{I}(\alpha)=\underline{I}(\alpha+1)$, for some α.

This is not, however, the only fixed point of Φ, and much of the interest of Kripke's construction lies in the properties of other fixed

points. To see that there are more fixed points, consider the set of functions defined by:

Definition 11:

For $\alpha=0$, $\underline{F}(\alpha)=\langle \underline{S}_{1,0}, \underline{S}_{2,0}\rangle$.

$\alpha=\beta+1$, $\underline{F}(\alpha)=\blacklozenge(\underline{F}(\beta))$.

α a limit ordinal,
$\underline{F}(\alpha)=\langle \bigcup_{\gamma<\alpha}\underline{S}_{1,\gamma}, \bigcup_{\gamma<\alpha}\underline{S}_{2,\gamma}\rangle$.

In the special case where $\underline{S}_{1,0}=\underline{S}_{2,0}=\Lambda$, $\underline{F}(\alpha)=\underline{I}(\alpha)$. To see that $\underline{F}$ can generate fixed points as $\underline{I}$ can, we need to consider the proof of Theorem 9: the only relevant part is 3a, where we must show that

i) $\underline{F}(0)\lessdot\underline{F}(1)$. (4.6)

ii) If $\underline{F}(0)\lessdot\underline{F}(\alpha)$, then $\underline{F}(0)\lessdot\underline{F}(\alpha+1)$.

In general, for $\underline{F}$, (4.6)i) does not hold. If for some $\underline{F}$ it does hold, however, then (4.6)ii) also holds. For suppose $\underline{F}(0)\lessdot\underline{F}(\alpha)$. Then, by monotonicity of $\blacklozenge$, $\blacklozenge(\underline{F}(0))\lessdot\blacklozenge(\underline{F}(\alpha))$, i.e. $\underline{F}(1)\lessdot\underline{F}(\alpha+1)$. But given $\underline{F}(0)\lessdot\underline{F}(1)$, we obtain $\underline{F}(0)\lessdot\underline{F}(\alpha+1)$, so (4.6)ii) holds. Let us say that a function $\underline{F}$ has an acceptable starting point if (4.6)i) holds for $\underline{F}$: in fact, unless $\underline{F}$ has an acceptable starting point, it is not well-defined, because $\blacklozenge$ is not defined for non-disjoint pairs, and at the limit we cannot guarantee that $\underline{F}(\alpha)$ remains disjoint. Then we can state the theorem:

Theorem 12: If $\underline{F}$ has an acceptable starting point, then for some α, $\underline{F}(\alpha)$ is a fixed point for $\blacklozenge$.

The distinctive feature of the fixed point generated by $\underline{I}$ is that every other fixed point extends it. It will henceforth be written as $\underline{FP}(\underline{I})$ $(=\langle \underline{FP}(\underline{I})_1, \underline{FP}(\underline{I})_2\rangle)$. Other fixed points will be written analogously as $\underline{FP}(\underline{F})$.

Theorem 13: $\underline{FP}(\underline{I}) \leqslant \underline{FP}(\underline{F})$, for all $\underline{FP}(\underline{F})$

Proof: We can prove that $\underline{I}(\alpha) \leqslant \underline{F}(\alpha)$, for all α, by strong induction on α.
Induction Hypothesis:
(β)(If $\beta < \alpha$, then $\underline{I}(\beta) \leqslant \underline{F}(\beta)$)

1. $\alpha = 0$.
   $\underline{I}(0) = \langle \Lambda, \Lambda \rangle$, $\underline{F}(0) = \langle \underline{F}_{1,0}, \underline{F}_{2,0} \rangle$
   so $\underline{I}(\alpha) \leqslant \underline{F}(\alpha)$.

2. $\alpha = \delta + 1$
   $\underline{I}(\delta) \leqslant \underline{F}(\delta)$, by Induction Hypothesis.
   $\Phi(\underline{I}(\delta)) \leqslant \Phi(\underline{F}(\delta))$, by monotonicity.
   But $\underline{I}(\alpha) = \Phi(\underline{I}(\delta))$, $\underline{F}(\alpha) = \Phi(\underline{F}(\delta))$.
   So $\underline{I}(\alpha) \leqslant \underline{F}(\alpha)$.

3. α is a limit ordinal.
   $\underline{I}_{1,\alpha} = \bigcup_{\beta < \alpha} \underline{I}_{1,\beta}$, $\underline{F}_{1,\alpha} = \bigcup_{\beta < \alpha} \underline{F}_{1,\beta}$.
   (β)(If $\beta < \alpha$ then $\underline{I}_{1,\beta} \subseteq \underline{F}_{1,\beta}$)
   by Induction Hypothesis.
   So $\underline{I}_{1,\alpha} \subseteq \underline{F}_{1,\alpha}$.
   Similarly, $\underline{I}_{2,\alpha} \subseteq \underline{F}_{2,\alpha}$.
   So $\underline{I}(\alpha) \leqslant \underline{F}(\alpha)$.

So we have shown $(\alpha)(\underline{I}(\alpha) \leqslant \underline{F}(\alpha))$. Since we know that there is some ordinal β for which $\underline{I}(\beta) = \underline{FP}(\underline{I})$ and $\underline{F}(\beta) = \underline{FP}(\underline{F})$, we obtain $\underline{FP}(\underline{I}) \leqslant \underline{FP}(\underline{F})$.

Thus at every fixed point, at least those sentences are true, and false, which are true, and false, at the minimal fixed point. All that happens at non-minimal fixed points is that some other sentences get truth-values which did not have them at the minimal fixed point. Now that we have the basic features of the construction, we can ask what happens to the Liar and the Truth-Teller. If we suppose that $\underline{M}$ is such that:

$$\begin{array}{l} \underline{v}(\underline{a}) = \underline{\sim Ta} \\ \underline{v}(\underline{b}) = \underline{Tb} \end{array} \qquad (4.7)$$

then we can establish the following results.

Theorem 14: For no function $\underline{F}$ does $\sim\underline{Ta}$ belong to the extension or anti-extension of $\underline{FP}(\underline{F})$, i.e. $\sim\underline{Ta} \notin \underline{FP}(\underline{F})_1 \cup \underline{FP}(\underline{F})_2$.

Proof: Suppose for some $\underline{F}$, $\sim\underline{Ta} \in \underline{FP}(\underline{F})_1$. If $\sim\underline{Ta} \in \underline{FP}(\underline{F})_1$, then $\sim\underline{Ta}$ is true in $\underline{MFP}(\underline{F})$ and so is $\underline{Ta}$. But then $\underline{FP}(\underline{F})$ is not a fixed point. Similarly, if $\sim\underline{Ta} \in \underline{FP}(\underline{F})_2$, $\underline{T}$'$\sim\underline{Ta}$' and $\underline{Ta}$ are false in $\underline{MFP}(\underline{F})$, and $\underline{FP}(\underline{F})$ is again not a fixed point.

Theorem 15: $\underline{Tb} \notin \underline{FP}(\underline{I})_1 \cup \underline{FP}(\underline{I})_2$

Proof: By induction.
Suppose (β)(If $\beta < \alpha$ then $\underline{Tb} \notin \underline{I}_{1,\beta} \cup \underline{I}_{2,\beta}$)

1. $\alpha = 0$.
   $\underline{Tb} \notin \underline{I}_{1,\alpha} \cup \underline{I}_{2,\alpha}$, trivially.
2. $\alpha = \delta + 1$.
   Suppose $\underline{Tb} \in \underline{I}_{1,\alpha}$.
   Then $\underline{Tb}$ is true in $\underline{M}\langle \underline{I}_{1,\delta}, \underline{I}_{2,\delta} \rangle$.
   But by the valuation rule for $\underline{T}$,
   this requires that $\underline{v}(\underline{b}) \in \underline{I}_{1,\delta}$,
   i.e. $\underline{Tb} \in \underline{I}_{1,\delta}$.
   This contradicts the Induction Hypothesis,
   so $\underline{Tb} \notin \underline{I}_{1,\alpha}$.
   Similarly, $\underline{Tb} \notin \underline{I}_{2,\alpha}$
   So $\underline{Tb} \notin \underline{I}_{1,\alpha} \cup \underline{I}_{2,\alpha}$.
3. α is a limit ordinal.
   $\underline{Tb} \notin \underline{I}_{1,\alpha} \cup \underline{I}_{2,\alpha}$,
   by definition of $\underline{I}(\alpha)$ and
   Induction Hypothesis.

Theorem 16: For some $\underline{F}_1$, $\underline{F}_2$, $\underline{Tb} \in \underline{FP}(\underline{F}_1)_1$, $\underline{Tb} \in \underline{FP}(\underline{F}_2)_2$.

Proof: Let $\underline{F}_1(0) = \langle \{\underline{Tb}\}, \Lambda \rangle$, $\underline{F}_2(0) = \langle \Lambda, \{\underline{Tb}\} \rangle$.
Both $\underline{F}_1(0)$ and $\underline{F}_2(0)$ are acceptable starting points; since $\underline{Tb}$ is true in $\underline{M}\langle \{\underline{Tb}\}, \Lambda \rangle$, $\langle \{\underline{Tb}\}, \Lambda \rangle \lessdot \blacklozenge(\langle \{\underline{Tb}\}, \Lambda \rangle)$, and since $\underline{Tb}$ is false in $\underline{M}\langle \Lambda, \{\underline{Tb}\} \rangle$, $\langle \Lambda, \{\underline{Tb}\} \rangle \lessdot \blacklozenge(\langle \Lambda, \{\underline{Tb}\} \rangle)$. Given Theorem 12, then, $\underline{Tb} \in \underline{FP}(\underline{F}_1)_1$, and $\underline{Tb} \in \underline{FP}(\underline{F}_2)_2$.

These results show that neither the Liar nor the Truth-Teller are true or false in the minimal fixed point, and there are fixed points in

which the Truth-Teller does get a value, but none in which the Liar does. This prompts the following definition:

Definition 17:

1. A sentence A is grounded iff $A \in FP(I)_1 \cup FP(I)_2$, and ungrounded otherwise.
2. A sentence A is paradoxical iff for no fixed point FP(F), $A \in FP(F)_1 \cup FP(F)_2$.

We can express the results of Theorems 14, 15, and 16 by saying that the Liar is paradoxical and the Truth-Teller ungrounded but unparadoxical. Before Kripke's definition, groundedness was a property informally ascribed to sentences whose truth-value could be seen to depend ultimately only on brute facts[34] and Kripke's definition gives a formal account of that. If we start from a model and assign no sentences to either the extension or anti-extension of T, then it seems that when we reach a fixed point, a sentence is assigned to the extension or anti-extension of T solely because of the facts of the model and the rule for asserting sentences containing 'true', and not because of any other considerations about truth. The minimal fixed point therefore represents an especially important interpretation of T not merely for the technical reason that it is minimal, but also for the conceptual reasons connected with groundedness.

By contrast with the minimal fixed point, the fixed points $FP(F_1)$ and $FP(F_2)$, above, in which the Truth-Teller is true and false respectively, seem to rest on arbitrary decisions, namely putting Tb into the extension of T in $F_1(0)$ and the anti-extension of T in $F_2(0)$. However, the minimal fixed point is not the only one of theoretical

---

[34] See, for instance, Hans Herzberger "Paradoxes of Grounding in Semantics".

interest. There are two classes which claim special distinction, the maximal fixed points, and the intrinsic ones.

Definition 18:

1. A fixed point $\underline{FP}(\underline{F})$ is maximal iff for every fixed point $\underline{FP}(\underline{F}')$, if $\underline{FP}(\underline{F}) \leq \underline{FP}(\underline{F}')$, then $\underline{FP}(\underline{F}) = \underline{FP}(\underline{F}')$.
2. A fixed point $\underline{FP}(\underline{F})$ is intrinsic iff, for every other fixed point $\underline{FP}(\underline{F}')$, $\underline{FP}(\underline{F})_1 \cap \underline{FP}(\underline{F}')_2 = \underline{FP}(\underline{F})_2 \cap \underline{FP}(\underline{F}')_1 = \Lambda$.

The maximal fixed points have the virtue that they assign truth-values to the greatest possible number of sentences. Since the minimal fixed point clearly has gaps, it might be thought desirable to fill them in as far as possible, though of course if the model gives rise to Liar sentences, the gaps cannot be filled completely. The truth-representing models whose existence is proved in MW correspond to maximal fixed points, whence comes their characteristic of giving the Truth-Teller either the value true or the value false. Although it might seem desirable to fill in as many truth-value gaps as possible, which explains the interest of the maximal fixed points as a group, it seems implausible that any of them represents the extension of 'true', because if there is a Truth-Teller in the model, there seems no good ground for picking a maximal fixed point in which the Truth-Teller is true, say, rather than one where it is false. Both have the advantage of giving it some value, but neither is more plausible than the other, so in the end neither is plausible.

The intrinsic fixed points are of interest because they fill in truth-value gaps, but only where it can be done non-arbitrarily. To see that there are sentences for which this can be done, consider the sentence (4.8):

$$\underline{v}(\underline{c}) = \underline{Tc} \vee \tilde{} \underline{Tc}. \qquad (4.8)$$

Using this sentence we can construct an intrinsic fixed point, as follows:

Theorem 19: If $\underline{F}(0)=\langle\{\underline{Tc}\vee\sim\underline{Tc}\},\Lambda\rangle$, $\underline{FP}(\underline{F})$ is an intrinsic fixed point.

Proof: First, $\underline{F}(0)$ is an acceptable starting point, since $\underline{Tc}$ and $\underline{Tc}\vee\sim\underline{Tc}$ are true in $\underline{M}\langle\{\underline{Tc}\vee\sim\underline{Tc}\},\Lambda\rangle$. Hence there is a fixed point $\underline{FP}(\underline{F})$. To see that it is intrinsic, we need to use transfinite induction. Assume that $\underline{F}'(0)$ is an acceptable starting point , and let $\underline{F}'(\alpha)\#\underline{F}(\alpha)$ just in case $\underline{F}'_{1,\alpha}\cap\underline{F}_{2,\alpha}=\underline{F}'_{2,\alpha}\cap\underline{F}_{1,\alpha}=\Lambda$. In this case let us call the interpretations consilient. Then the Induction Hypothesis is: $(\beta)(\text{If } \beta<\alpha \text{ then } \underline{F}'(\beta)\#\underline{F}(\beta))$

1. $\alpha=0$
   Clearly $\underline{F}'_{1,0}\cap\underline{F}_{2,0}=\Lambda$
   For no acceptable starting point does $\underline{Tc}\vee\sim\underline{Tc}$ belong to $\underline{F}'_{2,0}$, for if it did,
   $\sim\underline{Tc}$ and $\underline{Tc}\vee\sim\underline{Tc}$ would be true in $\underline{M}\underline{F}'(0)$,
   so it would not be the case that $\underline{F}'(0)\lessdot\blacklozenge(\underline{F}'(0))$.
   Hence $\underline{F}'(\alpha)\#\underline{F}(\alpha)$

2. $\alpha=\beta+1$
   $\underline{F}'(\beta)\#\underline{F}(\beta)$ by Induction Hypothesis
   Consider the pair $\underline{FU}=$
   $\langle\underline{F}'_{1,\beta}\cup\underline{F}_{1,\beta},\underline{F}'_{2,\beta}\cup\underline{F}_{2,\beta}\rangle$
   $\underline{F}'(\beta)\lessdot\underline{FU}$, and $\underline{F}(\beta)\lessdot\underline{FU}$.
   By the monotonicity of $\blacklozenge$,
   $\underline{F}'(\beta+1)\lessdot\blacklozenge(\underline{FU})$,
   and $\underline{F}(\beta+1)\lessdot\blacklozenge(\underline{FU})$.
   But $\blacklozenge(\underline{FU})$ is a disjoint pair, by definition.
   Hence $\underline{F}'(\alpha)\#\underline{F}(\alpha)$.

3. α is a limit ordinal.
   Using the Induction Hypothesis,
   trivially, $\underline{F}'(\alpha)\#\underline{F}(\alpha)$

Consequently $(\alpha)(\underline{F}'(\alpha)\#\underline{F}(\alpha))$, and hence $\underline{FP}(\underline{F})\#\underline{FP}(\underline{F}')$, i.e. $\underline{F}$ is an intrinsic fixed point.

Intrinsic fixed points have some formal points of interest: they form a complete lattice under $\lessdot$, for one thing, whose infimum is the minimal fixed point, since this is clearly an intrinsic fixed point, and is a lower bound on any other intrinsic fixed point. The supremum, the

greatest intrinsic fixed point, is an especially important fixed point, being the fixed point which fills the greatest possible number of gaps, without making any arbitrary assignments. Because of this interesting property, it shares with the minimal fixed point the candidacy for the best model of our intuitions about our use of 'true'.[35]

This completes the account of the basic construction, and we can easily see that the answer to the question "Does this step-by-step strategy finally distribute each sentence into either the true or the false?" is "No", at least if the model contains certain fishy kinds of items.[36] Moreover, the account clearly fails to meet a strong intuition which allows that some sentences may be neither true nor false, but holds that any truth ascription should be one or the other: if the model contains a denotation as in (4.7), not only is $\underline{\sim Ta}$ never in the extension or anti-extension of $\underline{T}$ in a fixed point, no sentence ascribing truth or non-truth to it can be either. For if a sentence $\underline{A}$ consisted of $\underline{\sim Ta}$ prefixed by any mixture of negations and truth ascriptions, and $\underline{A}$ was true or false in a fixed point, the sequential detaching of the negations and truth ascriptions would show that $\underline{\sim Ta}$ was true or false there too.

Recognising this alternative intuition, though still preferring his own intuition that $\underline{A}$ and $\underline{T'A'}$ always have the same value, Kripke offers

---

[35]There are some limitations to the plausibility of the largest intrinsic fixed point, which can be seen from considering $\underline{Tb} \vee \underline{\sim Tb}$. In any fixed point where it has a value, this sentence, like $\underline{Tc} \vee \underline{\sim Tc}$, is always true. However, for it to be true, $\underline{Tb}$ or $\underline{\sim Tb}$ must be true, so it is never true in an intrinsic fixed point. Thus the greatest intrinsic fixed point does not give every unambiguously valued sentence its appropriate value , though it does give values to all sentences which can receive one without making any arbitrary choices.

[36]In fact, as we shall shortly see, the answer can still be "No", even if the model contains no obviously fishy items.

some further interpretations, based on a prior construction of the standard hierarchy. The first move is to "close off" a fixed point.

Suppose we have a fixed point $\underline{FP}(\underline{F})$. Then the closure of $\underline{FP}(\underline{F})$ interprets $\underline{T}$ by $\langle \underline{FP}(\underline{F})_1, \overline{\underline{FP}(\underline{F})_1} \rangle$, where $\overline{\underline{FP}(\underline{F})_1}$ is the complement, relative to $\underline{D}$, of $\underline{FP}(\underline{F})_1$. Call this interpretation $\underline{Clos}(\underline{FP}(\underline{F}))$. Now clearly if this model contains paradoxical sentences such as $\ulcorner\underline{Ta}\urcorner$, $\underline{Clos}(\underline{FP}(\underline{F}))$ is not a fixed point, nor even an acceptable starting point, since $\ulcorner\underline{Ta}\urcorner \in \overline{\underline{FP}(\underline{F})_1}$. However, $\underline{Clos}(\underline{FP}(\underline{F}))$ enables us to express some facts about $\underline{FP}(\underline{F})$ that are not expressible in the fixed point. Any sentence $\underline{A}$ which is without truth value in $\underline{FP}(\underline{F})$ will belong to $\overline{\underline{FP}(\underline{F})_1}$, and so will $\ulcorner\underline{A}\urcorner$, so in the interpretation $\underline{MClos}(\underline{FP}(\underline{F}))$, both $\ulcorner\underline{T}\text{`}\underline{A}\text{'}$ and $\ulcorner\underline{T}\text{`}\ulcorner\underline{A}\text{'}$ will be true. Thus it seems that in the closure of a fixed point we can express the lack of truth value in the fixed point of sentences like $\underline{A}$.

What is more, closing off a fixed point can easily be included in our instructions for learning to use 'true', since we just add the instruction "And when you find you cannot add any more assertions to the interpretation of 'true', just deny everything that remains undecided". Of course, you have to wait for a fixed point, because something might lack a truth-value at an earlier stage just because of its complexity.

However, although the closure of a fixed point has some advantages over a fixed point, it also suffers from inexpressibility, since it is not a fixed point! In particular, if $\underline{A}$ is as above, $\ulcorner\underline{T}\text{`}\underline{A}\text{'}$ and $\ulcorner\underline{T}\text{`}\ulcorner\underline{A}\text{'}$ are true in $\underline{MClos}(\underline{FP}(\underline{F}))$, but that fact cannot be reported in that interpretation, because $\ulcorner\underline{T}\text{`}\underline{A}\text{'}$ and $\ulcorner\underline{T}\text{`}\ulcorner\underline{A}\text{'}$ both belong to $\overline{\underline{FP}(\underline{F})_1}$, so neither $\underline{T}\text{`}\ulcorner\underline{T}\text{`}\underline{A}\text{''}$ nor $\underline{T}\text{`}\ulcorner\underline{T}\text{`}\ulcorner\underline{A}\text{''}$ are true. Kripke suggests a further strategy: since in fact the interpretation $\underline{MClos}(\underline{FP}(\underline{F}))$ is a fully interpreted model, we can define a standard Tarskian truth predicate for it. Let that predicate be $\mathbf{T}$: clearly the extension of $\mathbf{T}$ will not coincide with that of $\underline{T}$ in $\underline{MClos}(\underline{FP}(\underline{F}))$, since $\ulcorner\underline{T}\text{`}\underline{A}\text{'}$, for instance, will belong to the former and not the latter. So in turn $\mathbf{T}$ can express

facts about $\underline{MClos}(\underline{FP}(\underline{F}))$ that were not expressible in it, just as it in turn could express facts about $\underline{FP}(\underline{F})$ that were inexpressible there. One odd thing about **T** concerns its treatment of the Liar and Truth-Teller. If $\underline{b}$ is the Truth-Teller sentence $\underline{Tb}$, and is without value in $\underline{FP}(\underline{F})$, then $\underline{Tb} \in \overline{\underline{FP}(\underline{F})}_1$. Hence $\sim\underline{Tb}$ is true in $\underline{MClos}(\underline{FP}(\underline{F}))$, and is therefore put in the extension of **T**. Consequently **T**'$\sim\underline{Tb}$' is true in the metalanguage. On the other hand, if $\underline{a}$ is the Liar sentence $\sim\underline{Ta}$, $\sim\underline{Ta} \in \overline{\underline{FP}(\underline{F})}_1$, $\sim\underline{Ta}$ is true in $\underline{MClos}(\underline{FP}(\underline{F}))$ and **T**'$\sim\underline{Ta}$' is true in the metalanguage. This means that the sentences expressing the truth of the negation of the Truth-Teller and the truth of the Liar are both true in the metalanguage, an odd quirk of the theory.

On the whole, then, it seems that either closing off a fixed point, or going on to introduce a Tarski truth-predicate, while enabling us to express certain features of the construction which are inexpressible there, cannot represent our primary uses of 'true'. Kripke suggests rather that the primitive use of 'true' is given by the basic construction, and that these further manoeuvres represent the result of further reflection about the primitive use. As I have already explained, I am sympathetic to the general position that not all intuitions about truth have to be reflected in a theory of our use of 'true'. Rather, our first responsibility is to account for our most natural uses of 'true'. Accordingly , I shall not discuss these additions, but will return to the basic construction to see how well it meets this responsibility.

The first thing to notice is that it admirably meets the 'riskiness' requirement. For suppose someone asserts a sentence which is rendered as $(\underline{x})(\underline{Px} \supset \sim\underline{Tx})$, and it unfortunately turns out that the extension of $\underline{P}$ is simply $\{(\underline{x})(\underline{Px} \supset \sim\underline{Tx})\}$. Then there cannot be a fixed point in which $(\underline{x})(\underline{Px} \supset \sim\underline{Tx})$ is true or false, for if $(\underline{x})(\underline{Px} \supset \sim\underline{Tx})$ is true in an interpretation $\underline{M}\langle\underline{S}_1, \underline{S}_2\rangle$, it has to belong to $\underline{S}_2$, by the valuation

rules, and if it is false in $\underline{M}\langle\underline{S}_1', \underline{S}_2'\rangle$, it has to belong in $\underline{S}_1$. So in every fixed point, it is neither true nor false, and hence is paradoxical. On the other hand, if the extension of $\underline{P}$ contained various harmless sentences, $(\underline{x})(\underline{Px} \supset \sim \underline{Tx})$ could readily be given a value.

Other prominent advantages are the extremely plausible distinction between mere ungroundedness and paradoxicality, which reflects the feeling that Liars are seriously worse than Truth-Tellers, and the satisfying match between the formal construction and the account of learning how to use 'true'. Given all these useful characteristics, the obvious question is "How does the formal theory relate to natural language?" and the equally obvious answer is to find the fixed point which best fits our intuitions: that is the formal interpretation of 'true'. However, things turn out not to be so straightforward. Kripke even goes so far as to suggest that there may simply be no fact of the matter as to which is the valuation scheme, or the fixed point. He does not elaborate on this agnosticism; what I want to do is to explore some of the problems which may have given rise to it, which are of two basic types. First, it turns out that no valuation scheme or fixed point has a monopoly of the various advantages, so that choice between them tends to be a matter of theoretical preference rather than obvious fact. Second, there is also some tension between the idea that some specific fixed point captures the extension of truth, and the various overall advantages that the formal construction presents.

Taking the various valuation schemes and fixed points first, we can see that whatever scheme we may settle for, there are only two serious candidates for the choice of fixed point, namely the minimal fixed point and the greatest intrinsic fixed point, so these are the only ones I shall consider. I shall now discuss some (though certainly not all) of the relative merits of the valuation schemes.

Weak Kleene

While differences between the other two schemes are fairly subtle, the Weak Kleene scheme is radically different. Its principal merit is its ability to contain not only a truth predicate, but also a falsity, and a neither-truth-nor-falsity one, too.[37] On the other hand, it purchases this expressive power at a serious cost, one which effectively rules it out as a contender.

First, we lose expressibility of truth conditions, since in Weak Kleene, typical Strong Kleene connectives are inexpressible: any Weak compound with a component whose value is u has value u too, whereas the Strong disjunction t u has value t. On the other hand, Weak disjunction is definable using Strong connectives:

$$\underline{A} \text{ weak disj. } \underline{B} \equiv \underline{A} \vee \underline{B} \ \&(((\sim \underline{A} \vee \underline{B}) \& (\underline{A} \vee \sim \underline{B})) \vee (\sim \underline{A} \vee \sim \underline{B})). \tag{4.9}$$

Second, in terms of accounting for our intuitions, typical Strong truth-values seem much more plausible. As the discussion of Martin's work showed, Weak existential quantification, for instance, seems to bear no resemblance to our intuitive assessment of "Something Enid said is true" when she may have said something true, and something neither true nor false.

Strong Kleene

By comparison with Weak Kleene, Strong suffers from its inability to express non-truth. Suppose, for instance, we wanted to express 'neither true nor false' by a predicate $\underline{N}$. Then the sentence $\underline{d}$,

$$\underline{v}(\underline{d}) = \sim \underline{T}\underline{d} \vee \underline{N}\underline{d} \tag{4.10}$$

causes the customary problem. Expressing for the moment the extension

---

[37] See the discussion of MW in the chapter on Martin.

and anti-extension in a fixed point of T and N by T(1), T(2), N(1), and N(2), we can see that any sentence falls into only one of T(1), T(2), and N(1). But if we ask which of those d falls into, we get the usual problem: if in T(1), then Td and ~Nd are true, so ~TdvNd is false; if in T(2) then ~Td and hence ~TdvNd are true; and if in N(1), Nd and hence ~TdvNd is true. In no case can we have a fixed point.

We can see that in the general case not even N alone can be represented. For suppose there is a predicate G, for which the sentence e is undefined:

$$v(e)=GevNe. \tag{4.11}$$

Suppose we formed a fixed point for N: if e belongs to N(1), Ne and GevNe are true, so e cannot belong to N(1). On the other hand, if e belongs to N(2), Ge is undefined, Ne is false, so GevNe is undefined and hence should belong to N(1). However, in the case where the basic language is fully defined, N can be represented by the trivial $\langle \mathbf{A}, D \rangle$.

However, the theoretical approach Kripke has adopted minimises this failure. For if we suppose that the primitive notions are truth and falsity, then lack of truth-value can be treated at a later stage, as a meta-theoretical concept . In fact, if lack of truth-value was treated as a primitive notion, the appropriate construction would have to have four values, truth, falsity, neither, and none of the above, in order for monotonicity to hold. In this case we would have three expressible values, and one inexpressible one.

This is perhaps an appropriate point at which to make an observation about truth-values in the basic model. Kripke starts with a standard three-valued model, and proves the existence of a representative truth predicate for the model. However, not all the concepts he goes on to define are intuitively satisfactory in the general case. For example, if some sentence Fa is neither true nor false in the model, it will never be put in the extension or anti-

extension of T, and hence will be, by definition, paradoxical, which is rather surprising, since it may just represent "Fred is bald". The requirement for these definitions to be satisfactory is that the basic model be fully defined, with no truth-value gaps. As such, the full theory is really a three-valued truth theory for a two-valued language.

Returning to the merits of the Strong Kleene scheme, we find that the comparison with the supervaluational approach is harder. The main advantage of the Strong version is a certain naturalness; the main disadvantage is some surprising inexpressibilities.

The naturalness can best be seen by considering the concept of groundedness. Strictly, groundedness is ambiguous: its definition relies on a given valuation scheme, but there is little doubt that the intuitive notion is closely tied to the Strong Kleene version. If a sentence is Strongly grounded, then we can easily trace the route through which it is connected to "the facts", and by comparison, Weakly grounded sentences seem to need too strong a connection, while supervaluationally grounded ones may end up at some tenuous kinds of facts. To see this latter claim, consider any sentence of the form $\underline{A} \vee \tilde{} \underline{A}$; since every such sentence is true in the minimal fixed point, they are all grounded, even if A is itself paradoxical. As a result, 'grounded' is now almost universally used to mean Strongly grounded.

What is surprising about the Strong Kleene scheme, however, is its treatment of certain logical and semantic laws. For instance, if Sent is a predicate true of sentences, then the sentence

$$(\underline{x})(\underline{\mathrm{Sent}}(\underline{x}) \supset \underline{\mathrm{T}}\underline{x} \vee \tilde{} \underline{\mathrm{T}}\underline{x}) \qquad (4.12)$$

might be expected to be true (as a logical law) or false (because some sentences are neither true nor false). However, in the Strong scheme, it is always ungrounded, because one requirement for it to be true is that it be true. If all other sentences have intrinsic truth-values, then it is true in the greatest intrinsic fixed point, but if there is,

say, a Truth-Teller in the model not even that is true. Finally, in the presence of a Liar, this apparent logical law is paradoxical, because one of its instances is.[38]

Supervaluations

To some extent these problems are overcome in the supervaluation scheme. In supervaluations, as we saw before, the logical laws are always true. Since in every classical valuation, for every sentence $\underline{x}$, $\underline{x}$ is either in the extension of $\underline{T}$ or out of it, either $\underline{Tx}$ or $\sim\underline{Tx}$ will be true. Hence, in the supervaluation (4.12) is true, even in the minimal fixed point. If we restrict the classical valuations in an intuitively plausible way, we can make further principles true. For instance, as things stand, a classical valuation can assign both $\underline{A}$ and $\sim\underline{A}$ to the extension of $\underline{T}$. If we impose the restriction that only classical valuations which assign consistent extensions to $\underline{T}$ are admitted, the following sentence is true:

$(\underline{x})\sim(\underline{Tx} \,\&\, \underline{Tneg}(\underline{x}))$, where $\underline{neg}(\underline{x})$ is the negation of $\underline{x}$. (4.13)

And if we insist that classical valuations assign maximal consistent extensions to $\underline{T}$ (or consistent extensions and anti-extensions) we can obtain

$(\underline{x})(\underline{Sent}(\underline{x}) \supset (\underline{Tx} \vee \underline{Tneg}(\underline{x})))$. (4.14)

though of course this does not mean that every sentence is true or false, as it might appear to do.

Thus we actually have a system of different supervaluation schemes, of which the most interesting are the basic one, and the last-mentioned,

---

[38]Note that it can never be false, because either the extension and anti-extension of truth together exhaust $\underline{L}$, or they do not; in the first case it is true, in the second, neither.

the maximal-consistent (m-c) one, for there seems to be little reason to stop at any half-way house.

However, not even the m-c supervaluation scheme can preserve our intuitions about every sentence. Consider

$$(\underline{x})(\underline{\text{Sent}}(\underline{x})\supset(\underline{Tx}\equiv\underline{T}(\underline{\text{true}}(\underline{x})))) \qquad (4.15)$$

where $\underline{\text{true}}(\underline{x})$ is the result of predicating $\underline{T}$ of $\underline{x}$.

Since nothing prevents a classical valuation from assigning a sentence $\underline{A}$ to the extension of $\underline{T}$, and omitting $\underline{T}$'$\underline{A}$', (4.15) will not be true in the minimal fixed point, nor, in the presence of a Liar, will it be true in the greatest intrinsic fixed point. We could impose further restrictions on the set of classical valuations, but at this point it is clear that to preserve our intuitions we will need to impose constraints which will themselves begin to resemble a complex theory of truth. Thus it will no longer be the case that our theory of truth starts from a simple set of basic assumptions; its basis will be as complex as the theory itself. Moreover, the sorts of considerations we will have to adduce are precisely the kind of thing we would like the theory to tell us.

Thus we can see that although the supervaluation schemes get the right truth-value on some sentences on which the Strong scheme gets the wrong one, they do not seem to get to the heart of the matter, and suffer from a certain loss of naturalness and plausibility. Choosing between these virtues is not obviously referrable to some fact of the matter, and may easily lead to scepticism about the existence of a real extension for true.

This scepticism is compounded by the second kind of problem mentioned above, where the claim that such-and-such a fixed point gives the extension of 'true' has to be reconciled with other advantages of the construction, especially with the ungrounded/paradoxical distinction and the analogy with the instructions for learning the use of 'true'.

The most serious conflict arises if a greatest intrinsic fixed point is proposed for the extension of 'true'. In that case, both the other advantages seem unavailable: the former, because then it is hard to see what interest the ungrounded/ paradoxical distinction could have for us, and the latter because no plausible instructions could help us reach it. If, after all, we are genuinely using one fixed point , what can be the significance of a distinction which refers to another, lesser one? And what could we make of the instruction which said, in effect: Construct all the fixed points, and then find the largest one which does not disagree with any other?

Even if we pick a minimal fixed point, some shades of these problems remain. The learning instructions are secure, of course, but for the distinction, again it is not clear what interest we should have in what might happen if we used other instructions. The main answer to this is that we readily understand a distinction, among those sentences to which our process has not awarded a value, between those that would, and those that would not, wreck that process if we tried to add them in a certain way. Nevertheless, this is an unusual kind of modality; it is not, like riskiness, a matter of how the world is, but rather a matter of how we might use our language.

These considerations taken together may inspire the following view of the construction and its relation to natural language. 'True' may not be a very well-defined concept, and the best-defined part of it might correspond to the general application of the Kripke construction, _i.e._ taking for extension and anti-extension of _T_ at one level the true and false at the previous level, and continuing to a fixed point. What might be ill-defined would be such things as the valuation scheme used to determine the true and the false at a level, and the fixed point we are aiming at. These open questions correspond to a group of general principles like maximising the award of truth-values, avoiding

arbitrariness, sticking to groundedness, and so forth. Thus it would not be the case that any particular fixed point represented our intuitions; rather, the whole structure would do so. Luckily, for most sentences, this vagueness would not matter, since there would be agreement on them across all or at least most fixed points.

This may in fact be the way things are: certainly it is possible that our intuitions about truth may disagree over certain examples. However, it seems a little soon to give up the search for a definite theory of the extension of 'true'. We can regard Kripke as having done two things to this end. First, he has established certain desiderata for such theories: the ability to handle riskiness, the importance of defining truth in terms of a model, and the ability to provide definitions of such notions as groundedness, paradoxicality and so forth. Second, he has provided a challenge for others: to produce a better theory which meets these requirements.

Chapter 5

Burge and Indexical Truth Predicates

It should be clear by now that the foregoing three-valued theories share some common deficiencies. In general, they tend to have unsatisfactory accounts of the Strengthened Liar, and a widespread inability to express various semantic facts about sentences which are neither true nor false, namely their non-truth and non-falsity.

These defects are essentially due to their common failure to express exclusion negation. We may share reasonable doubts whether any common use of negation expressions in English should be expressed by exclusion negation, but the function is comprehensible, and expressible in English, so that it is a deficiency in a theory if it fails to express it. Furthermore, something like exclusion negation is obviously responsible for the inference we find plausible, from "The Strengthened Liar is neither true nor false" to "The Strengthened Liar is not true", so that failure to express it is part of the dissatisfaction with the treatment of the Strengthened Liar in these theories. But, of course, none of them could contain it: as we saw with MW and the Kripke construction, exclusion negation prevents the formation of truth-representing $\underline{T}$ predicates, and the most obvious way of intoducing it into a supervaluation would be to add to the notion of a saturated set the condition that, for every $\underline{A} \notin \underline{G}$, $-\underline{A} \in \underline{G}$: this has the effect, when we take a supervaluation, that for every sentence $\underline{A}$ which is not true, $-\underline{A}$ is true. However, if we had a Liar sentence $\underline{U}$, such that $\langle \underline{U}, \underline{TU} \rangle$ and $\langle \underline{TU}, \underline{U} \rangle$ belong to $\underline{N}$, we could not form a consistent saturated set:

either $U \in G$, in which case both TU and -TU do too, or $U \notin G$, in which case TU does not belong either, so -TU must, and so both U and TU do. In either case, we can obtain no supervaluation, and the construction collapses. Another way of introducing exclusion negation would be to do so after taking supervaluations, defining it by the usual three-valued valuation. Then for any sentence A, either A or -A would be true. In particular, either TU or -TU would be true, and contradiction would follow as usual. Unlike the other Liar sentences, however, we could not disarm this by saying that this showed that TU and -TU had false presuppositions, and so must be neither true nor false, since one of them must be true.

Moreover, we can see that the possibility of representing truth, even if not non-truth, was gained not by moving to a three-valued language in the face of the Tarskian challenge: it was, rather, at the cost of expressibility of certain truth-functions. We can easily show that a Liar-type paradox is constructible in any multi-valued language if arbitrary truth-functions are expressible. Suppose we have an n-valued language with the truth-function $\neg$:

$$\underline{val}_{I,\alpha}(\neg A)=1 \text{ iff } \underline{val}_{I,\alpha}(A)\neq 1 \qquad (5.1)$$
$$=n \text{ otherwise}$$

Suppose in some model, the name a denotes as follows:

$$v(a)=\neg Ta \qquad (5.2)$$

Then if 1 is the preferred value, there can be no representing truth predicate in that language for that model. For suppose there were, and that $\neg$Ta belonged to the "extension" of T,[39] then Ta would be

---

[39]The "extension" of a predicate F would be analogous to the extension of a predicate in either two- or three-valued theories, and is roughly the set of denotations which, when F is predicated of the corresponding terms, gives the preferred value for the resultant sentence.

true. On the other hand, if ¬Ta did not belong to the "extension" of T, Ta would have some value i≠1, and ¬Ta would be true. In either case, the extension of T is not the same as the set of true sentences.

We saw in MW that further weakening of the truth-functional expressive power of a language leads to greater expressibility of semantic facts, but the latter cannot be got free. On the other hand, if we are bound to lose one kind of expressibility to gain the other, as it seems we are, we may well ask why we should bother to go to three values, and not look at restrictions in two-valued languages. It is open to question what advantages three values have at all.

In Kripke's construction, the main advantages of three values seem to be that they preserve a certain naturalness, and make the "book-keeping" easier, but there does seem to be some sense in which a third value is redundant. The point is this: given the extension of T alone, we can construct its anti-extension and hence know the remainder, because the anti-extension just contains all the negations of the members of the extension. Thus of the three values, true, false and neither, we only need to know what falls under the first, and we know what falls in each of the others. There really only seem to be two significant categories, true and not true. On the other hand, we could not apply a straightforward two-valued scheme to get a hierarchy like Kripke's which would reach a fixed point, since the ordinary valuations cannot give a monotonic function like ♦. To see this, we need only consider a model in which there is a Liar sentence, and compare the sets of sentences true in the model when the extension of T does and does not contain the Liar sentence: the set of sentence true when the extension of T contains the Liar will not contain it, and the corresponding set when it is not in the extension of T will contain it. We need to distinguish, for book-keeping, among the non-true sentences, those which have a true negation, and those which do not, and adopting the three-

valued scheme with an extension and an anti-extension for T is one way of doing this.

With Martin and van Fraassen, things are slightly different because the initial motivation for a three-valued system is to express various semantic relations which are thought to require three values. I am not going to comment extensively on whether the proposed systems are satisfactory with respect to these initial inspirations: in discussing Martin's theory I argued that his system was not a happy theory of semantic correctness, and even with van Fraassen's, which is less unsatisfactory, there is a problem that truth-value gaps arise without failure of presupposition. In general, it seems that even if we agree that sentences of certain kinds should lack truth-value, it is unclear that the theory of such sentences should be a logic in which lack of truth-value is a sign that a given sentence belongs to the relevant class.

If we disregard the prior motivation for the three values, and look just at the handling of the truth predicate, Martin's theory can be considered as a version of Kripke's. With van Fraassen's, some interesting points appear, concerning bivalence and trivalence, and their representations. In supervaluation theory, $A \vee \sim A$ is a logical truth, and hence does not represent the principle of bivalence, because in a three-valued language it could still be true, even if A is neither true nor false. Given some decisions on the treatment of T, a far better candidate is $TA \vee T\sim A$. Discussing this sentence, van Fraassen argues that we cannot adopt the valuation of TA whereby it is given the same value as A, because this would mean that the sentence could not express the non-truth of A, since it would be neither true nor false if A was also. Instead, we should treat TA as false if A is not true. In that case, $TA \vee T\sim A$ does seem to express bivalence, since it is true if A is true or false, and false if A is neither.

The sentence that seems to express trivalence is

$$TA \vee T\sim A \vee (\sim TA \& \sim T\sim A) \qquad (5.3)$$

which, like $A \vee \sim A$, is a logical truth. But this means that any sentence, A, whatever its value, makes (5.3) true. An alternative version is:

$$TTA \vee TT\sim A \vee T(\sim TA \& \sim T\sim A) \qquad (5.4)$$

which bears the same relationship to (5.3) as $TA \vee T\sim A$ does to $A \vee \sim A$. The trouble is, of course, that the Strengthened Liar, though neither true nor false, is such that $\sim TY \& \sim T\sim Y$ is neither true nor false. Moreover, for Y, (5.4) is false. Thus the Strengthened Liar presents the same problem with respect to trivalence that a "neither" sentence does for bivalence. The response to this, however, is not to introduce a fourth value. But if that is so, why should we have introduced a third value either?

It might be thought that supervaluational language must be a three-valued language, but we can readily introduce a two-valued supervaluation:

> Definition 1: The function S_k' is a two-valued supervaluation induced by a set k of classical valuations iff $S_k'(A)=t$ iff $v(A)=t$, for all $v \in k$
> $=f$ otherwise

In such supervaluations, all the logical truths will still be true, of course, so $A \vee \sim A$ will be true, but $TA \vee T\sim A$ need not be.[40] Suppose we still introduce T by necessitation, and construct a saturated set G to generate supervaluations. Then, just as before, a sentence A can be such that $A \notin G$, $\sim A \notin G$, $TA \notin G$, $T\sim A \notin G$. Then $TA \vee T\sim A$ is false as

---

[40] Herzberger discusses this kind of supervaluation in an unpublished paper "Supervaluations without Truth-Value Gaps".

before, and ~TA does not express the non-truth of A. Parenthetically, it is interesting to note that in two-valued supervaluations, sentences and their negations can both be not true, so negation is no longer truth-functional. Anyway, clearly the language has an ineffability problem. Nevertheless it is no worse than for three-valued languages.

A final problem for three-valued solutions to the Liar is that it is not even clear that they offer an intuitively satisfactory version of the Ordinary Liar, let alone the Strengthened Liar. It is true that if the Liar is true, there is a contradiction, and if it is false, there is one too. It does not necessarily follow that there is no contradiction in calling it neither true nor false. For the Liar sentence says that it is false, and if it is supposed to be neither true nor false, it is presumably not false, so in turn it must be false (since it says something which is not so). And off we go in circles again. This argument, which might be called the strengthening of the Ordinary Liar, is at least as plausible as any of the intuitions about the Strengthened Liar, and the means of meeting it in the different theories are essentially the same: we are not allowed to say of something that is neither true nor false, that it is thereby not false (rather than not true, as in the Strengthened Liar). Thus these theories leave the paradoxes very much at large, not even restraining the Ordinary Liar.

One author who has argued this point strongly is Tyler Burge, in his paper "Semantical Paradox". In particular Burge thinks that none of the approaches discussed above can satisfactorily account for a particular sequence of judgments which is characteristic of Liar-paradoxical sentences, exhibited in the following story.[41]

A student writes on a board in a room the sentence 'No

---

[41] This is a simplified version of Burge's example c, p. 179. Burge in turn derived it from A. N. Prior, "On a Family of Paradoxes".

> sentence written on the board in room 10 is now true': Unfortunately he is in room 10, and that is the only sentence written on the board. A passer-by, reflecting on the student's predicament, realizes that the sentence cannot be true, and says "No sentence written on the board in room 10 is now true." The student, hearing him, says "That is just what I wrote".

Schematically, we have here a) an assertion of a paradoxical sentence, b)the judgment that that assertion is not true, and hence c) the judgment that since that is what it says, it must be true. According to Burge, it is a central part of a theory of truth that it account for this reasoning, and he offers an attempt to do so. Roughly, he thinks that, given a), b) is justified by certain pragmatic features of truth-ascription, and that given a) and b), c) is justified by a change of the extension of the truth-predicate. The main part of his theory involves a choice of constructions based on a Tarski hierarchy of truth-predicates, which I shall describe. Since Tarski himself, and most writers following him, thought that a Tarskian theory would be most unnatural, it is of the greatest interest to find a proposal for which is claimed the greatest possible concurrence with our intuitions. Of course, one of the elements which Tarski thought unintuitive will remain, namely the elimination of vagueness and the full definition of all predicates. Setting this aside, still, few have thought a Tarskian theory could be a very natural account of our use of 'true', as I outlined in the Introduction.

Basically, Burge's theory consists of two parts, a set of constructions which use a hierarchy of predicates $\underline{\text{true}}_i$ and which define the extension of each of these predicates, and a set of priciples governing how ordinary sentences in English, using 'true', match sentences in the construction using '$\underline{\text{true}}_i$. I shall discuss Burge's three constructions C1, C2 and C3 first, and then the principles governing the translation from English into the constructions.

Besides the hierarchical truth-predicates there is another

hierarchy of predicates $\underline{P}_i$, for $\underline{\text{pathological}}_i$. The simplest construction to consider is C1 which quite closely follows Tarski's definition, except that it seems more useful to think of the hierarchy not as a sequence of languages, but as an infinite collection of truth-predicates $\underline{Tr}_i$, for a single base language $\underline{L}_0$, i.e. $\underline{L}=\underline{L}_0 \cup \{\underline{P}_i \mid 0<i<\infty\} \cup \{\underline{Tr}_i \mid 0<i<\infty\}$. Burge is not specific about this, but I take it to be his intent. He, properly, gives his constructions in terms of a satisfaction predicate, but for ease of exposition I shall just use a truth predicate.

C1:

1. $\underline{P}_i$'$\underline{A}$' iff $\underline{A}$ contains predicative occurrences of $\underline{P}_k$ or $\underline{T}_k$, for $k \geqslant i$.
2. $\underline{P}_i$'$\underline{A}$'$\supset\sim\underline{Tr}_i$'$\underline{A}$'
3. $\sim\underline{P}_i$'$\underline{A}$'$\supset$($\underline{Tr}_i$'$\sim\underline{A}$'$\equiv\sim\underline{Tr}_i$'$\underline{A}$')
4. ($\sim\underline{P}_i$'$\underline{A}$'&$\sim\underline{P}_i$'$\underline{B}$')$\supset$($\underline{Tr}_i$'$\underline{A}\&\underline{B}$'$\equiv\underline{Tr}_i$'$\underline{A}$'&$\underline{Tr}_i$'$\underline{B}$')
5. $\sim\underline{P}_i$'$\underline{A}$' ($\underline{Tr}_i$'$\underline{A}$'$\equiv\underline{A}$)[42]

As Burge points out (p.186), C1 agrees with Tarski about such sentences as 'This sentence is true', which cannot have truth_i conditions, but it goes further in allowing it to have truth conditions, for k>i. Burge feels this is justified by the fact that most paradoxical sentences do not seem to be seriously deviant; if they were, we would be justified in ruling them out entirely. Rather they are merely unfortunate, so we should be able to make semantic judgments about them, which C1 permits.

Two useful theorems follow:

Theorem 2: $\underline{Tr}_i$'$\underline{A}$'$\supset\underline{Tr}_{i+1}$'$\underline{A}$'

---

[42]Using satisfaction, we could lay down axioms governing $\underline{Tr}_i$'$(\underline{x})\underline{A}$'; essentially the rule follows the standard analogy with conjunction.

Proof: Suppose $\underline{Tr_i}\,'\underline{A}'$. Then $\tilde{}\underline{P_i}\,'\underline{A}'$, but hence $\tilde{}\underline{P_{i+1}}\,'\underline{A}'$. Since $\underline{Tr_i}\,'\underline{A}'\equiv\underline{A}$ and $\underline{Tr_{i+1}}\,'\underline{A}'\equiv\underline{A}$ both hold, $\underline{Tr_i}\,'\underline{A}'\equiv\underline{Tr_{i+1}}\,'\underline{A}'$.

Theorem 3: $\underline{Tr_i}\,'\underline{A}'>\underline{Tr_k}\,'\underline{Tr_i}\,'\underline{A}''$, $k>i$.

Proof: Suppose $\underline{Tr_i}\,'\underline{A}'$. Then $\tilde{}\underline{P_i}\,'\underline{A}'$, and hence for the highest predicative occurrence of $\underline{Tr_j}$ or $\underline{P_j}$ in $\underline{A}$, $j<i$. Hence in $\underline{Tr_i}\,'\underline{A}'$, the highest such occurrence is i. Consequently $\tilde{}\underline{P_k}\,'\underline{Tr_i}\,'\underline{A}''$, so if $\underline{Tr_i}\,'\underline{A}'$, $\underline{Tr_k}\,'\underline{Tr_i}\,'\underline{A}''$.

So once a sentence $\underline{A}$ is true_i, it is true_j, for all $j>i$ and so is the sentence '$\underline{A}$ is true_i'.

C2 and C3 loosen the restrictions on pathologicality in the following ways. First, C2 permits valid inferences from true sentences to be true too. Thus if 'Snow is white' is true_i, so is 'Snow is white or $\underline{A}$' even if $\underline{A}$ contains $\underline{Tr_k}$, $k>i$. C3 goes further and permits iteration of truth predicates: if 'Snow is white' is true_i so is '"Snow is white" is true_i' and '"Snow is white" is true_i' is true_i', and so on. However, I shall not go into the technical details of these constructions.

Any of C1, C2 and C3, then, gives axioms governing the predicates $\underline{Tr_i}$. How do these constructions represent the way we use 'true' in English? In the first place, occurrences of 'true' are interpreted as being implicitly occurrences of 'true_i': the next question is which uses of 'true' get assigned which indices. Burge offers two principles which govern this assignment, the Principle of Verity and the Principle of Justice.

The Principle of Verity is: "subscripts on 'true' are assigned ceteris paribus so as to maximize the interpreter's ability to give a sentence truth conditions by way of a truth schema" (p. 193). In particular, in any given case we should assign the lowest subscript compatible with Verity. We can consider an example. Suppose we have the sentence 'Everything Descartes said that does not concern mechanics

is true' and we want to know what index to assign to the occurrence of 'true'. The first point is that that assignment will be affected by the choice of construction, since whether under a given assignment a given sentence is pathological or not may vary with C1, C2 and C3. Suppose, then, that we are using C1, and suppose further that we have assigned indices to all the occurrences of 'true' in all Descartes' utterances on topics other than mechanics. Suppose the highest such is 93. Then we can satisfy Verity by any assignment greater than 93 for the occurrence in question: 93 or less would result in our sentence being pathological, and hence not having an a appropriate truth schema. Thus we should assign index 94.

The Principle of Justice is intended to override Verity in cases where two or more truth ascriptions fall in one another's scope, as in the example where Aristotle says 'What Plato is now saying is not true' and Plato says simultaneously, 'What Aristotle is saying is not true'. We have three options open: to assign Aristotle's use of 'true' a higher index than Plato's; to do the reverse; and to assign both the same index.

In the first case, we have two sentences:

$$\begin{array}{ll} \underline{a}: & \underline{p} \text{ is not true}_i. \\ \underline{p}: & \underline{a} \text{ is not true}_j. \quad j<i \end{array} \tag{5.5}$$

Under this assignment, $\underline{a}$ and $\underline{p}$ are pathological$_j$ and not true$_j$, but $\underline{p}$ is not pathological$_i$, and has a truth$_i$ schema, which is $\underline{Tr_i p} \equiv \sim \underline{Tr_j a}$. Hence $\underline{p}$ is true$_i$, and $\underline{a}$ is not true$_i$, because of the biconditional $\underline{Tr_i a} \equiv \sim \underline{Tr_i p}$. Thus Plato's remark succeeds, and Aristotle's fails. In the second case the reverse happens, and in the third we have the following:

$$\begin{array}{ll} \underline{a}: & \underline{p} \text{ is not true}_i. \\ \underline{p}: & \underline{a} \text{ is not true}_i. \end{array} \tag{5.6}$$

In this case, no schemata can be properly applied, and both sentences

are pathological$_i$. Thus Verity would justify either of the other two assignments because in each at least one sentence gets evaluated by a truth schema, but choosing one or the other seems arbitrary, so Justice requires the third assignment.

Unfortunately, these remarks are inadequate to give us a thorough indication of how indices are to be assigned. I shall discuss just one example which Burge analyses, and suggest another analysis which seems equally well justified by the principles. Burge takes two sentences, numbered iii) and iv), (p. 195)

iii) Mitchell is innocent and iv) is not true. (5.7)
iv) iii) is not true.

Burge suggests that since Mitchell is not innocent, we can say that iv) is true. The question is how this can be rendered using the indexical predicates. I shall give the version using C1: C2 and C3 are given in a footnote.

Burge starts by assigning i to both occurrences of 'true'. Then we can argue as follows. Since iv) contains 'true$_i$', it is pathological$_i$ and not true$_i$. Similarly, since iii) contains 'true$_i$', it is pathological$_i$ and not true$_i$. However, iv) is not pathological$_{i+1}$, so we have the biconditional 'iv) is true$_{i+1}$ iff iii) is not true$_i$', and hence iv) is true$_{i+1}$. So we can model the truth of iv), though it is worth noticing that no semantic evaluation by truth schema of iii) took place.

Now, Burge suggests, we naturally tend to go on: since iv) is true, the second conjunct of iii) cannot be, since it says iv) is not true. In C1 we cannot model this. For we find the Tarski biconditional '"iv) is not true$_i$" is true$_j$ iff iv) is not true$_i$' for any j>i, and since iv) is not true$_i$,'iv) is not true$_i$' is true$_j$. Burge suggests that the only way C1 (and in fact C2) could get this right would be for us to say that there is an equivocation over levels somewhere, but this is not

a good move![43] The moral of this, Burge suggests, is that C3 is better than C2 or C1 at rendering our intuitions correctly. I want to point up some other morals.

I assume that Burge gave the assignment i to both occurrences of 'true' by some application of the Principle of Justice, on the grounds that each of iii) and iv) attempts to evaluate the other, and neither should be favoured, just as in the example of p and a. However, the examples are not exactly parallel.

Consider the assignment in which 'true' in iii) gets i+1 and in iv) gets i. Then in C1, as before, iv) is $true_{i+1}$. Since this is just what the second conjunct of iii) denies, it, in turn, is not $true_{i+2}$. Consequently, under this assignment even C1 models our intuitions perfectly. On the other hand, in the reverse assignment, C1 is incapable of rendering our intuitions, and depending on the exact levels, even C2 and C3 will have trouble.

So the situation is as follows. We have three different assignments, one even-handed and two not. Of these three, two get one truth-value right (that of iv)) and one wrong (that of the second conjunct of iii)), while the other gets both right. Consequently the situation is very different from the Plato and Aristotle case, where one had to choose between two which were apparently better than a third, and Justice was called in to pick the third. Here, there is every reason to pick the single, best assignment.

This leads to two subsidiary morals, and one major. First, it is

---

[43] In C2, an evaluation of iii) takes place, assuming that the falsity of 'Mitchell is innocent' has been established by level i. In that case iii) is genuinely not $true_i$, rather than by default, as it were. Otherwise C2 works like C1. In C3, iii) is not $true_i$ as for C2, and hence iv) is $true_i$, and the second conjunct of iii) not $true_i$, so C3 works fine.

obvious that Justice is not sufficiently well explicated for us to know how to apply it on any given occasion. Second, until we do know exactly how to apply it, we cannot really draw any morals about the relative merits of C1, C2 and C3.

Most important of all, however, we realise that both Verity and Justice are the wrong kind of rule to give for the role they have to play. When we use the predicate 'true', the index which should be assigned to each occurrence of its use depends on the context of utterance, and in this sense depends on pragmatic features. On the other hand, it does not depend on pragmatic features such as conversational convention, or speaker desires and intentions. Once the facts of the context of utterance are fixed, so is the assignment of level. This is a slight idealisation, but it is one whose rejection calls for much more argument than Burge offers.[44] Technically, what seems to be needed is the following. The facts of the matter determine a model of a certain kind, in which the interpretations of some predicates contain occurrences of Tr with no subscript, corresponding to ordinary language facts like Fred's saying '"Snow is white" is true'. What Burge must provide is a systematic way of transforming a model of that kind into a model in which every such occurrence of Tr gets a subscript, and then show that the extensions of each of the Tr_i can be satisfactorily defined. Until we have some such systematic account, we cannot really say how his theory treats any sentence at all, except that Verity and Justice do seem to embody plausible constraints on such an account.

---

[44] It is an idealisation because there may be pragmatic conventions which enable us to restrict our truth predicate in certain ways, just as we clearly use quantifiers in restricted ways on many occasions. However, these conventions should not be confused with semantics for the truth-predicate: indeed, until we have the semantics, we cannot go on to discuss the precise effect of pragmatic phenomena.

The task of constructing such an account is not trivial. C1 provides the easiest version, because there we decide on levels independently of truth-values. That is, assigning a level to the non-specified truth-predicate in $\underline{Tr}'\underline{Tr}_2\underline{a} \vee \underline{Tr}_3\underline{b}'$ is straightforward in C1, but in C2 and C3 we have to know the truth-values of $\underline{Tr}_2\underline{a}$ and $\underline{Tr}_3\underline{b}$ before making the assignment. I shall discuss some of the difficulties of this process below.

Supposing that an account of the right kind could be produced, how does Burge's theory describe what is happening in a), b) and c) above? We can model a) 'No sentence written on the board in room 10 is now true', b) 'No sentence written on the board in room 10 is now true', and c) 'The sentence written on the board is true', with some loss of generality but otherwise harmlessly as:

$$\begin{aligned} &\underline{a}': \ \underline{a}' \text{ is not true}_i. \\ &\underline{b}': \ \underline{a}' \text{ is not true}_j. \\ &\underline{c}': \ \underline{a}' \text{ is true}_k. \end{aligned} \tag{5.8}$$

How should i, j and k be chosen to model our intuitions? Burge's suggestion is the following. First, we have no absolute index to substitute for i so we leave it as it is. Understanding how we are to treat j relies on the following implicature which Burge claims: "sentences being referred to or quantified over are to be evaluated with the truth schema for the occurrence of 'true' in the evaluating sentence" (p. 180, emphasis in the original). I take that to mean the following: if we have a sentence like ''$\underline{A}$' is true$_i$', then $\tilde{}\underline{P}_i\,'\underline{A}'$ and hence $\underline{Tr}_i\,'\underline{A}' \equiv \underline{A}$ will be implicated. If that is so, then $\underline{a}'$ causes trouble for the implicature, because '$\underline{a}$' is not true' would implicate $\tilde{}\underline{P}_i\underline{a}'$ and $\underline{Tr}_i\underline{a}' \equiv \tilde{}\underline{Tr}_i\underline{a}'$ which are clearly false.

Thus as I understand Burge, when someone asserts $\underline{a}'$, we realize that the implicature fails, and acknowledge it by asserting '$\underline{a}$' is not true$_i$' but without the implicature, merely recording the failure of $\underline{a}'$

to have truth_i conditions. Thus in b', j=i. However, since a' is not $\text{pathological}_{i+1}$ it will have truth_{i+1} conditions:

$$\underline{Tr_{i+1}a'} \equiv \sim \underline{Tr_i a'} \qquad (5.9)$$

So a' turns out to be true_{i+1} allowing us to set k=i+1 in c'.

I must confess at this point that Burge's use of an implicature is completely unclear to me. The problem which he wishes to account for is that in b' we want to record the non-truth of a', but use an identical sentence token to do so. Burge carefully discusses various options, such as explaining the difference between a' and b' as due to some semantic shift either in reference, negation or the truth predicate, and rejects them all. Thus he takes the shift to be explained "in terms of change in implicatures or background assumptions on the part of those propounding or interpreting the relevant sentences" (p.184).

I have two comments to make about this attempt. First, the problem of accounting for a' and b' exists independently of anyone asserting a (or a'). If we merely ask what the truth value of a' is, we have to explain how we can say that a' is not true_i by using the very sentence a'. In fact, of course, if someone seriously asserted a, pragmatic considerations would lead us to understand something entirely different by the assertion, since nobody can assert a literally, knowing it to be paradoxical, and still be abiding by conversational conventions. Thus we interpret anyone saying it as either meaning it non-literally, or being mistaken about the facts, and thinking it is non-paradoxical. Since, however, a' and b' present a problem without there being any question about conditions of assertion, it seems unlikely that b' can be explained in terms of the failure of pragmatic conventions that are supposed to hold for a'.

The second comment I have is that even if the change of implicature that Burge describes occurs it cannot explain a change in truth-value. Changes of implicature may change our assessments of truth-values, given

somebody's assertion, but not the actual truth-value. To extend an example of Grice's, suppose A says "Fred broke up the furniture: he was a little drunk",[45] then the obvious implicature is "Fred was seriously drunk". Suppose, however, A goes on to say "Poor Fred, whenever he gets the least bit drunk, his repressed anger comes out and he starts rampaging around". This clearly cancels the implicature of the first remark. So an audience that did not know Fred well might reasonably think first that "Fred was a little drunk" was false and then that it was true. But none of this affects the case of whether Fred was a little drunk or not. When we look at a', the problem is not that we first think that it is not true and then that it is true; it is rather that the non-true assertion a' and the true assertion b' are given in one and the same sentence.

I think, however, that there is a way of achieving some understanding of a, b and c in terms of Burge's theory without resort to the conversational implicature offered, and it is the following. Given an indexical predicate, we can make lots of semantic judgments that do not have corresponding assertions under our intuitions about the simple predicate 'true'. Suppose, for instance, that we consider '"Snow is white" is true' is true': call that sentence s_1. In C1 and C2, at least, the occurrence of 'true' is s_1 must be given indices 1 and 2 respectively. Consequently, s_1 is pathological$_1$, and not true$_1$. However, the report that s_1 is not true (simpliciter) is false. So not all sentences that are true$_i$ for some i in the hierarchy correspond to true sentences of English. This holds true for C3 as well, thought the examples are more complicated to construct.

The question then arises, "Which evaluations in the hierarchy are reportable in English?", and part of an answer might run as follows.

---

[45]This is borrowed from H. P. Grice, "Logic and Conversation".

Suppose you have a sentence $\underline{Tr_i t}$ where i is the appropriate index applied to express '$\underline{t}$ is true'. Then no judgment $\sim\underline{Tr_j t}$ is translated back as '$\underline{t}$ is not true' if $j<i$. Thus in our example, the translation of '$\underline{s}_1$ is true' is presumably $\underline{Tr_2 s_1}$ and neither of the (true) sentences $\sim\underline{Tr_1 s_1}$ and $\sim\underline{Tr_2 s_1}$ is translated back as '$\underline{s}_1$ is not true'. Since in general i will be assigned in such a way that $\underline{t}$ is not pathological$_i$, the truth of $\underline{Tr_i t}$ is ascertained by the schema $\underline{Tr_i t}=\underline{X}$ where $\underline{X}$ is the denotation of $\underline{t}$. By Theorem 2, if $\underline{t}$ is true$_i$, it is also true$_j$, $j>i$, so all judgments $\underline{Tr_j t}$ for $j>i$ will uniformly go back to English as '$\underline{t}$ is true'.

However, for the paradoxical case like $\underline{a}'$, $\underline{a}'$ is not true$_i$, but is true$_{i+1}$ and given the above rule, <u>both</u> of these sentences can be translated back, the first as '$\underline{a}$ is not true', the second as '$\underline{a}$ is true'. Then Burge can find a way of justifying $\underline{b}$ and $\underline{c}$, but the justification for $\underline{b}$ is based not on a pragmatic feature, but on semantic rules about representing 'true' in the constructions using $\underline{Tr_i}$ and vice versa.[46]

Unfortunately, Burge's ground for claiming that 'true' is an indexical predicate seem insufficient. I should point out that there are two very different ways of regarding the relationship between 'true' and the $\underline{Tr_i}$ predicates. In the one I have described, 'true' is a genuinely indexical predicate, just as 'is here' is, or 'is now'. That

---

[46]There are still problems attached to this. The judgment '$\underline{a}$ is not true' for which the hierarchy does produce some justification, as I have outlined, is still itself not true in the same way that $\underline{a}$ is not true, namely it is not true$_i$. Thus for the judgment $\underline{b}$, both the sentences '$\underline{b}$ is not true' and '$\underline{b}$ is true' are justified by the hierarchy, just as they both are for $\underline{a}$. Thus we could see Burge's use of pragmatic principles as an attempt to rule out the judgment '$\underline{b}$ is not true' as out of order in some way. My point still stands: pragmatics cannot change truth. If Burge could give a semantic account which justifies rejecting '$\underline{b}$ is not true', all well and good, though it seems impossible that he should.

is, different occurrences of 'The clock is here' or 'The race is now' can have different truth-values because of the different contexts of use. On the other view, the English predicate 'true' is non-indexical, but its extension is determined by a hierarchy of distinct truth predicates $\underline{Tr}_i$. We could, for instance, describe Kripke's construction not as giving a series of extensions for a single truth predicate $\underline{T}$, but as giving the extension of a series of truth predicates $\underline{T}_i$ (though it is not really as simple as that, because of the occurrence in the model of $\underline{T}$). On this second view, the value of '$\underline{s}_1$ is true' is determined by the ultimate verdict the hierarchy settles on: it might be false when the occurrence of 'true' gives indices 1 or 2, but since it settles down after that to be true_i, we take it to be true, simpliciter. On this view of Burge's theory then, the Liar sentence is just plain true, because after some funny stuff, it settles down to true_j, $j>i+1$. Clearly on this account the non-truth of $\underline{a}$ cannot be reported in English any more than the non-truth of $\underline{s}_1$ can be.

So if we are to explain $\underline{b}$ and $\underline{c}$, it seems we must adopt the genuinely indexical view. But cases like $\underline{b}$ and $\underline{c}$ are the only pieces of evidence that 'true' is indexical. In general, other indexical features of language share three characteristics:

(5.10)

a) Different tokens of the same sentence-type can have different truth-values.

b) In reported speech indexical sentences go through systematic shifts as in 'He said that he was there".

c) We can eliminate indexical elements in favour of non-indexical ones.

Of these three features, truth ascriptions share only one, once we have accounted for other sources of indexicality in sentences. The third feature seems particularly important, in view of the first, since unless there is a way of publicly specifying the referent of 'here', for

instance, 'I am here' communicates no information. This is not to say that all indexical claims are ultimately eliminable altogether, but just that on given occasions, ambiguity is eliminable by reference to some public framework. Not only is it the case that 'true' does not have this property, but apparently it must not have it. For presumably elimination of the indexicality of 'true' would involve some reference to the particular $\underline{Tr}_i$ relevant on some occasion. But if we were allowed to refer to particular $\underline{Tr}_i$ we could express the sentence:

This sentence is true at no level. (5.11)

a sentence which Burge calls a Super Liar. Of such sentences, he says that they represent a misunderstanding of his account, because even in that sentence 'true' will have an index i; they are akin to 'I am here at some place' (p. 192). Maybe that is true of the Super Liar sentence: the point is that we do have for 'here', but do not have for 'true' a non-indexical reference scheme, and to that extent 'true' is very unlike other indexical expressions.

There might perhaps be no serious harm in holding that 'true' was an unusual indexical expression if there were clear-cut cases in which it was indexical, but the cases Burge offers seem so dubious as to open him to the charge of offering an ad hoc solution. First, I do not think that our native intuitions plump readily for the claim b), that a Liar sentence is not true: I have suggested above, in my Introduction, that the reliable limit of our intuitions is reached when we see that holding that the Liar is true or not true leads to contradiction. At that point we are reduced to non-comprehension, and make no judgments at all. Furthermore, I doubt that anyone's intuition is that the Liar is clearly true, which is Burge's verdict on it. What has happened is that numerous writers have held that the Liar is not true, and it does seem plausible to say that if you are going to say that, you cannot justify not going on to say that it is true; but invoking an indexicality in

'true' to explain what semantic theorists have held is going a long way from intuition. A final doubt that 'true' should be held to be indexical is cast by the following consideration. With other indexical expressions, we recognize that there is no contradiction involved in saying that one occurrence is true and another false. Nobody naturally would hold that saying that the Liar is true and saying it is not true is contradictory, however: it is precisely because it is contradictory and yet apparently justifiable that makes the Liar a paradox.

So it is methodologically unsound to hold that 'true' is indexical, at least for the reasons that Burge offers. Nevertheless, we can still consider his theory as a proposal for a Tarskian theory of truth, by taking the second interpretation of the relation between 'true' and the predicate $\underline{Tr}_i$. How well does the proposal meet the standard objections against the Tarskian approach, given in the Introduction?

The first objection, that we simply do not use a hierarchical predicate, is obviously met: in practice we do not, but Burge can say that the predicate we do use is a systematically ambiguous one, ranging over a hierarchy. I am less happy with the device Burge adopts to meet the second objection, that there are global uses of 'true' which cannot be modelled by any hierarchical technique.

Burge's suggestion is that a sentence like 'Every sentence is true or not' be interpreted not as having any particular index on its occurrence of 'true', but rather as a schematic sentence with an open place for the subscript, to be filled in as the context requires (p. 192). Thus we could express it as:

$$(\underline{s})(\underline{Tr}_i\underline{s} \vee \tilde{}\underline{Tr}_i\underline{s}) \qquad (5.12)$$

If someone then judges this sentence to be true, they can be thought of as asserting

$$(\underline{s})(\underline{Tr}_{i+1}\underline{s} \vee \tilde{}\underline{Tr}_{i+1}\underline{s}) \qquad (5.13)$$

Once again, this suggestion is seriously inadequate to cope with the different cases which occur and some of the difficulties may be more than just technical ones. Suppose we have:

1) Apples are tasty. 2) 1) is true. 3) 2) is true. etc.

(5.14)

All the sentences in the box are true. (5.15)

Intuitively, all the sentences in the box are true so (5.15) is true. However, under C1 and C2, at least, the truth of (5.15) cannot be represented using ordinary indexical truth predicates. For suppose we express it by

$(\underline{s})(\underline{Bs} \supset \underline{Tr_i s})$ (5.16)

where $\underline{B}\alpha$: α is in the box.

Then whatever value we choose for i, there is bound to be a $j>i$ such that some sentence in the box contains $\underline{Tr_j}$, and hence for that sentence $\underline{Bs}$ and $\sim\underline{Tr_i s}$ hold. Consequently no choice of i can make (5.16) true.

Several problems appear if we try to claim that (5.15) is really schematic. It seems plausible to say that 'Every sentence is either true or false' or $\underline{P_i}\,'\underline{A}' \supset \underline{Tr_i}\,'\underline{A}'$, for instance, are schematic. There is indeed no particular index we would care to attach to them, and they are true whatever index we do attach to them. They seem to be abstract principles in about the right way to be schematic. (5.15), however, is a very different kind of sentence. First, whether or not it can be expressed by (5.16) will be a matter of fact: if an appropriate collection of sentences is in the box, there will be a choice of i for which (5.16) represents (5.15) quite satisfactorily. Thus whether or not (5.15) is a schematic utterance is a matter of fact. Furthermore, if it is schematic in some model, it cannot be expressed equivalently to

(5.13), because (5.16) will still be false for each and every i. 'Every sentence is true or false' might be regarded as a schematic utterance of each and every instance of (5.12): 'Every sentence in the box is true' cannot be regarded as a schematic utterance of each and every instance of (5.16).

Thus we cannot tell, independently of the results of a model, which sentences are to be understood schematically and which not, and we have no uniform way of interpreting schematic utterances in terms of non-schematic ones. A final problem is that we may be left with the re-emergence of paradox. Consider the sentence

Either this sentence is not true or
one of the sentences in the box is not true. (5.17)

Intuitively, this is paradoxical: if it is true, both disjuncts are false, and if it is false, one of its disjuncts is true. We cannot however represent it by the sentence $\underline{d}$, where

$$\underline{v}(\underline{d})=\sim\underline{Tr}_i\underline{d}\vee(\underline{Es})(\underline{Bs}\&\sim\underline{Tr}_j\underline{s}) \tag{5.18}$$

because for every choice of j, $(\underline{Es})(\underline{Bs}\&\sim\underline{Tr}_j\underline{s})$ is true, so that (5.18) would represent (5.17) as true, counterintuitively. So it must be schematic. We can perhaps represent the second disjunct by the schematic $\sim\underline{Tr}_i\underline{b}_i$ where $\underline{b}_i$ is the i-th sentence in the box, but representing the first disjunct seems impossible without using a new predicate $\underline{Tr}_s$ for 'true (schematically)'. Then we might represent (5.17) by $\underline{d}'$, where

$$\underline{v}(\underline{d}')=\sim\underline{Tr}_s(\underline{d}')\vee\sim\underline{Tr}_i\underline{b}_i \tag{5.19}$$

but then we do not know how to evaluate $\underline{Tr}_s(\underline{d}')$. We could take Burge's hint that saying of a schematic sentence that it is true is to raise the indices by 1; this gives

$$\underline{v}(\underline{d}')=\sim(\sim\underline{Tr}_s(\underline{d}')\vee\sim\underline{Tr}_{i+1}\underline{b}_i)\vee\sim\underline{Tr}_i\underline{b}_i \tag{5.20}$$

However, this process clearly can never eliminate $\underline{Tr}_S$ entirely, and without some account of what its truth conditions are, we cannot be satisfied that paradox has been eliminated.

It is true that in C3, (5.15) can be handled without schematics: in fact the index on all occurrences of 'true' in (5.14) and (5.15) is just 1. However, a different example shows that C3 cannot escape all such problems. Consider the sequence:

$$\begin{aligned} \underline{a}_1&: \ \underline{a}_1 \text{ is not true} \\ \underline{a}_2&: \ \underline{a}_1 \text{ is true} \\ \underline{a}_{2n-1}&: \ \underline{a}_{2n-2} \text{ is true and } \underline{a}_{2n-1} \text{ is not true} \\ \underline{a}_{2n}&: \ \underline{a}_{2n-1} \text{ is true} \end{aligned} \tag{5.21}$$

C3 permits iteration of indices only for non-pathological cases. If we assign 1 to the occurrence of 'true' in $\underline{a}_1$, $\underline{a}_1$ is clearly pathological$_1$ and not true$_1$. Since $\underline{a}_1$ is pathological$_1$, the occurrence of 'true' in $\underline{a}_2$ must be of index 2: then $\underline{a}_2$ is non-pathological$_2$, and in fact true$_2$. In general, the index on the second occurrence of 'true' in $\underline{a}_{2n-1}$ is n, and that on the occurrence of 'true' in $\underline{a}_{2n}$ is n+1. Then the remark that either one of the $\underline{a}_{2n}$ are not true or that very remark is not true will be in the same situation that (5.17) is in with respect to C1 and C2. Thus C3 may give better results for some sentences, but cannot avoid problems with schematics altogether.

In the end, it really does not seem that reference to schematics meets the objection about global truth ascriptions. The schematic form of 'All sentences are true or false' admittedly has the right truth-value, insofar as schematic utterances have them, but when we look at 'God is omniscient' (i.e. knows all truths) then we are inclined to say that if it is true, God should know that he is omniscient, too. Since the schematic form seems to imply merely that God knows all the non-schematic truths, it fails to represent this feature.

Overall, I feel that Burge has failed to satisfy certain

requirements on a theory of truth which Kripke's article made essential: that such a theory should provide a definition which determined the extension of truth given a model, and that it should readily handle transfinite examples like (5.15). Furthermore, Burge's account rests on some methodologically unsound premisses. The problems with global truth-predications are less serious methodologically, though serious enough, of course, for a theory of truth. We are still in the position of looking for a theory which might plausibly be called a Tarskian theory of truth which nevertheless matches our intuitions rather better than Tarski's own.

Chapter 6

Truth and Dependence

I want, finally, to develop an alternative to the foregoing theories, and shall start by summing up some points of agreement and disagreement with Kripke and Burge.

As should by now be clear, I agree with Kripke that it is important to offer a theory of truth which is defined with respect to a basic model. This is one area in which Burge's theory is seriously deficient, as I argued in the last chapter. Second, I follow Kripke in using levels only as a calculating device, so that ordinary uses of 'true' are not presumed to have an implicit index. Again, as I have already argued, the support for the claim that 'true' is intrinsically indexical is so slight that it would require a very convincing theory to persuade us of its truth, and I am not in a position to provide one.

On the other hand, Kripke's account has its disadvantages: first, as Burge argued, the use of three-valued logics is insufficiently motivated, and only produces such results as it does by restricting expressibility, and second, it treats logical laws in an implausible way. However, a few points about the first disadvantage have to be cleared up. As I argued, Kripke's theory is, in a sense, a three-valued truth theory for a two-valued language, and in many ways my own theory ends up the same. My objection to Kripke, then, is not that the resultant theory does not give a truth-value to some sentences: it is, rather, that the twin disadvantages of failure to express exclusion negation and inability to make logical laws true result from not taking

bivalence seriously, and not taking it as far as possible. Thus I shall start with a two-valued language, and make the truth-predicate reflect as many of the facts of that language as possible, rather than treating it as a degenerate three-valued language.

I would like to put this disagreement in the terms of an earlier discussion of semantic principles. The Principle of Bivalence, at least construed as a principle about meaningful sentences, is much more like a stereotype than a defining principle. The situation may be very different for propositions, where it may be defining, but it has long been recognized that many kinds of sentences can be highly meaningful, yet fail to be true or false. Consequently, I have no a priori objection to a semantic theory which holds that there is some additional class of sentences which are meaningful but fail to be true or false. My objection is rather that Kripke gives up bivalence on some sentences for which we have a strong intuition that bivalence holds, and if another theory can accomodate that intuition, so much the better for it.

These and similar considerations also provide the starting point for a closely connected set of theories of truth discussed by Herzberger, in "Notes on Naive Semantics" and "Naive Semantics and the Liar Paradox", Gupta, in "Truth and Paradox", and Belnap, in "Gupta's Rule of Revision Theory of Truth", and before giving my own theory, I shall sketch this set of theories, and then later discuss the relationship between them and my own.

The central characteristic of these theories is the revision rule for 'true'. Suppose $\underline{L}$ is a standard first order language with a logical predicate $\underline{T}$, and $\underline{M}$ is a base model for $\underline{L}$, namely a two-valued model which interprets everything except $\underline{T}$. Suppose we add to $\underline{M}$ an arbitrary set $\underline{U}$ as the extension of $\underline{T}$: then we can define the revision of $\underline{U}$, $\tau^1_{\underline{M}}(\underline{U})$, as the set of sentences true in the model $\underline{M}+\underline{U}$, and then revise that set again, to $\tau^2_{\underline{M}}(\underline{U})$, the set of sentences true in $\underline{M}+\tau^1_{\underline{M}}(\underline{U})$. We can

go on repeating this process ad nauseam. Now if the model contains a Liar sentence, for instance, it will steadily switch in and out of the extension of T, but no matter how obscure a set U might be, gradually more and more sentences settle down as the revision continues. Eventually we will need to use a limit rule, and it is here that divergences between the different theories appear. Obviously at the limit it is sensible to keep in the extension of T those sentences that have settled in it by then, and to keep out those that have settled out. The question is what to do with those that have not settled down at all yet, and different revision rules decide this question in different ways. In this initial discussion, I shall give Gupta's original rule, and discuss different versions later.

Gupta's suggestion is that at limit ordinals we refer to U again to decide the wobbly cases. To define the appropriate function we need some preliminary definitions.

Definition 1: P is locally stably true at α for U iff $(E\beta)(\beta<\alpha$ and $(\gamma)($If $\beta\leq\gamma<\alpha$ then $P\in\tau_M^{\gamma}(U))$.

P is locally stably false at α for U iff $(E\beta)(\beta<\alpha$ and $(\gamma)($If $\beta\leq\gamma<\alpha$ then $P\notin\tau_M^{\gamma}(U))$.

Definition 2: By recursion

$\alpha=0$ $\quad\tau_M^{\alpha}(U)=U$, where $U\subseteq L$

$\alpha=\beta+1$ $\quad\tau_M^{\alpha}(U)=\{P:P$ is true in $M+\tau_M^{\beta}(U)\}$

α is a limit ordinal

$$\tau_M^{\alpha}(U)=X\cup(U-Y)$$

where $X=\{P:P$ is locally stably true at $\alpha\}$

$Y=\{P:P$ is locally stably false at $\alpha\}$

Now what is interesting about $\tau_M(U)$ is that while in the general case it will not settle down to a single stable value, it will eventually reach a stage at which the same sets are repeated in sequence with most of the sentences being settled one way or the other, and only

a few wobbling in a regular pattern. What is more, no matter how strange U might be, most of the same sentences will settle down in the end.

In certain models, in fact, every starting point leads to the same set. In such cases the model is called Thomasonian, and it represents a criterion of non-pathologicality, because we can regard the truth predicate as wholly defined by the model. In models containing a Truth-Teller, this will not happen: from some starting points the Truth-Teller will be in the extension of T, and from others it will be out of it, though from each starting point a single extension set results. Increased degrees of pathologicality result in a greater divergence of one starting point from another, but a large measure of uniformity still runs through the truth extensions.

Relative to a given model M, then, we can define kinds of stability that a sentence may exhibit.

Definition 3: P is stably true(false) in M relative to U iff $(E\beta)(\gamma)$(If $\gamma > \beta$ then $P \in \tau_M^{\gamma}(U)$ (or $P \notin \tau_M^{\gamma}(U)$)).

Definition 4: P is stably true (false) in M iff P is stably true (false) in M relative to all U.

Since the definition of τ_M relies on regular two-valued valuation rules, sentences such as

$$(x)(Tx \vee \tilde{} Tx) \tag{6.1}$$

will clearly be stably true, as will all the logical laws. In addition, Herzberger and Gupta have shown that all the Kripke grounded sentences are stably true or false, so the theory obviously has many advantages. One minor disadvantage is that, as we shall shortly see, the sentence

$$(x)(Tx \vee T(neg(x))), \tag{6.2}$$

where neg(x) denotes the negation of x,

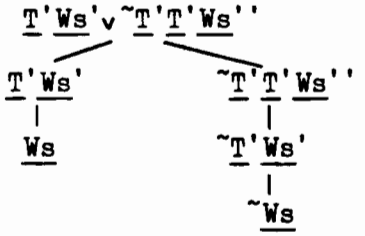
is not stably true.

I shall turn now from this brief sketch of the revision theory to present my own. The most immediate difference is that whereas both the Kripke and Gupta theories start with notional extensions for T and then proceed to revise those extensions by moving up through level after level, I present a downward or dependence theory, in which we start with a sentence and ask what would have to be true for it to be true, and what in turn that depends on, and so on. Ideally, this process gives simpler and simpler sentences, and at some point we can say that the simple sentences are true and that our original sentence thereby inherits its truth.

One advantage of such a theory over the upward theories is its correspondence to our actual method of determining the truth of sentences. We do not, after all, determine the truth of 'Everything Fred said is true' by constructing a minimal fixed point and checking to see whether it is in the extension of T there: rather we look at what Fred said and ask 'Is each of these true?' If he said things of some complexity, then in turn we must analyze them, and see if what they depend on is true, and so forth. This advantage is, in itself, rather slight, since so long as an upward theory gave a plausible definition of 'true', it would not matter whether we actually determined truth by another method. However, it is important that there be another method, for the following reason. If a given upward theory had no corresponding downward version, then it would cast doubt on the legitimacy of the theory because it could not correspond to a usable theory of truth. If we could not determine the truth of 'Everything Fred said is true' by a comparatively local investigation of what Fred said, but instead also had to find out who won the 1932 World Series, what the current precipitation in China was, and how old Julius Caesar's mother was when she died, then there would be a serious flaw in the theory. (Unless, of course, these were all things Fred made claims about!)

An additional advantage of the downward mode is that even in cases where it is equivalent to an upward theory, it may provide considerable insight into that theory, and allow us to prove results which are otherwise difficult to prove within the theory. Downward versions of the Kripke construction are available in papers by Lawrence Davis ("An Alternate Formulation of Kripke's Theory of Truth") and Steve Yablo ("Grounding, Dependence and Paradox"). The former is limited only to giving a downward version of the Strong Kleene minimal fixed point, but the latter is considerably more general. Both, however, are restricted to grounded sentences, but in the following sections I shall not only discuss both grounded sentences and those that get truth-values in non-minimal fixed points, but also establish some correspondence with the Gupta revision process, all in a unified framework which enables us to compare easily various theories which differ quite markedly in their upward versions.

What, intuitively, is the dependence of one sentence on others? Suppose we start with a sentence A like 'Either it is true that snow is white or it is not true that it is true that snow is white'. Clearly the first dependence we recognize is the disjunction: one or the other of 'It is true that snow is white' and 'It is not true that it is true that snow is white' has to be true for A to be true. For the former to be true, 'Snow is white' must be true, and for the latter , 'It is not true that snow is white' must be. This last, in turn, depends on 'Snow is not white' being true. Schematically, we have

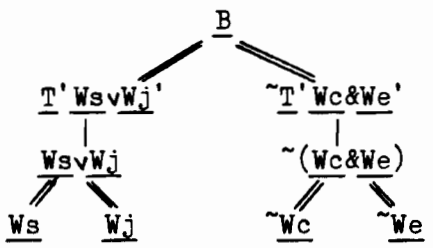


(6.3)

Since snow is white, Ws, T'Ws' and T'Ws'∨~T'T'Ws'' are all true, and none of ~Ws, ~T'Ws' and ~T'T'Ws'' is.

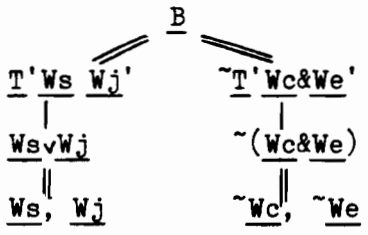
In this dependence we can easily recognize two different aspects. One is the dependence of T'Ws'∨~T'T'Ws'' on T'Ws' and ~T'T'Ws'', the other is that of T'Ws' on Ws, for instance. Roughly, we can call the former logical and the latter semantic. These two aspects will, in general, alternate, since a truth predication may depend semantically on a logically compound sentence: once we have analyzed the logical dependence of this sentence there may be some truth predications whose semantic dependence we need to expose, and so on. Suppose we represent logical dependence by double lines, and semantic dependence by single lines. Consider, for instance, the sentence B: 'It is true that either snow is white or jade is, but it is not true that coal and ebony are.' The downward analysis now gives:

(6.4)



In this diagram, the double lines still represent vastly different kinds of dependences, conjunctive and disjunctive. For simplicity, I shall henceforth use conjunctive branching only,[47] so that the preceding diagram becomes:

(6.5)



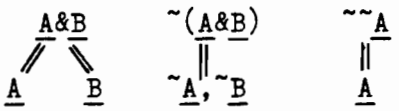
Thus we understand this diagram to say that the truth of B depends

---

[47]Naturally, disjunctive branching could have been used: later definitions and proofs are easier with conjunctive branching.

ultimately on one of Ws and Wj being true, and one of ~Wc and ~We being true.

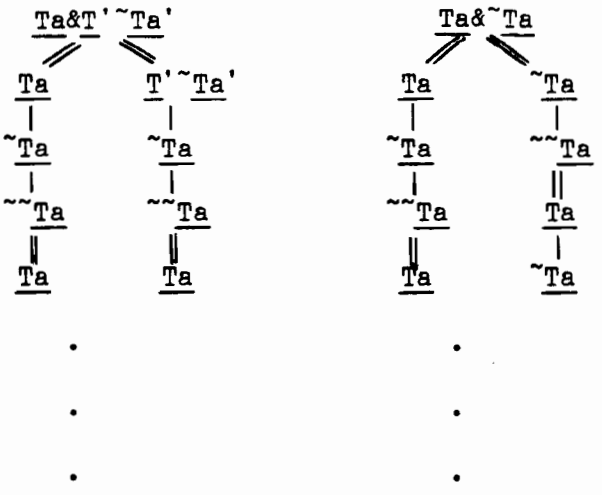
For simplicity I shall restrict the investigation to propositional connectives only. This does not result in any serious restriction on the usefulness of the theory, because many different kinds of pathologicality can still be represented, and the gain in economy is very great. I shall briefly discuss the treatment of quantifiers later. Furthermore, I shall treat conjunction as primitive. Thus there are only three basic forms of logical dependence:



(6.6)

Turning now to semantic dependence, the basic rule being applied is the disquotation rule: Tt depends on den(t), the denotation of t, and ~Tt depends on neg(den(t)), the negation of den(t). However, there is an added complication in the way the disquotation is used, which rests on an important aspect of the whole notion of dependence which I have not yet discussed.

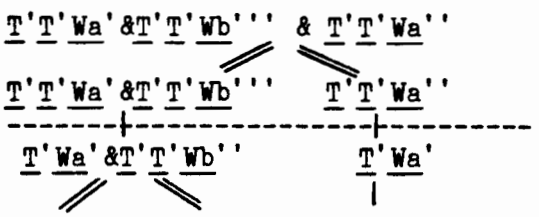
This aspect is revealed when we examine the dependences of the two sentences Ta&T'~Ta' and Ta&~Ta where, as usual, a is a Liar sentence, i.e. den(a)=~Ta.



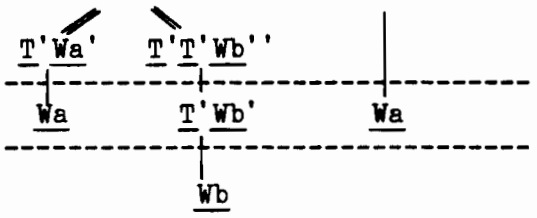
It is, perhaps, natural in the dependences so far discussed, to take each branch separately, see what that requires for the truth of the original sentence, and then sum up the requirements for all branches. These examples show that that policy cannot accomodate our wishes for these two sentences. If we try it in these cases, we come up with essentially the same result: for either sentence to be true, it would be necessary for both Ta and ~Ta to be true. This is obviously impossible, so neither sentence can be true. However, this conceals a crucial difference between them. The first sentence is equivalent to Ta, since it merely says that the Liar is true in two different ways, and should presumably receive the same treatment as the Liar, whereas the second is an out-and-out contradiction, and its dependence should register that fact.

What has happened is that some crucial information has been ignored, namely what each branch contains after each successive use of the disquotation rule. If we examine this information, the resemblance of the first sentence to the Liar and of the second to a logical falsehood becomes closer. For the first sentence merely requires after each disquotation either that Ta be true, or that ~Ta be true, just as the Liar does itself, while the second requires that both Ta and ~Ta be true each time.

To preserve this information, and the discriminations that it permits, it is necessary not only that logical analysis and disquotation alternate on each branch, but also that they do so in a regimented way on the dependence structure as a whole. First, all logical analysis is completed, then disquotation is applied, then logical analysis, and so on. So we finally obtain structures looking like this:



(6.7)



where the dotted lines mark stages of (universal) disquotation.

These preliminary remarks explain most of the following definitions, except that instead of dependence structures, I only define a function on sets of sets of sentences. This method results in a loss of some structural information, such as what sentences in the structure depend on which others, but it turns out that the essential information is how various sentences at one level depend en masse on other sentences at other levels, and this method has the virtue of emphasizing the level-by-level dependence rather than dependencies within each branch.

So let L be a standard first-order language with negation and conjunction, but no quantifiers. In addition, let it contain quotation names of all its sentences, plus a monadic logical constant T. An ordered pair M=⟨D,v⟩ is a model for L if L⊆D, and v is a regular two-valued valuation function in which quotation names get their intended interpretation but T is uninterpreted. Then, relative to L and M, we can define a series of dependence relations between sentences, sets of sentences and sets of sets of sentences.

The basic relations between sets of sentences are the literal conjunctive dependence relation C and the disquotation relation D. To define the former, however, we need first to define a more primitive relation, the "one-step" conjunctive dependence relation C', and the notion of a literal:

> **Definition 5:** A sentence P is a literal iff P is an atomic sentence or the negation of one. Let Lit be the set of literals of L.

> **Definition 6:** If A is a non-empty set of sentences then B is a conjunctive dependent of A (C'(A,B)) iff:

1. for every literal P, if $P \in A$ then $P \in B$

2. if $P \& Q \in A$, $P \in B$ or $Q \in B$, but not both

3. if $\sim(P \& Q) \in A$, $\sim P \in B$ and $\sim Q \in B$

4. if $\sim\sim P \in A$, $P \in B$

5. nothing else is in B.

Definition 7: If A, B are finite sets of sentences, then B is a literal conjunctive dependent of A ($C(A,B)$), iff there is a sequence of sets of sentences $\langle A_1, \ldots\ldots A_n \rangle$ such that

1. $A_1 = A$, $A_n = B$

2. for all A_i, $C'(A_i, A_{i+1})$

3. for all $P \in B$, P is a literal.

Effectively, a literal conjunctive dependent of a set of sentences bears a relation to that set similar to the relation one conjunct in a conjunctive normal form for a sentence bears to that sentence.

Definition 8: If A is a non-empty set of literals, B is the disquotation of A ($D(A,B)$) iff

1. if $Tt \in A$, $v(t) \in B$

2. if $\sim Tt \in A$, $neg(v(t)) \in B$, where $neg(A) = \sim A$

3. all other members of A belong to B

4. nothing else belongs to B.

The two relations between sets of sentences, C and D, give rise to two functions of sets of sets of sentences:

Definition 9: If $X \in \mathfrak{P}(L)$, and $X \neq \Lambda$, then $A \in C(X)$ iff $A \subseteq Lit$ and $(\exists B)(B \in X$ and $C(B,A))$.

Definition 10: If $X \subseteq \mathfrak{P}(Lit)$ and $X \neq \Lambda$, then $A \in D(X)$ iff $A \subseteq L$ and $(\exists B)(B \in X$ and $D(B,A))$.

Finally we can define, for a set of sets of sentences X the

dependence function R, by recursion.[48]

Definition 11: If $\underline{X} \leq \mathfrak{D}(\underline{L})$, then

1. $\underline{R}(\underline{X},0)=\underline{C}(\underline{X})$
2. $\underline{R}(\underline{X},n+1)=\underline{C}(\underline{D}(\underline{R}(\underline{X},n)))$.

Using R, now, we can represent the dependence tree of (6.7), letting P abbreviate T'T'Wa'&T'T'Wb'''&T'T'Wa'':

$$\underline{R}(\{\{\underline{P}\}\},0)=\{\{\underline{T}'\underline{T}'\underline{Wa}'\&\underline{T}'\underline{T}'\underline{Wb}'''\},\{\underline{T}'\underline{T}'\underline{Wa}''\}\} \tag{6.8}$$

$$\underline{R}(\{\{\underline{P}\}\},1)=\{\{\underline{T}'\underline{Wa}'\},\{\underline{T}'\underline{T}'\underline{Wb}''\}\}$$

$$\underline{R}(\{\{\underline{P}\}\},2)=\{\{\underline{Wa}\},\{\underline{T}'\underline{Wb}'\}\}$$

$$\underline{R}(\{\{\underline{P}\}\},3)=\{\{\underline{Wa}\},\{\underline{Wb}\}\}$$

$$\underline{R}(\{\{\underline{P}\}\},n)=\underline{R}(\{\{\underline{P}\}\},3), \text{ for all } n>3.$$

For those accustomed to a tree-style semantics, the development so far has a striking omission, namely the absence of any discussion of a closure rule. Even the most innocent sentence, such as '"Snow is white" is true', has an infinitely long dependence function: in fact, it is the following.

$$\begin{aligned}\underline{R}(\{\{\underline{T}'\underline{Ws}'\}\},0)&=\{\{\underline{T}'\underline{Ws}'\}\}\\ \underline{R}(\{\{\underline{T}'\underline{Ws}'\}\},1)&=\{\{\underline{Ws}\}\}\\ \underline{R}(\{\{\underline{T}'\underline{Ws}'\}\},n)&=\{\{\underline{Ws}\}\} \quad n>1.\end{aligned} \tag{6.9}$$

Henceforth, I shall seldom, if ever, give the full form of the dependence function, but rather give its values sequentially, divested of inessential braces, and sometimes with added lines for clarity. Thus $\underline{R}(\{\{\underline{T}'\underline{Ws}'\}\})$ will be shown as:

---

[48] I shall frequently use 'the dependence function of P (or A)' to refer to the dependence function of $\{\{\underline{P}\}\}$ (or $\{\underline{A}\}$).

$$\begin{array}{l} \underline{T'\underline{Ws}'} \\ \underline{Ws} \\ \cdot \\ \cdot \\ \cdot \end{array} \qquad (6.10)$$

The following two sections are devoted to a discussion of closure rules, but until that discussion, the dependence functions will be left in their current infinite state. Given some of the closure rules proposed, it would be easy to close the function for $\underline{T'\underline{Ws}'}$ after reaching $\underline{Ws}$, as is only natural, so the excess material can be thrown away. However, until it is quite clear what parts of the functions may be thrown away without loss, we shall drag all of them around with us.

To start the discussion of closure rules, it is worth seeing why a naive application of the closure rule for straight propositional conjunctive trees will not work. By that rule, if a branch contained $\underline{A}$ and $\underline{\sim A}$, the branch was closed, and if every branch closed, the sentence was logically true. However, the dependence function for the Liar sentence, $\underline{a}$, is:

$$\begin{array}{l} \underline{\sim Ta} \\ \underline{Ta} \\ \underline{\sim Ta} \\ \cdot \\ \cdot \\ \cdot \end{array} \qquad (6.11)$$

Consequently, applying that rule here would have as a result that the Liar is logically true, which is rather startling. A little reflection will show why this rule is inappropriate, since successive members in the function will not necessarily be simultaneously satisfiable, though if the sentence is true, they should be.

What instead we can do is give closure rules which are equivalent to various different policies in the Kripke construction, and also ones which compare with stability in the Gupta revision process, and I shall now go on to discuss these.

6.1 Closure Rules for Kripke Constructions

As the discussion of Kripke's theory showed, it is important that a theory of truth be able to represent groundedness in a natural way. Consequently, the first closure rule I shall discuss is one which does precisely that.

The intuitive notion of groundedness is that a sentence is grounded if one can trace its relation to some collection of facts which are sufficient to determine its truth value. Since the relation of a given sentence to the successive values of its dependence function is closely related to the claim that the sentence is true just in case a series of conjunctive normal forms are true, the obvious closure rule is to say that a dependence function is closed if at some level the appropriate conjunctive normal form is true in the basic model. This notion is captured in the following definition:

> Definition 12: A dependence function $\underline{R}(\underline{X})$ is a-closed in a model $\underline{M}$ iff for some α, $(\underline{A})$(If $\underline{A}\epsilon\underline{R}(\underline{X},\alpha)$, then $(E\underline{P})(\underline{P}\epsilon\underline{A}$ and $\underline{val}_{\underline{M}}(\underline{P})=t)$), where $\underline{val}_{\underline{M}}(\underline{P})=t$ iff $\underline{P}$ is true in $\underline{M}$.

Note that if $\underline{P}$ is in $\underline{A}$, where $\underline{A}$ belongs to some level of a dependence function, it must be a literal, and if it is true in the model, it cannot ascribe or deny truth, since $\underline{M}$ is not defined for $\underline{T}$. A typical a-closed dependence function would be the one for $\underline{T}'\underline{Ws}'\vee(\underline{T}'\underline{T}'\underline{Ws}''\&\sim\underline{T}'\underline{T}'\sim\underline{Ws}'')$, which is:

$$\begin{array}{l}\{\underline{T}'\underline{Ws}',\ \underline{T}'\underline{T}'\underline{Ws}''\},\ \{\underline{T}'\underline{Ws}',\ \sim\underline{T}'\underline{T}'\sim\underline{Ws}''\} \\ \quad\{\underline{Ws},\ \underline{T}'\underline{Ws}'\},\ \ \{\underline{Ws},\ \sim\underline{T}'\sim\underline{Ws}'\} \\ \qquad\{\underline{Ws}\} \\ \qquad\vdots\end{array} \tag{6.12}$$

so long as $\underline{M}$ makes $\underline{Ws}$ true.

To show that the a-closure of the dependence function of either $\underline{P}$ or $\sim\underline{P}$ is equivalent to $\underline{P}$ being grounded, we must first establish various

lemmas concerning the relationship of ascent in the Kripke hierarchy to descent in dependence functions.

Lemma 13: For all $\alpha>0$, all $\underline{A}$ and all $\underline{B}$, if $D(\underline{A},\underline{B})$, then $\underline{A}$ $\underline{\text{Int}}$ $\underline{S}_{1,\alpha+1}$ iff $\underline{B}$ $\underline{\text{Int}}$ $\underline{S}_{1,\alpha}$, where $\underline{A}$ $\underline{\text{Int}}$ $\underline{B}$ iff $A \cap B \neq \Lambda$.

Proof: See Appendix to chapter.

Lemma 14: For all $\alpha>0$, and for all $\underline{A}$, $\underline{A}$ $\underline{\text{Int}}$ $\underline{S}_{1,\alpha}$ iff for all $\underline{B}$, if $C(\underline{A},\underline{B})$ then $\underline{B}$ $\underline{\text{Int}}$ $\underline{S}_{1,\alpha}$.

Proof: See Appendix.

Lemma 15: For all m, if $\alpha+n-m>0$, $\underline{A}$ $\underline{\text{Int}}$ $\underline{S}_{1,\alpha+n}$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},m)$, then $\underline{B}$ $\underline{\text{Int}}$ $\underline{S}_{1,\alpha+n-\dot{m}}$).

Proof: See Appendix.

What 15, the crucial lemma, says is that if a set intersects with the extension of 'T' at some level, then we can descend any number of levels in the dependence function, and the sets there will intersect with the extension of 'T' at the correspondingly lower level, so long as we stay above zero. This has an interesting consequence for the levels in the Kripke hierarchy:

Lemma 16: If $\underline{P}$ is grounded true, then $\underline{P}$ is in $\underline{S}_{1,\alpha}$ for finite α.

Proof: Suppose $\underline{P}$ is grounded true, then there is a lowest level where $\underline{P}$ is in $\underline{S}_1$, $\alpha+n$, where α is a non-successor ordinal. Suppose α is not 0. Then, by using Lemma 15, setting m=n, $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\{\underline{P}\}\},n)$ then $\underline{B}$ $\underline{\text{Int}}$ $\underline{S}_{1,\alpha}$). Let $\underline{F}$ be the set of sentences $(\cup\underline{R}(\{\{\underline{P}\}\},n)) \cap \underline{S}_{1,\alpha}$. Since $\cup\underline{R}(\{\{\underline{P}\}\},n)$ is finite, $\underline{F}$ is finite, and hence there is a level $\beta<\alpha$ at which the last of its members became true. But applying Lemma 15 again, upwards, we find that $\underline{P}$ $\underline{\text{Int}}$ $\underline{S}_{1,\beta+n}$ where $\beta+n<\alpha$. This contradicts the hypothesis, so $\alpha=0$.

This means that not only is quantification sufficient to necessitate transfinite levels, it, or something like it, appears necessary. All grounded truth-values for the propositional case are decided at finite levels. (Obviously the case is the same for grounded false sentences, since sentences are true iff their negations are false and vice versa.)

Finally we obtain the equivalence theorem for a-closure:

Theorem 17: $\underline{P}$ is grounded iff either the dependence function for $\{\{\underline{P}\}\}$ is a-closed, or the dependence function for $\{\{\sim\underline{P}\}\}$ is a-closed.

Proof:

1. Assume $\underline{P}$ is grounded.

   a. Assume $\underline{P}$ is true.
   Then, for some n, $\underline{P}$ intersects $\underline{S}_{1,n}$.
   By Lemma 15, setting m=n-1,
   $(\underline{B})$(If $\underline{B}\in\underline{R}(\{\{\underline{P}\}\},n-1)$ then $\underline{B}$ Int $\underline{S}_{1,1}$).
   But $\underline{Q}\in\underline{S}_{1,1}$ iff $\underline{val}_{\underline{M}}(\underline{Q})=t$.
   Hence $(\underline{B})$(If $\underline{B}\in\underline{R}(\{\{\underline{P}\}\},n-1)$
   then $(\underline{E}\underline{Q})(\underline{Q}\in\underline{B}$ and $\underline{val}_{\underline{M}}(\underline{Q})=t)$).
   So $\underline{R}(\{\{\underline{P}\}\})$ is a-closed.

   b. Assume $\underline{P}$ is false.
   Then $\sim\underline{P}$ is true.
   Hence, as above, $\underline{R}(\{\{\sim\underline{P}\}\})$ is a-closed.

2. For the converse, reverse the steps above.

Technical complications aside, the comparison with the Kripke hierarchy is thus straightforward. To determine whether a sentence is grounded, just construct its dependence function and that of its negation until you reach a level where every set of sentences at that level contains at least one sentence which is true in the base model: naturally, it will not be a truth ascription. If neither function reaches such a level, the sentence is ungrounded.

In fact, the process could be simplified: in constructing the

dependence functions one could close off any set containing a sentence true in the model, because any descendant of such a set will contain the same sentence, it being a non-truth ascribing literal. Thus, for instance, if we have the sentence $\underline{T'T'Fa''\&T'Fa'}$, where $\underline{Fa}$ is true in the model, we need only construct the following:

$$\begin{array}{cc} \{\underline{T'T'Fa''}\} & \{\underline{T'Fa'}\} \\ \{\underline{T'Fa'}\} & \{\underline{Fa}\} \\ \{\underline{Fa}\} & X \end{array} \qquad (6.13)$$

However, we can do more than model (Strong) groundedness: we can see whether a sentence is true in the minimal fixed points for other schemes as well. Consider the closure rules:

Definition 18: A dependence function $\underline{R}(\underline{X})$ is

1. $\underline{\text{b-closed}}$ $\underline{\text{in}}$ $\underline{M}$ iff for some α $(\underline{A})$(If $\underline{A}\in\underline{R}(\underline{X},\alpha)$ then $(\underline{E}\underline{P})(\underline{P}\in\underline{A}$ and $\underline{val}_{\underline{M}}(\underline{P})$=t) and $(\underline{P})$(If $\underline{P}\in\underline{A}$ then $\underline{val}_{\underline{M}}(\underline{P})$=t or f)).
2. $\underline{\text{c-closed}}$ $\underline{\text{in}}$ $\underline{M}$ iff for some α $(\underline{A})$(If $\underline{A}\in\underline{R}(\underline{X},\alpha)$ then $(\underline{E}\underline{P})(\underline{P}\in\underline{A}$ and either $\underline{val}_{\underline{M}}(\underline{P})$=t or $\sim\underline{P}\in\underline{A}$).

Thus a dependence function is b-closed if you can find some level where in every set there is at least one sentence true in the model, and all the others are true or false. Consequently, none of them can be truth ascriptions. On the other hand, c-closure demands the existence of a level where every set either has a true member, or some sentence and its negation: typically, the latter pair may be a truth ascription and its negation. A typical sentence which has a b-closed dependence function might be $\underline{T'Fa\vee Gb'}$ where $\underline{Fa}$ is true in $\underline{M}$ and $\underline{Gb}$ is false. Its dependence function looks like:

$$\begin{array}{c} \underline{T'Fa\vee Gb'} \\ \underline{Fa},\ \underline{Gb} \\ \cdot \\ \cdot \\ \cdot \end{array} \qquad (6.14)$$

On the other hand, $\underline{T}\text{'}\underline{Fa} \vee \underline{Tb}\text{'}$, where $\underline{b}$ is the Truth Teller, has a function which is a-closed but not b-closed, since $\underline{Tb}$ gets no value in $\underline{M}$.

$$\begin{array}{c} \underline{T}\text{'}\underline{Fa} \vee \underline{Tb}\text{'} \\ \underline{Fa} \vee \underline{Tb} \\ \vdots \end{array} \qquad (6.15)$$

A typical c-closed sentence might be $\underline{Tb} \vee \underline{T}\text{'}\underline{Ta} \vee \underline{\sim Ta}\text{'}$ where $\underline{a}$ is a Liar sentence. This gives:

$$\begin{array}{c} \underline{Tb}, \underline{T}\text{'}\underline{Ta} \vee \underline{\sim Ta}\text{'} \\ \underline{Tb}, \underline{Ta}, \underline{\sim Ta} \\ \vdots \end{array} \qquad (6.16)$$

Clearly, this latter dependence function is not a-closed. A final characteristic c-closed function is the one for $\underline{T}\text{'}\underline{Ta}\text{'} \vee \underline{T}\text{'}\underline{\sim Ta}\text{'}$, an instance of the general form $\underline{T}(\underline{a}) \vee \underline{T}(\underline{neg}(\underline{a}))$:

$$\begin{array}{c} \underline{T}\text{'}\underline{Ta}\text{'}, \underline{T}\text{'}\underline{\sim Ta}\text{'} \\ \underline{Ta}, \underline{\sim Ta} \\ \vdots \end{array} \qquad (6.17)$$

We can prove two analogues of Theorem 17:

Theorem 19: $\underline{P}$ is true in the Weak Kleene minimal fixed point iff $\underline{R}(\{\{\underline{P}\}\})$ is b-closed.

Proof: See Appendix.

Theorem 20: $\underline{P}$ is true in the m-c supervaluational minimal fixed point iff $\underline{R}(\{\{\underline{P}\}\})$ is c-closed.

Proof: See Appendix.

These three closure rules allow some interesting reflections on the Kripke hierarchy. In the three schemes, the minimal fixed point is

reached by starting from the same point, and applying diffent valuation rules at each successive level. In the dependence function, however, we can find equivalents to each scheme by using the same reduction from level to level, but different finishing points. Thus one way of constrasting the three schemes is to say that they only differ in the kind of fact that is sufficient to ground a sentence, the Weak being the most demanding, the supervaluational the least so. Furthermore, we can say that there is a strong sense in which the logic of all three schemes is the same: instead of regarding choice between them as a choice of different logics, we can just regard each as a device for keeping a desired property in the truth sets at successive levels. This tends to bear out some cryptic remarks Kripke makes in his "Outline of a Theory of Truth", p. 700, fn. 18.

As noted before, in the discussion of the Kripke construction in Chapter Four, the Strong Kleene scheme seems to give the most plausible account of groundedness, and the supervaluation scheme gets more truth values correct. Also it is quite obvious that the supervaluation scheme extends the Strong Kleene one. This could of course be proved directly, but it is particularly evident in the difference between a-closure and c-closure.

Finally, we can also use the dependence function to look at non-minimal fixed points. Once again, at this point I shall revert to the Strong Kleene scheme, but similar results hold for the other two.

Reflection on the proofs which lead to Theorem 17 show that only at one point is appeal made to the fact that we are working with the function generated by $\underline{M}\langle \mathfrak{A}, \mathfrak{A} \rangle$, namely where we note that $\underline{Q} \in \underline{S}_{1,1}$ iff $\underline{val}_{\underline{M}}(\underline{Q})=t$ Otherwise, all the other claims remain true for non-minimal fixed points, especially:

Lemma 15: $\underline{A}$ $\underline{Int}$ $\underline{S}_{1,\alpha+n}$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ $\underline{Int}$ $\underline{S}_{1,\alpha+n-m}$. (6.18)

Theorem 16 generalized: If $\underline{P}$ is true in a fixed point, then $\underline{P}$ is in $\underline{S}_{1,\alpha}$ for finite α.

So suppose we have an acceptable starting point $\underline{F}(0)$, and want to know whether $\underline{P}$ will be true in the resulting fixed point $\underline{FP}(\underline{F})$. All we need to find is some level n in $\underline{R}(\{\{\underline{P}\}\})$ where all the members intersect with $\underline{F}_{1,1}$. If there is one, $\underline{P}$ is in $\underline{F}_{1,n+1}$, and once in, stays there. It follows that if every member of $\underline{R}(\{\{\underline{P}\}\},n)$ intersects with $\underline{F}_{1,1}$, every member of $\underline{R}(\{\{\underline{P}\}\},m)$, $m>n$, must do so too, since $\underline{P}$ belongs to $\underline{F}_{1,m+1}$.

Conversely, if we have a sentence $\underline{P}$ and want to find out if there is a fixed point in which it is true, we can adopt the following policy: take any final segment of $\underline{R}(\{\{\underline{P}\}\})$, and use it to construct a pair of sets as follows. For each set $\underline{B}$ occurring there, if $\underline{B}$ contains a sentence $\underline{Q}$ such that $\underline{val}_{\underline{M}}(\underline{Q})=t$, do nothing. Otherwise, pick either a truth ascription or the negation of one (if there are no sentences of either kind, abandon the task). If you have picked a sentence $\underline{Tt}$, put $\underline{den}(\underline{t})$ in the first member of the pair. If you have picked a sentence $\underline{\sim Tt}$, put $\underline{den}(\underline{t})$ in the second member. When this is completed for all $\underline{B}$, see whether the result is an acceptable starting point: if it is, then $\underline{P}$ will be true in the resulting fixed point.

Of course this method will not generally be effective, but in simple cases it is easy to follow. The reason for taking a final segment of $\underline{R}(\{\{\underline{P}\}\})$ rather than just some level is that finding the sentences to construct the acceptable starting point may require reference to more than one level. Consider, for instance, the set of sentences $\underline{a}_i$, where:

$$\begin{aligned} &\underline{v}(\underline{a}_0)=\underline{Ta}_1 \\ &\underline{v}(\underline{a}_i)=\underline{Ta}_{i+1} \end{aligned} \qquad (6.19)$$

The dependence function for $\underline{Ta}_0$ just looks like

$$\begin{array}{l} \underline{Ta}_0 \\ \underline{Ta}_1 \\ \underline{Ta}_2 \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \end{array} \qquad (6.20)$$

Clearly $\underline{Ta}_0$ is true in any fixed point where the extension of $\underline{T}$ contained a final sequence of the $\underline{Ta}_i$, but you could not tell that just from looking at one level: that would just tell you that the desired starting point must have some $\underline{Ta}_j$ in the first member, but nothing more.

We can even tell whether a sentence is intrinsically true, i.e. is true in an intrinsic fixed point. Clearly, we can in principle expand the previous method to generate all the acceptable starting points from which $\underline{P}$ will end up true. First we check that there is no acceptable starting point from which $\underline{P}$ is false. Now the proof that $\langle\{\underline{Tc} \vee \tilde{} \underline{Tc}\}, \mathbb{A}\rangle$ gave a fixed point (given on page 96) relied on only one essential fact: that that starting point was consilient with every other. If the starting point is consilient, the fixed point is. Consequently, we need only ask whether any of the acceptable starting points which make $\underline{P}$ true are consilient with all others.

This method is not noticeably easier than constructing all the fixed points, seeing which are intrinsic, and seeing if $\underline{P}$ is true in any of them, but again it emphasizes that the dependence function can illuminate the process, because we find that the result can be achieved by looking only at the starting points, i.e. at the kinds of "facts" that must obtain or be assumed for the result to obtain.

Of the various fixed points in the Kripke hierarchy, neither the maximal fixed points nor the greatest intrinsic fixed point have any correlate in the dependence function for individual sentences, for the simple reason that they are characterized not by their treatment of any particular sentence, but by how they treat the whole set of sentences. For a particular sentence, there is no discernible difference between

its being true in a maximal fixed point and its being true in an arbitrary fixed point which is extended by that point, and similarly for the intrinsic fixed point. These exceptions aside, the dependence function approach yields a much more unified account of the whole hierarchy, in which debate about whether to call a given sentence true or not can be reduced to questions about the desirable characteristics of the kind of facts to which truth should be reducible.

6.2 Dependence and Gupta's Revision Rule

Not only does the dependence function unify the different parts of the Kripke construction: it also clarifies the relation between that construction and the revision rule theory. One question, for instance, is how close the latter theory is to the m-c supervaluation scheme, since they are similar in their treatment of sentence like $\underline{A} \vee \tilde{} \underline{A}$. Before discussing that question, however, I shall first establish a fundamental theorem relating the revision rule and the dependence function. The theorem closely resembles Lemma 15, but there are some important restrictions at limit ordinals which did not apply there.

Lemma 21: If $D(\underline{A},\underline{B})$ then $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+1}(\underline{U})$ iff $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$, for all successor ordinals α.

Proof: See Appendix.

Lemma 22: $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$ iff $(\underline{B})$(If $C(\underline{A},\underline{B})$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$), for all successor ordinals α.

Proof: See Appendix

In general, since $\underline{U}$ need not be consistent, complete, nor even faithful, that is, it may contain $\tilde{}\underline{R}\underline{t}_1 \ldots \underline{t}_n$, when $\langle \underline{v}(\underline{t}_1) \ldots \underline{v}(\underline{t}_n) \rangle \in \underline{v}(\underline{R})$, these lemmas are not true at $\alpha=0$, nor, because of the limit rule, when

α is a limit ordinal. However, for successor cases, they give the fundamental theorem:

Theorem 23: If α is a non-successor ordinal, then for all $m<n$, $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+n}(\underline{U})$ iff $(\underline{B})$(If $\underline{B}\in\underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$).

Proof: See Appendix.

One of the most frequently used forms of Theorem 23 is:

$$\underline{P}\in\tau_{\underline{M}}^{\alpha+n}(\underline{U}) \text{ iff } (\underline{B})(\text{If } \underline{B}\in\underline{R}(\{\{\underline{P}\}\},n-1) \text{ then } \underline{B} \text{ Int } \tau_{\underline{M}}^{\alpha+1}(\underline{U})). \qquad (6.21)$$

Thus $\underline{P}$ belongs in the extension of $\underline{T}$ at one level just in case there is a level in its dependence function where every member intersects with the extension of $\underline{T}$ at a level one after the last non-successor ordinal. Clearly, as with the Kripke hierarchy, revision upwards is closely related to dependence downwards.

Having established this relationship, we can turn to the relationship between stability and c-closure.

Lemma 24: If $\underline{R}(\{\{\underline{P}\}\})$ is c-closed then for some m and any successor ordinal α, $(\underline{B})(\underline{U})$(If $\underline{B}\in\underline{R}(\{\{\underline{P}\}\},m)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$).

Proof: See Appendix.

We can now readily obtain a theorem relating c-closure and stability:

Theorem 25: If $\underline{R}(\{\{\underline{P}\}\})$ is c-closed, then $\underline{P}$ is locally stably true at every successor limit ordinal for any starting set, where a successor limit ordinal has a last preceding limit ordinal or is the first limit ordinal.

Proof: Assume $\underline{R}(\{\{\underline{P}\}\})$ is c-closed. By Lemma 24 and (6.21) we can readily obtain $\underline{P}$ $\tau_{\underline{M}}^{\alpha+n}(\underline{U})$, for any limit ordinal α (or $\alpha=0$) and any $n>m$, and any $\underline{U}$. Consequently, at the next limit ordinal, $\underline{P}$ is locally stably true for any $\underline{U}$.

Unfortunately, this is as close as we can come in connecting the two. There are sentences which are locally stably true at successive limit ordinals, but not stably true. Consider the example $\underline{T}'\sim\underline{Ta}'\vee\underline{T}'\underline{Ta}'$, where $\underline{a}$ is a Liar sentence. This has the dependence function:

$$\begin{array}{l} \underline{T}'\sim\underline{Ta}'', \underline{T}'\underline{Ta}' \\ \quad \sim\underline{Ta}, \underline{Ta} \\ \quad\ \underline{Ta}, \sim\underline{Ta} \\ \qquad \vdots \end{array} \tag{6.22}$$

Clearly this is c-closed. On the other hand, if we take the starting set $\underline{U}=\mathbb{A}$, we obtain the following stages:

α	$\tau_M^{\alpha}(\underline{U})$	Remainder (6.23)
0	-	$\underline{Ta},\underline{T}'\underline{Ta}',\sim\underline{Ta},\underline{T}'\sim\underline{Ta}'\ldots$
1	$\sim\underline{Ta},\ldots$	$\underline{Ta},\underline{T}'\underline{Ta}',\underline{T}'\sim\underline{Ta}',\ldots$
2	$\underline{Ta},\underline{T}'\sim\underline{Ta}',\underline{T}'\sim\underline{Ta}'\ \underline{T}'\underline{Ta}',\ldots$	$\sim\underline{Ta},\underline{T}'\underline{Ta}',\ldots$
3	$\sim\underline{Ta},\underline{T}'\underline{Ta}',\underline{T}'\sim\underline{Ta}'\ \underline{T}'\underline{Ta}',\ldots$	$\underline{Ta},\underline{T}'\sim\underline{Ta}',\ldots$
ω	$\underline{T}'\sim\underline{Ta}'\ \underline{T}'\underline{Ta}'$	$\underline{Ta},\sim\underline{Ta},\underline{T}'\underline{Ta}',\underline{T}'\sim\underline{Ta}',\ldots$
ω+1	$\sim\underline{Ta},\ldots$	$\underline{Ta},\underline{T}'\sim\underline{Ta}',\underline{T}'\underline{Ta}',\underline{T}'\underline{Ta}'\ \underline{T}'\sim\underline{Ta}',\ldots$
ω+2	$\underline{Ta},\underline{T}'\sim\underline{Ta}',\underline{T}'\sim\underline{Ta}'\ \underline{T}'\underline{Ta}',\ldots$	$\sim\underline{Ta},\underline{T}'\underline{Ta}',\ldots$

etc.

After a moment's hesitation, $\underline{T}'\sim\underline{Ta}'\vee\underline{T}'\underline{Ta}'$ settles down, and at ω is locally stably true. Neither $\underline{Ta}$ nor $\sim\underline{Ta}$, however, is locally stably true, and their fate is decided by $\underline{U}$: both go out of $\tau_M^{\omega}(\underline{U})$. Consequently, at ω+1, $\underline{T}'\sim\underline{Ta}'\vee\underline{T}'\underline{Ta}'$ is not in $\tau_M^{\alpha}(\underline{U})$, and this pattern is repeated at each limit ordinal in turn. Consequently, $\underline{T}'\sim\underline{Ta}'\vee\underline{T}'\underline{Ta}'$, though it has a c-closed dependence function, is not stably true.

Clearly the divergence between stable truth and c-closure is here largely tied up with the limit rule being used: if the rule did not use $\underline{U}$ to decide the fate of $\underline{Ta}$ and $\sim\underline{Ta}$, it might be possible to keep $\underline{T}'\sim\underline{Ta}'\vee\underline{T}'\underline{Ta}'$ in $\tau_M^{\alpha}(\underline{U})$, and then this example at least would not show the non-equivalence of c-closure and stability.

We can approach a new limit rule which does tie the two together more closely by first considering a generalization of Gupta's rule. Overall, the revision process contains a random element, represented by the arbitrariness of $\underline{U}$, and a regulated element, the step-by-step revision process. After the latter has gone as far as it can, the random element is reintroduced, and the process is repeated until it settles down to a steady repetition. Belnap's suggestion in "Gupta's Rule of Revision Theory of Truth" is that we maximize the random element by introducing a "bootstrapping policy". A bootstrapping policy is just a function Γ from the limit ordinals to the power set of the sentences of $\underline{L}$, which we use to settle the extension of $\underline{T}$ at limit ordinals, just as before we used $\underline{U}$. Thus we can define a new function $\tau_{\underline{M}}^{\alpha}(\Gamma)$, parallel to Definition 2.

Definition 26:

For $\alpha=0$ $\quad \tau_{\underline{M}}^{\alpha}(\Gamma)=\Gamma(\alpha)$

$\alpha=\beta+1$ $\quad \tau_{\underline{M}}^{\alpha}(\Gamma)=\{\underline{P}:\underline{P}$ is true in $\tau_{\underline{M}}^{\beta}(\Gamma)\}$

α a limit ordinal

$$\tau_{\underline{M}}^{\alpha}(\Gamma)=\underline{X}\cup(\Gamma(\alpha)-\underline{Y})$$

where $\underline{X}=\{\underline{P}:\underline{P}$ is locally stably true at α for $\Gamma\}$

$\underline{Y}=\{\underline{P}:\underline{P}$ is locally stably false at α for $\Gamma\}$

Naturally, 'locally stably true (false) at α for Γ' is defined analogously to the definition given for $\underline{U}$. Further, we can define stably true in Γ and bootstrap stably true parallel to Definitions 3 and 4.

What is perhaps surprising is that this rather radical random element does not greatly change the set of stably true sentences: a few sentences are stably true for Gupta's limit rule, but not bootstrap stably true. Naturally, since the Gupta limit rule is a special case of a bootstrapping policy, corresponding to taking the constant function $\Gamma(\alpha)=\underline{U}$, $\underline{T}'\sim\underline{Ta}'\vee\underline{T}'\underline{Ta}'$ is no more bootstrap stably true than it is stably

true. However, a plausible restriction on bootstrapping policies brings us much closer. This is the restriction to acceptable bootstrappers. A bootstrapping policy is acceptable just in case $\tau_{\underline{M}}^{\alpha}(\Gamma)$ is consistent and complete at each limit ordinal α. This limit rule is mentioned by Gupta in fn. 10 of "Truth and Paradox".

The intuition behind this restriction is partly derived from examples like $\underline{T}'\tilde{}\underline{Ta}'\vee\underline{T}'\underline{Ta}'$, and partly from general considerations about consistency and completeness. As we saw, in the revision stages either $\underline{T}'\tilde{}\underline{Ta}'$ or $\underline{T}'\underline{Ta}'$ was always in the extension of $\underline{T}$, and it was essentially this fact that led to $\underline{T}'\tilde{}\underline{Ta}'\vee\underline{T}'\underline{Ta}'$ being locally stably true at ω. At the limit, however, the rule said "Keep $\underline{T}'\tilde{}\underline{Ta}'\vee\underline{T}'\underline{Ta}'$, but throw out $\underline{T}'\tilde{}\underline{Ta}'$ and $\underline{T}'\underline{Ta}'$", even though the latter pair were responsible for the disjunction being there at all. Furthermore, since the revision process always gives a consistent and complete extension, and the revision process is so central to the way we calculate the extension, it is plausible to suggest that the limit rule should also give a consistent and complete extension.

It is useful to note an important distinction: a bootstrapper Γ can be such that $\Gamma(\alpha)$ is always consistent and complete, without being acceptable. Suppose, for instance, $\Gamma(\omega)$ contained $\tilde{}\underline{T}'\underline{Ta}'$, $\tilde{}\underline{T}'\tilde{}\underline{Ta}'$ and $\tilde{}(\underline{T}'\tilde{}\underline{Ta}'\vee\underline{T}'\underline{Ta}')$, as it well might. In this case $\tau_{\underline{M}}^{\omega}(\Gamma)$ will contain $\underline{T}'\tilde{}\underline{Ta}'\vee\underline{T}'\underline{Ta}'$, because that is locally stably true, but also $\tilde{}\underline{T}'\underline{Ta}'$ and $\tilde{}\underline{T}'\tilde{}\underline{Ta}'$, because they are in the bootstrapper. Whether a given bootstrapper is acceptable can only be determined by sitting down and working through the revision process stage by stage.

Given acceptable bootstrappers, then, we can establish closer connections between stability and c-closure. Clearly, all the preceding proofs for arbitrary $\underline{U}$ hold for arbitrary Γ. To go on, we need some preliminary definitions and lemmas.

Definition 27: If $\underline{B}_i$ is a finite set of sentences $\{\underline{P}_1 \ldots \underline{P}_n\}$, then $\underline{disj}(\underline{B})$ is the disjunction $\underline{P}_1\vee\underline{P}_2\vee\ldots\vee\underline{P}_n$.

If $\underline{R}(\{\underline{A}\},n)=\{\underline{B}_1 \ldots \underline{B}_m\}$, then $\underline{conj}(\underline{A},n)$ is the conjunction $\underline{disj}(\underline{B}_1)\&\ldots\&\underline{disj}(\underline{B}_m)$.

Lemma 28: For all n, $\underline{P}$ is locally stably true at limit ordinal α for Γ iff $\underline{conj}(\{\underline{P}\},n)$ is locally stable true at α for Γ.

Proof: See Appendix.

Henceforth I shall assume that Γ is an acceptable bootstrapper, unless there are explicit remarks to the contrary. Two important results follow, one a result internal to the revision theory, but unavailable without the dependence function, and the second relating c-closure and stability for acceptable bootstrappers, or a-stability.

Theorem 29: If $\underline{P}$ is locally stably true at limit ordinal α in Γ, $\underline{P}$ is stably true in Γ.

Proof: Assume $\underline{P}$ is locally stably true at α in Γ.
I shall show that, for all $\gamma>\alpha$, $\underline{P}\in\tau_{\underline{M}}^{\gamma}(\Gamma)$, by induction.
Induction Hypothesis
(β)(If $\alpha\leqslant\beta<\gamma$ then $\underline{P}\in\tau_{\underline{M}}^{\beta}(\Gamma)$)

1. γ is a limit ordinal.
   By definition, and IH, $\underline{P}\in\tau_{\underline{M}}^{\gamma}(\Gamma)$
2. γ is a successor $=\delta+n$, where δ is a limit ordinal $\delta>\alpha$, and $n>0$.
   So $\underline{P}$ is locally stably true at δ in Γ.
   Consequently $\underline{conj}(\{\underline{P}\},n)$ is locally stably true at δ in Γ, by Lemma 28.
   So $\underline{conj}(\{\underline{P}\},n)\in\tau_{\underline{M}}^{\delta}(\Gamma)$.
   By consistency and completeness, however,
   $(\underline{B})$(If $\underline{B}\in\underline{R}(\{\underline{P}\},n)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\delta}(\Gamma)$).
   Moreover, in acceptable starting points,
   the restriction in Theorem 23 to $m<n$ can be lifted.
   Hence $\underline{P}$ $\tau_{\underline{M}}^{\delta+n}(\Gamma)$.

Hence $\underline{P}\in\tau_{\underline{M}}^{\gamma}(\Gamma)$, for all γ, i.e., $\underline{P}$ is stably true in Γ.

It is crucial to recognize that the reason Lemmas 21 and 22 were restricted to successor ordinals, and that Theorem 23 had the restriction m<n was the possibility that at limit ordinals $\tau_{\underline{M}}^{\alpha}(\Gamma)$ was neither consistent nor complete. In acceptable bootstrappers, this possibility no longer exists, so that these lemmas and the theorem hold more generally: the only restriction is that $\alpha \neq 0$ in 21, and that m<n and $\alpha+n-m \neq 0$ in 23. These last remaining restrictions hold because although $\Gamma(0)$ is consistent and complete, we have done nothing to make it faithful to the model. Hence knowing that $\underline{A}$ Int $\tau_{\underline{M}}^{1}(\Gamma)$ does not necessarily mean that if $D(\underline{A},\underline{B})$, $\underline{B}$ Int $\tau_{\underline{M}}^{0}(\Gamma)$. Of course, the restriction m<n is to ensure that $\alpha+n-m$ is well-defined.

Finally we obtain the relation between a-stability and c-closure.

Theorem 30: If $\underline{R}(\{\{\underline{P}\}\})$ is c-closed, $\underline{P}$ is a-stably true.

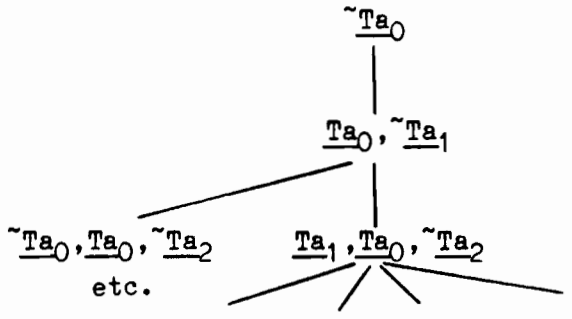
Proof: If $\underline{R}(\{\{\underline{P}\}\})$ is c-closed, $\underline{P}$ is locally stably true at w in Γ, for any Γ, by Theorem 25. Hence, by Theorem 29, $\underline{P}$ is stably true in Γ, for any Γ. Hence $\underline{P}$ is a-stably true.

Unfortunately, the converse does not hold: there are sentences which are a-stably true, but which do not have c-closed dependent functions. One example, due to Gupta, is the following. Suppose $\underline{M}$ is such that:

$$\underline{v}(\underline{a}_i) = \tilde{}\underline{Ta}_0 \& \underline{Ta}_{i+1}, \quad i \geq 0. \tag{6.24}$$

Then we can represent the dependence function for $\tilde{}\underline{Ta}_0$ as the following tree, in which some descendants are suppressed.

(6.25)



etc.

Clearly, at each line there will be a collection of sets containing $\underline{Ta}_0$ and $\underline{\sim Ta}_0$, and one containing $\underline{Ta}_0 \ldots \underline{Ta}_n$, $\underline{\sim Ta}_{n+1}$. By definition, then, the dependence function is not c-closed. However, it can be shown that $\underline{\sim Ta}_0$ is a-stably true, as follows. Take any Γ: either $\Gamma(0)$ is such that $\underline{Ta}_i \in \tau^1_{\underline{M}}(\Gamma)$, for some i, or none of them do. Suppose the former. By inspection, for all $n>i+1$, the sets belonging to $\underline{R}(\{\underline{\sim Ta}_0\},n)$ either contain $\underline{Ta}_0$ or $\underline{\sim Ta}_0$, or they contain $\underline{Ta}_i$. Either way, every member intersects with $\tau^1_{\underline{M}}(\Gamma)$. Hence for all $n>i+2$, $\underline{\sim Ta}_0 \in \tau^n_{\underline{M}}(\Gamma)$ and so $\underline{\sim Ta}_0$ is locally stably true at w in Γ. Suppose none of the $\underline{Ta}_i$ are in $\tau^1_{\underline{M}}(\Gamma)$: then every set in the dependence function intersects $\tau^1_{\underline{M}}(\Gamma)$, since each set contains $\underline{\sim Ta}_j$ for some j. So again $\underline{\sim Ta}_0$ is locally stably true at w in Γ. But then, by Theorem 29, $\underline{\sim Ta}_0$ is stably true in all Γ, i.e. $\underline{\sim Ta}_0$ is a-stably true.

This example points up an important difference between Kripke grounded sentences and a-stable sentences. For $\underline{\sim Ta}_0$, no matter what extension of $\underline{T}$ you start with, $\underline{\sim Ta}_0$ will become true and stay true after a finite number of revision steps. However, there is no finite limit by which every extension you start with will make $\underline{\sim Ta}_0$ true. In groundedness, however, there has to be a finite level by which all starting sets make a given sentence true.

Some reflection on the relationship of revision processes, either Kripke- or Gupta-style, to the dependence function enables us to think of this difference in a fruitful way, and leads us to a different closure rule. Initially, the dependence function was based on an intuition which was, in a way, static. Given a model which fixed everything but the interpretation of $\underline{T}$, all we could say was that if a

given sentence was to belong to the extension of T, certain other sentences had to as well. However, we now can see whether revision of a given interpretation of T will lead, ultimately, to one in which our given sentence is true. All we have to do is consider the value of conj({P},i), if P is our starting sentence. If there comes a point in the sequence after which all the conj({P},i) are true given the initial interpretation, we know that revision will make P true in the final interpretation. The difference between Kripke grounding and a-stability amounts to this: for Kripke grounding, it is necessary that one particular conj({P},i) have an appropriate property which guarantees that all the following levels turn out true, and with a-stability, this need not be so.

We can define a new closure rule by considering how one might try to find an extension which could not be revised to make the sentence P true. Suppose we start with P, and go through R({{P}}) assigning values to the sentences occurring there in such a way as to try to make the occasional conj({P},i) false. If every such attempt fails, that is, if there comes a point in each attempt when it is no longer possible to make any of the conj({P},i) false, then R({{P}}) is closed. If, on the other hand, we can systematically assign values to the sentences so that at every level there is a lower level where conj({P},i) is false, then R({{P}}) is open. This leads to the following definition.

Definition 31: R({A}) is d-open in M iff there is an infinite sequence $B=\langle B_1, B_2 \ldots\rangle$ which satisfies the following conditions.

1. Every B_i is in R({A},j) for some j.
2. For every i, there is a j,k such that B_j is in R({A},k) and k>i.
3. No B_i contains a sentence P such that $val_M(P)=t$.
4. $\cup B$ is consistent.

Otherwise, R({A}) is d-closed in M.

Using this definition we can see that $\underline{R}(\{\{\sim\underline{Ta}_0\}\})$ is d-closed, because only one member at each level could be used to build the infinite set, and if we pick any one such, then there is no subsequent set which can be used to form the appropriate consistent set for condition (4). We can go on to show that d-closure implies a-stable truth:

Theorem 32: If $\underline{R}(\{\{\underline{P}\}\})$ is d-closed, then $\underline{P}$ is a-stably true.

Proof: Suppose $\underline{R}(\{\{\underline{P}\}\})$ is d-closed, and suppose, for *reductio*, that there is an infinite number of levels for which $\underline{conj}(\{\underline{P}\})\in\tau_{\underline{M}}^{1}(\Gamma)$, for some Γ. Then for each such level m, there is a $\underline{B}\in\underline{R}(\{\{\underline{P}\}\},m)$, for which $\underline{B}\ \tau_{\underline{M}}^{1}(\Gamma)=\Lambda$. Consequently, there is an infinite $B = \langle\underline{B}_1,\ldots\rangle$, each of whose members is in $\underline{R}(\{\{\underline{P}\}\},m)$, for some m, none of whose members contains a sentence true in the model, and for which $\cup B$ is consistent. Thus $\underline{R}(\{\{\underline{P}\}\})$ is d-open. But this contradicts the hypothesis, so there can only be a finite number of levels where $\underline{conj}(\{\underline{P}\})\in\tau_{\underline{M}}^{1}(<\Gamma)$. Suppose the last such is m. Then, for all n>m, $\underline{conj}(\overline{\{\underline{P}\}},n)\in\tau_{\underline{M}}^{1}(\Gamma)$. But then by Definition 27 and (6.21), for all n>m, $\underline{P}\in\tau_{\underline{M}}^{n+1}(\Gamma)$. Consequently $\underline{P}$ is locally stably true at ω in Γ, and by Theorem 29, is stably true in Γ, and, since Γ was arbitrary, $\underline{P}$ is a-stably true.

Further, we can also show that d-openness and instability have a connection:

Theorem 33: If $\underline{R}(\{\{\underline{P}\}\})$ is d-open, $\underline{P}$ is not locally stably true at ω for some Γ.

Proof: Suppose $\underline{R}(\{\{\underline{P}\}\})$ is d-open. Then there is a set B of the specified kind. Construct $\Gamma(0)$ in such a way that $B\cap\Gamma(0)=\Lambda$, and it is faithful to $\underline{M}$, that is, among the literals which are not truth ascriptions, $\underline{P}$ is in $\Gamma(0)$ just in case $\underline{P}$ is true in the model. Now consider an arbitrary $\underline{B}_i\in B$: let it belong to $\underline{R}(\{\{\underline{P}\}\},m)$. As usual, there is a $\underline{C}$ and a $\underline{D}$ such that $\underline{C}\in\underline{R}(\{\{\underline{P}\}\},m-1)$, $D(\underline{C},\underline{D})$ and $C(\underline{D},\underline{B}_i)$. By Lemma 22, $\underline{D}\cap\Gamma(0)=\Lambda$. Moreover, since $\Gamma(0)$ is faithful, $\underline{C}\cap\tau_{\underline{M}}^{1}(\Gamma)=\Lambda$. Consequently,

not all members of $\underline{R}(\{\{\underline{P}\}\},m-1)$ intersect $\tau^1_{\underline{M}}(\Gamma)$. But this is true for an infinite number of levels. Hence for an infinite number of levels, $\underline{P} \notin \tau^n_{\underline{M}}(\Gamma)$.

Unfortunately, again there is a counterexample to the claim that $\underline{R}(\{\{\underline{P}\}\})$ is d-closed iff $\underline{P}$ is a-stably true. There are sentences which are locally unstable at ω in some Γ, but stabilize thereafter. The example which I shall give is extremely complicated, and I shall approach the matter gently by considering the infinite series of sentences $\underline{a}_i$ given in (6.19). Clearly, $\underline{R}(\{\{\underline{Ta}_0\}\})$ is d-open, since there are numerous ways of picking consistent sets from different levels. However, I want to look carefully at the set $\{\underline{Ta}_1, \underline{Ta}_2, \underline{Ta}_4, \underline{Ta}_7 \ldots\}$, whose i-th member has index $i(i-1)/2 + 1$. Suppose $\Gamma(0)$ contains all the $\underline{Ta}_j$ except those ones. Clearly $\underline{Ta}_0$ is unstable at ω in Γ: there is an infinite number of levels where $\underline{conj}(\{\underline{Ta}_0\})$ does not intersect with $\tau^1_{\underline{M}}(\Gamma)$. Consequently, each and every $\underline{Ta}_j$ is also unstable at ω in Γ. However every disjunction $\underline{Ta}_j \vee \underline{Ta}_k$, $j \neq k$, is locally stably true at ω in Γ. To see how this happens, consider $\underline{Ta}_0 \vee \underline{Ta}_2$, and its dependence function:

$$\begin{array}{l} \underline{Ta}_0, \underline{Ta}_2 \\ \underline{Ta}_1, \underline{Ta}_3 \\ \underline{Ta}_2, \underline{Ta}_4 \\ \underline{Ta}_3, \underline{Ta}_5 \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \end{array} \qquad (6.26)$$

The following sentences are not in $\tau^1_{\underline{M}}(\Gamma)$: $\underline{Ta}_0, \underline{Ta}_1, \underline{Ta}_3, \underline{Ta}_6$, and the i-th has index $i(i-1)/2$. Inspection of $\underline{R}(\{\{\underline{Ta}_0\ \underline{Ta}_2\}\})$ shows that only the pair $\underline{Ta}_1$, $\underline{Ta}_3$ are both out of $\tau^1_{\underline{M}}(\Gamma)$: for every other pair given, one or other is in $\tau^1_{\underline{M}}(\Gamma)$. The reason for this is simple: for a given pair in this dependence function, $\underline{Ta}_i$, $\underline{Ta}_{i+2}$, both to be out of $\tau^1_{\underline{M}}(\Gamma)$, we have to be able to find a value of i in the series

0,1,3,6,...i(i-1)/2,... where the difference between the two terms is exactly two. But, in general, the difference between the (i-1)th term and the i-th term is i-1. Consequently, for any i>2, the difference between successive terms is bound to be greater than 2, and hence one or other of $\underline{Ta}_i$, $\underline{Ta}_{i+2}$ is bound to be in $\tau^1_{\underline{M}}(\Gamma)$.

This result can be extended quite generally to any pair $\underline{Ta}_j$, $\underline{Ta}_k$ where k-j=m: once past a certain point, no pair $\underline{Ta}_{j+n}$, $\underline{Ta}_{k+n}$ can possibly both be out of $\tau^1_{\underline{M}}(\Gamma)$. So all the $\underline{Ta}_j \vee \underline{Ta}_k$ are locally stably true at w in $\mathcal{A}$. This means that when $\Gamma(w)$ comes to distribute the $\underline{Ta}_i$ to give $\tau^w_{\underline{M}}(\Gamma)$, it must respect all those stabilities, and as a result, at most one of them can be left out of $\tau^w_{\underline{M}}(\Gamma)$. If two were, say the j-th and k-th, then $\tau^w_{\underline{M}}(\Gamma)$ would not be consistent, since it would have $\underline{Ta}_j \vee \underline{Ta}_k$, and also $\sim\underline{Ta}_j$ and $\sim\underline{Ta}_k$. Hence $\underline{Ta}_0$ is bound to be locally stably true at $w.2$ in Γ, and hence stably true in Γ.

Now this argument does not mean that $\underline{Ta}_0$ is a counterexample to the claim that a-stability and d-closure are equivalent, because $\underline{Ta}_0$ is not a-stable and has a d-open dependence function. However, this argument forms the basis for constructing a counterexample. In the following, the various denotations are designed so that the dependence function is d-open, but the various bootstrappers which can make the sentence locally unstable at w have the characteristic of the set $\{1,2,4,7,\ldots i(i-1)/2+1,\ldots\}$ that successive members are further and further apart. Suppose $\underline{M}$ contains denotations as follows:

$$\text{i)}\quad \underline{v}(\underline{a}_{i,j}) = \underline{Ta}_{i,j+1} \qquad 1 \leqslant i,j < \omega \tag{6.27}$$

$$\text{ii)}\quad \underline{v}(\underline{b}_0) = \underline{Tb}_1 \vee \underline{Ta}_{1,1}$$

$$\underline{v}(\underline{b}_i) = \underline{T}^{i+1}\underline{b}_{i+1} \vee \underline{T}^{i+1}\underline{a}_{i+1,1} \vee (\sim\underline{T}^i\underline{a}_{i,1} \vee \sim\underline{T}^{i-1}\underline{a}_{i,1} \ldots \vee \sim\underline{Ta}_{i,1})$$

where $\underline{T}^n\underline{t} = \underline{T}'\underline{T}'\ldots\underline{t}''$ ' with n occurrences of $\underline{T}$.

Let me sort out this complicated apparatus. Each $\underline{a}_{i,1}$ generates an

infinite chain just like $\underline{a}_0$ above: we just have an infinite set of such chains. They serve two purposes: to prevent the $\underline{b}_i$ from ending up grounded, and to permit the kind of argument used above. The $\underline{b}_i$, on the other hand, are what force the $\Gamma(0)$ for which our sentence will be unstable to have the characteristic spacing. Omitting their dependence on the $\underline{a}_i$, the $\underline{b}_i$ have dependences as follows:

$$\underline{Tb}_0 \tag{6.28}$$

$$\underline{Tb}_1,\ \underline{Ta}_{1,1}$$

$$\underline{T}'\underline{Tb}_2{}',\ \underline{T}'\underline{Ta}_{2,1}{}',\ \sim\underline{Ta}_{1,1},\ \underline{Ta}_{1,2}$$

etc.

$$\underline{Tb}_i \tag{6.29}$$

$$\underline{T}^{i+1}\underline{b}_{i+1},\ \underline{T}^{i+1}\underline{a}_{i+1,1},\ \sim\underline{T}^{i}\underline{a}_{i,1},\ \sim\underline{T}^{i-1}\underline{a}_{i,1}\cdots\sim\underline{Ta}_{i,1}$$

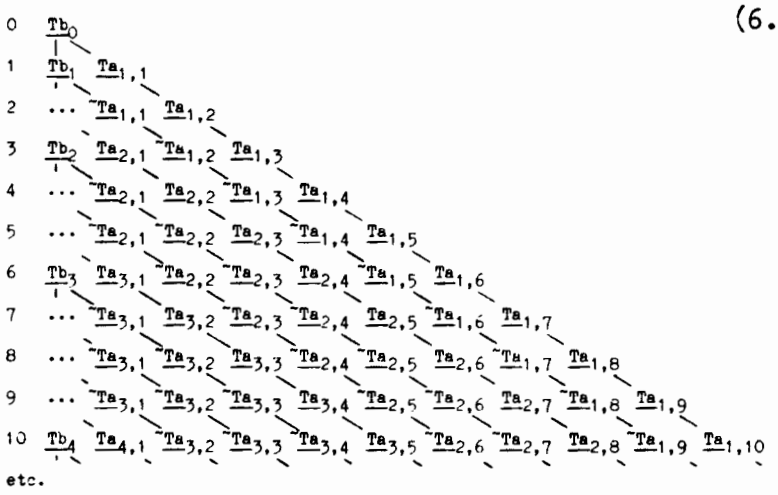
$$\underline{T}'\underline{Tb}_{i+1}{}',\ \underline{T}'\underline{Ta}_{i+1,1}{}',\ \sim\underline{Ta}_{i,1}\ +\text{terms in } \sim\underline{Ta}_{i,j}$$

$$\underline{Tb}_{i+1},\ \underline{Ta}_{i+1,1},\ \sim\underline{Ta}_{i,2}\ +\text{terms in } \sim\underline{Ta}_{i,j}$$

etc.

Thus if $\underline{b}_{i+1}$ first occurs at the first line, $\underline{b}_{i+2}$ first occurs at the (i+1)-th line, and similarily for $\underline{a}_{i+1,1}$ and $\underline{a}_{i+2,1}$. One crucial feature of this dependence is that every line containing $\underline{T}^j\underline{b}_{i+1}$, $j>1$ also contains an occurrence of $\sim\underline{Ta}_{i,1}$, but $\underline{Tb}_{i+1}$ has $\underline{Ta}_{i+1,1}$. A summary of the dependence function of $\underline{Tb}_0$ follows, with unimportant terms in $\underline{T}^j\underline{b}_i$ and $\underline{T}^j\underline{a}_i$ omitted (the omission being marked by dots).

$\underline{R}(\{\{\underline{Tb}_0\}\})$: (6.30)



First we need to show that $\underline{R}(\{\{\underline{Tb}_0\}\})$ is d-open. The set B contains the unique set $\underline{B}_m$ from $\underline{R}(\{\{\underline{Tb}_0\}\},m)$ where m has the values $0,1,3,6,\ldots n(n+1)/2$, $0\leqslant n<\omega$. In other words, we take all the levels where there is an occurrence of $\underline{Tb}_i$. Clearly B satisfies conditions (1) and (2) in Definition 31. To show that it satisfies the third, consider the i-th member of B. It contains:

$$\underline{Tb}_i,\underline{Ta}_{i,1},\sim\underline{Ta}_{i-1,2}\cdots\underline{Ta}_{i-1,i+1}\cdots\sim\underline{Ta}_{i-j,n}\cdots\underline{Ta}_{1,p}. \tag{6.31}$$

We are particularily interested in the lowest and highest values of n for which $\underline{Ta}_{i-j,n}$ occur, whether negated or unnegated. For j=0, they are 1 and 1, for j=1, they are 2 and i+1, and in general they are $2+(j-1)(2i-j+2)/2$ and $1+j(2i-j+1)/2$, for $j\geqslant 1$. Now to show consistency, we can show that no sentence occurring in the i-th member of $\underline{B}$ occurs in the (i+1)-th member. Clearly none of the $\underline{Tb}_i$ are the same. Furthermore, the highest n on $\underline{Ta}_{i-j,n}$ in the i-th member is $1+j(2i-j+1)/2$, and the lowest on $\underline{Ta}_{(i+1)-(j+1),n}$ in the (i+1)-th member is obtained by replacing i by i+1 and j by j+1 in $2+(j-1)(2i-j+2)/2$, namely $2+j(2i-j+3)/2$: the difference between these is j+1.

Since none of the members of B have any members in common, $\cup B$ is consistent. So $\underline{R}(\{\{\underline{Tb}_0\}\})$ is d-open, and hence $\underline{Tb}_0$ is not locally stably true at Ψ for some Γ. Consider the Γ determined by B. Let it

be Γ'. I shall show that $\underline{Tb_0}$ is stable at $\omega.2$ in Γ', and that all Γ in which it is unstable, the same occurs.

Clearly, at ω, all the $\underline{conj}(\{\underline{Tb_0}\},i)$ are unstable in Γ'. However, by exactly similar reasoning to the case for the $\underline{Ta_i}$ before, every disjunction $\underline{conj}(\{\underline{Tb_0}\},i)\vee\underline{conj}(\{\underline{Tb_0}\},j)$ is stably true at ω, since there comes a stage where both disjuncts cannot simultaneously be out of $\tau^1_{\underline{M}}(\Gamma)$. Consequently, $\underline{Tb_0}$ will be locally stably true at $\omega.2$ in Γ', and stably true in Γ'.

However, inspection reveals that B, and hence Γ', has a special property. Consider another set of the $\underline{B_m}$, B', whose union is consistent. Let the members of this set be $\underline{B_0}'\ldots\underline{B_j}'\ldots$,and let the level at which they occur be $m(0),\ldots m(j),\ldots$. Then, if $m(j)$ lies between $n(n+1)/2$ and $(n+1)(n+2)/2$ then $m(j+1)-m(j)$ is at least $n+1$. That is, no other consistent set can choose sets from levels which are closer than the corresponding levels in B. Hence every Γ in which $\underline{Tb_0}$ is locally unstable at ω has the spacing property at $\Gamma(0)$, and hence leads to $\underline{Tb_0}$ being stable.

One interesting observation about $\underline{Tb_0}$ is that it shows that even in the propositional case ascent to levels higher that ω may be necessary for a sentence to stabilize. This is different from the Kripke case, where sentences always grounded at a finite level.

Thus it turns out that even d-closure and a-stability are not equivalent. Two points about d-closure should be made, however. First, it seems to be considerably closer to a-stability than c-closure. This can be made specific in the following way. I described a Thomason model as one in which all starting points led to a unique truth-extension, and remarked that they represented a plausible measure of pathologicality: no genuinely pathological self-reference can occur in a Thomason model. Now if the only self-reference in a model $\underline{M}$ is of the order characterized by (6.24), then the model is Thomasonian: no matter what

starting set one chooses, every one of the $\sim\underline{Ta}_i$ is a-stably true, and, by supposition, every other sentence is either a-stably true or a-stably false. Hence c-closure disagrees with a-stability on sentences which can get values in Thomason models, and since there is such a strong intuition that the values sentences obtain in a Thomason madel are the right values, untainted by pathology, c-closure seems an inadequate closure rule. On the other hand, although $\underline{Td}_0$ is always in the extension of $\underline{T}$, it depends on sentences which can themselves stay unstable. Supposing, for instance, that in Γ' at w we put all the $\underline{T}^n\underline{b}_i$ in the extension of $\underline{T}$. Then we are at liberty to leave all the $\underline{Ta}_{i,j}$ out, or to distribute them here and there. This liberty can be continued at higher levels, so the $\underline{Ta}_{i,j}$ need never be stably true. Consequently, $\underline{Tb}_0$ cannot have a value in a Thomason model, and though I have not shown that no sentence whose dependence function is d-open can obtain a value in a Thomason model, it is doubtful that there could be one.

The second point relating d-closure to a-stability is the following. We can easily modify the definition of d-closure so that a-stability implies the new closure property:

Definition 34: $\underline{R}(\{\{\underline{P}\}\})$ is e-open iff it is d-open and successive levels in $\underline{R}(\{\{\underline{P}\}\})$ from which the members of some B are drawn differ by less than m, for some m. Otherwise $\underline{R}(\{\{\underline{P}\}\})$ is e-closed.

Theorem 35: If $\underline{P}$ is a-stably true, then $\underline{R}(\{\{\underline{P}\}\})$ is e-closed.

Proof: (Sketch only.) Suppose $\underline{R}(\{\{\underline{P}\}\})$ is e-open. Then $\underline{P}$ is locally unstable at w in Γ, for some Γ, as for d-openness. Furthermore, each m-long conjunction $\underline{conj}(\{\underline{P}\},n)\&\underline{conj}(\{\underline{P}\},n+1)\&\ldots\&\underline{conj}(\{\underline{P}\},n+m-1)$ is locally stably false, because one conjunct will be false at each level. Since each is locally stably false at w, all are stably false in Γ. Hence, at every limit level, one of every m successive $\underline{conj}(\{\underline{P}\},n)$ must be out of $\tau^{\alpha}_{\underline{M}}(\Gamma)$. Hence $\underline{P}$ is unstable in Γ.

Consequently, if $\underline{P}$ is a-stably true, $\underline{R}(\{\{\underline{P}\}\})$ is e-closed.

Unfortunately now the converse is unproven. What we can see, though, is that a-stability lies somewhere between d-closure and c-closure, and the difference between these two is not very great.

The chief open question in the revision process is what the appropriate limit rule is, and on what grounds preferences for different limit rules can be established. What comparison with the dependence function shows is that the revision process and dependence function agree exactly on what happens up to ω, or from one limit level up to the next, but that there is a corresponding difficulty in the downward theory in matching the upward theory.

The problem can be put in the following way. If we take a starting set, and then revise it to ω, we can think of a limit rule as a rule for handing us a new starting set, based on properties of the old one. Now suppose we think of the dependence function for a given sentence, and suppose that sentence is locally unstable at ω in the starting set we have taken. All the changes which the limit rule introduces ought to be derivable from information in the dependence function, at least with respect to the changes which are relevant to the sentence being considered. In fact, in discussing $\underline{Tb}_0$, it was precisely certain features of $\underline{R}(\{\{\underline{Tb}_0\}\})$ which I took to show that $\underline{Tb}_0$ would stabilize by $\omega.2$. However, we still lack a general characterization of the properties the dependence function has to possess which correspond to the action of a limit rule. Nevertheless, without a dependence function it would be virtually impossible to discuss some of the examples given.

Thus, use of the dependence function has enabled us to see how the Kripke and Gupta constructions are related, and to prove several important results both about the relationship of reivsion to dependence and purely about the revision process itself.

6.3 Dependence and 'True'

At this point, it seems necessary to say something about the relationship of the dependence function to natural language uses of 'true'. The claim can be put straightforwardly: all successful (i.e. intuitively correct) uses of 'true', given agreement with the facts, correspond to formal sentences whose dependence functions are d-closed.

I shall first defend my choice of d-closed dependence functions. These functions have the advantage over a- and b-closed functions of making $\underline{T}\text{'}\underline{A}\text{'}\vee\tilde{}\underline{T}\text{'}\underline{A}\text{'}$ true for any $\underline{A}$, and over c-closed functions of getting the right value on $\tilde{}\underline{Ta}_0$ in (6.24). Compared with a-stability, the choice is based on slim differences indeed. Since it is virtually impossible to have an intuitive opinion about the truth of $\underline{Tb}_0$, the choice between them cannot be based on that, and assuming a-stability and d-closure agree on essential cases, there is not much to say one way or another. One small intuitive source of preference for d-closure is that a slight change in the denotations for the $\underline{b}_i$ would make $\underline{Tb}_0$ unstable rather than stably true. Suppose, for instance, the denotations are as follows:

$$\underline{v}(\underline{b}_i')=\underline{T}^m\underline{b}_{i+1}\vee\underline{T}^m\underline{b}_{i+1,1}\vee(\tilde{}\underline{T}^m\underline{a}_{i,1}\vee\tilde{}\underline{T}^{m-1}\underline{a}_{i,1}\cdots\vee\tilde{}\underline{Ta}_{i,1}) \qquad (6.32)$$

Then $\underline{Tb}_0'$ turns out to be d-open, but also e-open, and hence unstable. Since this only seems to differ from $\underline{Tb}_0$ in how long a trail of $\tilde{}\underline{Ta}_{i,1}$ each $\underline{b}_i$ drags after it, it seems surprising that this is not stably true and $\underline{Tb}_0$ is. All this may perhaps emphasize how marginal these cases are in explaining ordinary uses of 'true'!

For ease of exposition, I shall say that a sentence whose dependence function is d-closed is dependence-true or d-true. Further, we can introduce d-false, for sentences whose negations have dependence functions which are d-closed. Now clearly not every sentence is either d-true or d-false: the Liar and Truth-Teller are immediate examples. Furthermore, not even $\tilde{}\underline{T}\text{'}\tilde{}\underline{Ta}\text{'}$ and $\tilde{}\underline{T}\text{'}\underline{Tb}\text{'}$ are either d-true or d-false.

Consequently, the language has gaps: some uses of 'true' and 'false' will be held not to be successful. A familiar response, at this point, is to hold that 'The Liar is not true' is a consequence of my theory, and ought to be successfully sayable. I want to respond to this challenge slowly.

First, it is clear that sentences not in the immediate neighbourhood of pathology are treated in a natural way, including the logical laws. Moreover, for such sentences the Tarski biconditionals are true: $\underline{R}(\{\{\underline{A} \equiv \underline{T}'\underline{A}'\}\},0)$ is

$$\{\underline{A}, \tilde{}\underline{T}'\underline{A}'\} \quad \{\underline{T}'\underline{A}', \tilde{}\underline{A}\} \tag{6.33}$$

So if $\underline{A}$ is d-true, $\underline{T}'\underline{A}'$ will be also, and so $\underline{A} \equiv \underline{T}'\underline{A}'$ will be, and if $\underline{A}$ is d-false, $\tilde{}\underline{A}$ and $\tilde{}\underline{T}'\underline{A}'$ will be d-true, and so will $\underline{A} \equiv \underline{T}'\underline{A}'$.

On the other hand, for the Liar and Truth-Teller we get:

$$\begin{array}{cc} \{\tilde{}\underline{Ta}\} & \{\underline{Ta}\} \\ \{\underline{Ta}\} & \{\tilde{}\underline{Ta}\} \\ \vdots & \end{array} \qquad \begin{array}{c} \{\underline{Tb}, \tilde{}\underline{Tb}\} \\ \{\underline{Tb}, \tilde{}\underline{Tb}\} \\ \vdots \end{array} \tag{6.34}$$

In fact, for the Liar, the Tarski biconditional is d-false, since $\underline{R}(\{\{\tilde{}(\underline{Ta} \equiv \tilde{}\underline{Ta})\}\})$ is:

$$\begin{array}{c} \underline{Ta}, \tilde{}\underline{Ta} \\ \tilde{}\underline{Ta}, \underline{Ta} \\ \vdots \end{array} \tag{6.35}$$

This completely matches our intuition that "$\underline{a}$ is true iff $\underline{a}$ is not true" is false, and the Truth-Teller comes out with a d-true biconditional, reflecting the unenlightening remark that "$\underline{b}$ is true iff $\underline{b}$ is true".

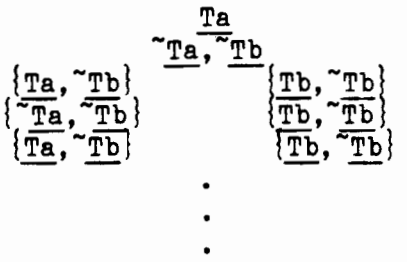
Thus the Tarski biconditionals do not all turn out true, though of course the intuition behind disquotation which is closely linked to them is central to the construction of the dependence function. This release

from the thrall of the biconditionals allows us to make sensible responses to Buridan-style cases. Suppose we have the sentences a and b

a: Either a or b is false. (6.36)
b: God exists.

In the absence of direct information about the truth of b, the standard argument goes: Suppose b is true, then if a is true, it is false, and if false, true, so b is false and a true. On the other hand, the dependence function for Ta is:

(6.37)



So Ta is only true in case God does not exist. This is surely the intuitively correct response: either God exists, or God does not exist. If the former, then a is paradoxical and we have no idea what truth-value to give it, and if the latter, it is staightforwardly true.

Thus the dependence function allows us to understand how ordinary truth assertions are uninfected by paradox: we do not determine them using Tarski biconditionals, but rather by using the function. We can now go on to say why the challenge given above is misplaced.

What I have offered is a theory which accounts for our basic, intuitive uses of 'true'. In doing so, I introduce a predicate 'd-true', and claim, with Kripke, that the sense in which 'The Liar sentence is not true' proper is a reflective, theoretical use. In this use, 'true' actually corresponds to 'd-true', rather than the basic 'true'. Burge claims, in a footnote of "Semantical Paradox", p. 174, that this is unsatifying and that there is "no intuitive ... basis for the claim that the natural language 'true' changes its sense or logic as a result of reflection."

I want to argue that in this case, it is quite clear that the use of 'true' is ambiguous: in one sense 'The Liar sentence is not true' is improper, and in the other it is proper. Whether this ambiguity is a consequence of reflection is irrelevant: the theory being offered explains it as due to a natural language/metalanguage distinction, and claims that there are intuitions which correspond precisely to the basic language.

Thus the proper use of 'true' in this case is when it is taken to mean 'd-true', and the improper use is when it is understood as basic 'true'. For there is a perfectly clear intuition that it is as bad to say 'The Liar is not true' as it is to assert the Liar itself, or to say 'It is not true that the Liar is not true' and so on. Part of what we recognize here is that not all iterations of '...is true' represent semantic ascent, which confers the ability to make genuine semantic assessment. The most apt analogy here is treading water in an abyss: if we compare for instance $\underline{R}(\{\{\ulcorner\underline{T}\urcorner\ulcorner\underline{Ta}\urcorner\})$ with $\underline{R}(\{\{\ulcorner\underline{Ta}\urcorner\})$, they are indistinguishable below the zeroth level, and stretch away to infinity, so expecting $\ulcorner\underline{T}\urcorner\underline{Ta}$ to evaluate $\ulcorner\underline{Ta}$ seems absurd. Faced with this problem, we simply refrain from assenting to any such sentence. Similarly, we refrain from asserting either the truth or falsity of the Buridan sentence $\underline{a}$, because we have no evaluation open to us.

My proposal is in one sense programmatic. It should include evidence for the claim that 'd-true' can be included in the language along with 'true'. Clearly any such addition would itself be gappy, and hence necessitate a further metalanguage and so on. However, the theory would not need to have coincident gaps: 'd-true' can be fully defined where 'true' has gaps, and only have gaps itself for 'This sentence is not d-true' and so forth.

Though a more comprehensive theory would indeed include all these elements, nevertheless it is important to get the basic language

straight first, and worry about higher levels later. What is more, even without the higher levels, there is a well-defined set of intuitions which it explains in a satisfactory way: the fact that there are further phenomena to be explained should not be allowed to cover this up.

At this point I should explain that there is a problem with introducing quantifiers into the account of $\underline{R}(\{\underline{A}\})$. The natural approach would be to add to the definition of $C'(\underline{A},\underline{B})$ the condition:

$$\text{If } (\underline{x})\underline{Q}\in\underline{A} \text{ then } \underline{Q}(\underline{t}/\underline{x})\in\underline{B} \tag{6.38}$$

This would, of course, pose the usual problem about names and infinite domains, but even assuming that this problem was met in one or another of the customary ways, this condition does not invariably give the right results. Suppose the model $\underline{M}$ contains the following assignment to a predicate $\underline{F}$:

$$\underline{v}(\underline{F})=\{\underline{A},\ \underline{T}'\underline{A}',\ \underline{T}'\underline{T}'\underline{A}'',\ldots\} \tag{6.39}$$

Let us suppose, for convenience, that $\underline{D}=\{\underline{d},\underline{A},\underline{T}'\underline{A}',\ldots\}$ and that $\underline{v}(\underline{a})=\underline{d}$. Then we can take the dependence function for $(\underline{E}\underline{x})(\underline{F}\underline{x}\&\sim\underline{T}\underline{x})$ by stages. Eliminating $(\underline{E}\underline{x})$ we get

$$\{\underline{F}\underline{a}\&\sim\underline{T}\underline{a},\underline{F}'\underline{A}'\&\sim\underline{T}'\underline{A}',\ldots\} \tag{6.40}$$

Now distributing the conjunctions, we get an infinite collection of sets each containing just one of each of $\underline{F}\underline{a}$ and $\sim\underline{T}\underline{a}$, $\underline{F}'\underline{A}'$ and $\sim\underline{T}'\underline{A}'$ and so on. The only set of significant interest is

$$\{\underline{F}\underline{a},\sim\underline{T}'\underline{A}',\sim\underline{T}'\underline{T}'\underline{A}'',\ldots\} \tag{6.41}$$

This in turn depends on:

$$\{\underline{F}\underline{a},\sim\underline{A},\sim\underline{T}'\underline{A}',\sim\underline{T}'\underline{T}'\underline{A}'',\ldots\} \tag{6.42}$$

and this just depends on itself, repeating level after level. Now, if $\underline{A}$

happens to be true in the model, all $\underline{A}$, $\underline{T}'\underline{A}'$, $\underline{T}'\underline{T}'\underline{A}''$ and so on are true too, and so $(\underline{E}\underline{x})(\underline{F}\underline{x}\&\sim\underline{T}\underline{x})$ should be false. However, we cannot discern that from the dependence function, which just repeats, and gives us no option to build up truths and falsehoods level by level. The trouble seems to be that (6.41) does indeed depend on (6.42), but to get to the truth of $(\underline{E}\underline{x})(\underline{F}\underline{x}\&\sim\underline{T}\underline{x})$ we need to separate out the dependencies of each of $\underline{A}$, $\underline{T}'\underline{A}'$, $\underline{T}'\underline{T}'\underline{A}''$, and so on, and not lump them all into one set.

On the other hand, the rule is not completely mistaken because if we look at the dependence function for $(\underline{x})(\underline{F}\underline{x}\supset\underline{T}\underline{x})$ it gives just the right results:

$$\begin{matrix} \{\sim\underline{F}\underline{a},\underline{T}\underline{a}\} \\ \{\sim\underline{F}\underline{a},\underline{T}\underline{a}\} \\ \text{etc.} \end{matrix} \quad \begin{matrix} (\underline{x})(\underline{F}\underline{x}\supset\underline{T}\underline{x}) \\ \{\sim\underline{F}'\underline{A}',\underline{T}'\underline{A}'\} \\ \{\sim\underline{F}'\underline{A}',\ \underline{A}\} \end{matrix} \quad \begin{matrix} \\ \{\sim\underline{F}'\underline{T}'\underline{A}'',\underline{T}'\underline{T}'\underline{A}''\}\text{etc.} \\ \{\sim\underline{F}'\underline{T}'\underline{A}'',\ \underline{T}'\underline{A}'\}\ \text{etc.} \end{matrix} \tag{6.43}$$

Each member of $\underline{R}(\{\{(\underline{x})(\underline{F}\underline{x}\ \underline{T}\underline{x})\}\})$ will be reduced to $\{\sim\underline{F}'\underline{T}\ldots'\underline{T}'\underline{A}''\ ',\underline{A}\}$ after a finite number of levels, so the whole function would be d-closed, which is an appropriate evaluation. It seems that the general solution of this problem is to introduce finer divisions than the simple dependence sets, but the means of doing this remain obscure. It should be clear, however, that though the current theory cannot treat quantified sentences in a satisfactory way, it can nevertheless cover a wide variety of cases, including nearly all the important variants of the Liar. Even Kripke's important quantified examples have propositional analogues with extended conjunctions and disjunctions, so that the technical restriction does not render the dependence function devoid of significance.

Finally, I want to argue that the dependence function enables us to take a more unified view of our intuitions about 'true' than Kripke's construction suggested, and lessens the scepticism which various problems led to. Briefly, it seemed that various plausible intuitions about 'true' were represented by incompatible alternatives: Strong

Kleene gave some advantages, supervaluations gave others, minimal fixed points others again, and so on. Now there is still room for debate about what the best closure rule may be, but the dependence function shows that many of these alternatives are not as incompatible as appeared. If we take d-closure as the correct explication of our use of 'true', then we can see that, say, the Strong Kleene grounded sentences have a particularly interesting property, having a-closed dependence functions, but that is not incompatible with d-closure being the right account. Similarly, supervaluation schemes characterize another interesting group of sentences, and the maximizing intuition is also given more rein than in the Kripke construction.

As a result, d-closure of the dependence function can claim to represent the basic intuitions about 'true' in an illuminating and satisfying way.

6.4 Appendix: Proofs

Lemma 13: For all $\alpha>0$, $(\underline{A})(\underline{B})$(If $D(\underline{A},\underline{B})$, then $\underline{A}$ Int $S_{1,\alpha+1}$ iff $\underline{B}$ Int $S_{1,\alpha}$)

Proof: Suppose $D(\underline{A},\underline{B})$. Then for each $\underline{P}\in\underline{A}$, $\underline{P}$ is a literal of one of the following forms:

1. $\underline{Tt}$
   If $\underline{P}\in S_{1,\alpha+1}$, then by definition of $S_{1,\alpha}$ and the valuation rules, $\underline{v(t)}\in S_{1,\alpha}$, and conversely.
   So $\underline{P}\in S_{1,\alpha+1}$ iff $\underline{v(t)}\in S_{1,\alpha}$.
2. $\underline{\sim Tt}$
   If $\underline{P}\in S_{1,\alpha+1}$, $\underline{Tt}\in S_{2,\alpha+1}$, so $\underline{v(t)}\in S_{2,\alpha}$ and $\underline{neg(v(t))}\in S_{1,\alpha}$, and conversely.
   So $\underline{P}\in S_{1,\alpha+1}$ iff $\underline{neg(v(t))}\in S_{1,\alpha}$.
3. $\underline{Rt_1\ldots t_n}$ or $\underline{\sim Rt_1\ldots t_n}$, $\underline{R}\neq\underline{T}$
   $\underline{P}\in S_{1,\alpha+1}$ iff $\underline{P}\in S_{1,\alpha}$, since $\alpha>0$.

However, every member of $\underline{B}$ is either a) $\underline{v(t)}$ where $\underline{Tt}$ is in $\underline{A}$, b) $\underline{neg(v(t))}$, where $\underline{\sim Tt}$ is in $\underline{A}$, or c) is in $\underline{A}$.
Hence $\underline{A}$ Int $S_{1,\alpha+1}$ iff $\underline{B}$ Int $S_{1,\alpha}$.

Lemma 14: For all $\alpha>0$, $(\underline{A})$($\underline{A}$ Int $S_{1,\alpha}$ iff $(\underline{B})$(If $C(\underline{A},\underline{B})$ then $\underline{B}$ Int $S_{1,\alpha}$))

Proof:

1. Suppose $\underline{A}$ Int $S_{1,\alpha}$, and $C(\underline{A},\underline{B})$.
   Then there is a sequence $\langle\underline{A}_1,\ldots\underline{A}_n\rangle$ such that $\underline{A}=\underline{A}_1$, $\underline{B}=\underline{A}_n$, and $C'(\underline{A}_i,\underline{A}_{i+1})$, $1\leqslant i<n$.
   It suffices to show that if $\underline{A}_i$ Int $S_{1,\alpha}$ then $\underline{A}_{i+1}$ Int $S_{1,\alpha}$.
   Suppose $\underline{A}_i$ Int $S_{1,\alpha}$. Then for some $\underline{P}\in\underline{A}_i$, $\underline{P}\in S_{1,\alpha}$.
   Then either
   a. $\underline{P}$ is a literal.
      If $\underline{P}$ is a literal, $\underline{P}\in\underline{A}_{i+1}$, so $\underline{A}_{i+1}$ Int $S_{1,\alpha}$
   b. $\underline{P}$ is of the form $\underline{Q\&R}$.

Then either $\underline{Q}$ or $\underline{R}$ is in $\underline{A}_{i+1}$, and both are in $\underline{S}_{1,\alpha}$
So $\underline{A}_{i+1}$ Int $\underline{S}_{1,\alpha}$

c. $\underline{P}$ is of the form $\sim(\underline{Q}\&\underline{R})$
Then both $\sim\underline{Q}$ and $\sim\underline{R}$ are in $\underline{A}_{i+1}$,
and one or other is in $\underline{S}_{1,\alpha}$.
So $\underline{A}_{i+1}$ Int $\underline{S}_{1,\alpha}$

d. $\underline{P}$ is of the form $\sim\sim\underline{Q}$
Then $\underline{Q}$ is in $\underline{A}_{i+1}$ and $\underline{S}_{1,\alpha}$
So $\underline{A}_{i+1}$ Int $\underline{S}_{1,\alpha}$.

So $\underline{A}_{i+1}$ Int $\underline{S}_{1,\alpha}$, hence $\underline{B}$ Int $\underline{S}_{1,\alpha}$

2. Suppose $(\underline{B})$(If $C(\underline{A},\underline{B})$, then $\underline{B}$ Int $\underline{S}_{1,\alpha}$).
Suppose, for reductio, $\underline{A}\cap\underline{S}_{1,\alpha}=\Lambda$.
Define $\underline{A}'$, as follows

a. if $\underline{P}$ is a literal in $\underline{A}$, then $\underline{P}$ is in $\underline{A}'$

b. if $(\underline{Q}\&\underline{R})$ is in $\underline{A}$, then, if $\underline{Q}\in\underline{S}_{1,\alpha}$, $\underline{Q}$ is in $\underline{A}'$
otherwise $\underline{R}$ is in $\underline{A}'$

c. if $\sim(\underline{Q}\&\underline{R})$ is in $\underline{A}$, then both $\sim\underline{Q}$ and $\sim\underline{R}$ are in $\underline{A}'$

d. if $\sim\sim\underline{Q}$ is in $\underline{A}$, $\underline{Q}$ is in $\underline{A}'$

e. nothing else is in $\underline{A}'$

By inspection we can determine two facts.
First, $C'(\underline{A},\underline{A}')$.
Second, $\underline{A}'\cap\underline{S}_{1,\alpha}=\Lambda$.
i.e. if $\underline{A}\cap\underline{S}_{1,\alpha}=\Lambda$, then there is an $\underline{A}'$ such that
$C'(\underline{A},\underline{A}')$ and $\underline{A}'\cap\underline{S}_{1,\alpha}$.
Consequently there is a $\underline{B}$ such that $C(\underline{A},\underline{B})$ and $\underline{B}\cap\underline{S}_{1,\alpha}=\Lambda$.
But this contradicts the hypothesis.
So $\underline{A}$ Int $\underline{S}_{1,\alpha}$.

Lemma 15: For all m, if $\alpha+n-m>0$, $\underline{A}$ Int $\underline{S}_{1,\alpha+n}$ iff
$(\underline{B})$(If $\underline{B}\in\underline{R}(\{\underline{A}\},m)$, then $\underline{B}$ Int $\underline{S}_{1,\alpha+n-m}$)

Proof: By induction on m.

1. Base Clause.
   By Lemma 14
   $\underline{A}$ Int $\underline{S}_1, \alpha+n$ iff $(\underline{B})$(If $C(\underline{A},\underline{B})$, then $\underline{B}$ Int $\underline{S}_1, \alpha+n$)
   By definition of $\underline{R}$ and $\underline{C}$,
   $(\underline{B})(C(\underline{A},\underline{B})$ iff $\underline{B} \in \underline{R}(\{\underline{A}\},\overline{O}))$
   Hence
   $\underline{A}$ Int $\underline{S}_1, \alpha+n$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},O)$, then $\underline{B}$ Int $\underline{S}_1, \alpha+n$)

2. Induction Clause.
   Assume, for p=m-1,
   $\underline{A}$ Int $\underline{S}_1, \alpha+n$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},p)$, then $\underline{B}$ Int $\underline{S}_1, \alpha+n-p$)

   a. Suppose $\underline{A}$ Int $\underline{S}_1, \alpha+n$. Then
      $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},p)$ then $\underline{B}$ Int $\underline{S}_1, \alpha+n-p$)
      Suppose $\underline{B} \in \underline{R}(\{\underline{A}\},m)$
      Then, by definition of $\underline{R}$, there is a $\underline{C}$ and a $\underline{D}$ such that:
      $C(\underline{C},\underline{B}), D(\underline{D},\underline{C})$ and $\underline{D} \in \underline{R}(\{\underline{A}\},p)$.
      Hence $\underline{D}$ Int $\underline{S}_1, \alpha+n-p$,
      and by Lemmas 13 and 14
      $\underline{C}$ Int $\underline{S}_1, \alpha+n-m$ and $\underline{B}$ Int $\underline{S}_1, \alpha+n-m$
      Hence $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},m)$, then $\underline{B}$ Int $\underline{S}_1, \alpha+n-m$)

   b. Suppose $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ Int $\underline{S}_1, \alpha+n-m$)
      Let $\underline{B} \in \underline{R}(\{\underline{A}\},p)$
      Then there is a $\underline{C}$ such that $D(\underline{B},\underline{C})$
      But by definition of $\underline{R}$
      $(\underline{D})(C(\underline{C},\underline{D})$ iff $\underline{D} \in \underline{R}(\{\underline{A}\},m))$
      Hence $(\underline{D})$(If $C(\underline{C},\underline{D})$, then $\underline{D}$ Int $\underline{S}_1, \alpha+n-m$)
      By Lemma 14
      $\underline{C}$ Int $\underline{S}_1, \alpha+n-m$, and hence $\underline{B}$ Int $\underline{S}_1, \alpha+n-p$.
      So $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},p)$, then $\underline{B}$ Int $\underline{S}_1, \alpha+n-p$)
      Hence, by the Induction Hypothesis,
      $\underline{A}$ Int $\underline{S}_1, \alpha+n$.

   So $\underline{A}$ Int $\underline{S}_1, \alpha+n$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ Int $\underline{S}_1, \alpha+n-m$)

This completes the induction, and the proof.

Theorem 19: $\underline{P}$ is true in the Weak Kleene minimal fixed point iff $\underline{R}(\{\{\underline{P}\}\})$ is b-closed.

Proof: (Sketch only.) The proof parallels the proof of 17, except that everywhere that proof uses $\underline{A}$ Int $\underline{S}_1, \alpha$ or $\underline{B}$ Int $\underline{S}_1, \alpha$, this proof needs $\underline{A}$ Int $\underline{S}_1, \alpha$ and $\underline{A} \leq \underline{S}_1, \alpha \cup \underline{S}_2, \alpha$, and

similarly for $\underline{B}$. Inspection of the lemmas shows that the proof then goes through.

Theorem 20: $\underline{P}$ is true in the m-c supervaluation minimal fixed point iff $\underline{R}(\{\{\underline{P}\}\})$ is c-closed.

Proof: (Sketch only.) The proof again parallels that of Theorem 17, though this time the relevant property is that $\underline{A}$ intersects with every maximal, consistent assignment to $\underline{T}$ at level α. Call this property $\pi(\alpha)$. One case worth looking at in detail is the relation D and change in levels.

Suppose a set $\underline{A}=\{\underline{P}_1 \ldots \underline{P}_n\}$ has $\pi(\alpha)$. Then every admissible valuation makes one of $\underline{TP}_1 \ldots \underline{TP}_n$ true, and so the supervaluation makes $\underline{TP}_1 \vee \ldots \vee \underline{TP}_n$ true. Consequently the set $\{\underline{TP}_1 \ldots \underline{TP}_n\}$ has $\pi(\alpha+1)$. Suppose $\underline{P}_i$ is of the form $\underline{Ft}$. Then at any level greater than zero, the value of $\underline{TP}_i$ is fixed by the supervaluation: whether true or false, $\{\underline{TP}_1 \ldots \underline{P}_i \ldots \underline{TP}_n\}$ will have $\pi(\alpha+1)$. Suppose $\underline{P}_j$ is of the form $\sim\underline{Q}_j$: then, since any valuation in which $\underline{TP}_j$ is true will be one in which $\underline{TQ}_j$ is false, and $\sim\underline{TQ}_j$ true, $\{\underline{TP}_1 \ldots \underline{P}_i \ldots \sim\underline{TQ}_j \ldots \underline{TP}_n\}$ will also have $\pi(\alpha+1)$. Hence any set $\underline{B}$ which bears D to $\underline{A}$ will have $\pi(\alpha+1)$. To reverse the argument, similar considerations apply.

Consequently we can obtain: $\underline{A}$ has $\pi(\alpha)$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},n)$ then $\underline{B}$ has $\pi(\alpha-n)$). But $\{\underline{P}\}$ has $\pi(\alpha)$ just in case $\underline{P}$ gets t in the supervaluation at α, and $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\{\underline{P}\}\},\alpha-1)$ then $\underline{B}$ has $\pi(1)$) can only be true if every $\underline{B}$ contains either a sentence true in the model or both a sentence and its negation, since $\underline{B}$ contains only literals. Hence the theorem holds.

Lemma 21: If $D(\underline{A},\underline{B})$, then $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+1}(\underline{U})$ iff $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$, for all successor ordinals α.

Proof: Suppose $D(\underline{A},\underline{B})$, so $\underline{A}$ only contains literals.

1. Assume $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+1}(\underline{U})$

   If $\underline{P} \in \underline{A}$ and $\underline{P} \in \tau_{\underline{M}}^{\alpha+1}(\underline{U})$, there are three cases:

   a. $\underline{P}$ is of the form $\underline{Tt}$.

      Then $\underline{v}(\underline{t}) \in \tau_{\underline{M}}^{\alpha}(\underline{U})$ and $\underline{v}(\underline{t}) \in \underline{B}$.

      Hence $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

b. P is of the form $\sim Tt$.

Then $v(t) \notin \tau_M^{\alpha}(U)$,

and, since $\tau_M^{\alpha}(U)$ is complete,

$neg(v(t)) \in \tau_M^{\alpha}(U)$.

But $neg(v(t)) \in B$

Hence B Int $\tau_M^{\alpha}(U)$.

c. P is of the form $Rt_1 \ldots t_n$, or $\sim Rt_1 \ldots t_n$ $(R \neq T)$.

Then $P \in \tau_M^{\alpha}(U)$, since $\alpha \neq 0$.

But $P \in B$.

Hence B Int $\tau_M^{\alpha}(U)$.

2. Assume B Int $\tau_M^{\alpha}(U)$.

If $Q \in B$ and $Q \in \tau_M^{\alpha}(U)$, then there is a P in A such that one of the following holds:

a. P is Tt and Q is $v(t)$.

Then $P \in \tau_M^{\alpha+1}(U)$.

b. P is $\sim Tt$ and Q is $neg(v(t))$.

By consistency of $\tau_M^{\alpha}(U)$, $v(t) \notin \tau_M^{\alpha}(U)$

So $Tt \notin \tau_M^{\alpha+1}(U)$, and hence

$P \in \tau_M^{\alpha+1}(U)$

c. P and Q are either $Rt_1 \ldots t_n$ or $\sim Rt_1 \ldots t_n$ $(R \neq T)$.

Then $P \in \tau_M^{\alpha+1}(U)$.

In each case, we obtain A Int $\tau_M^{\alpha+1}(U)$.

Note: In this proof, neither of the second subcases holds generally if α is a limit ordinal, and if $\alpha=0$, neither of the third subcases holds.

Lemma 22: For all successor ordinals α, A Int $\tau_M^{\alpha}(U)$ iff (B)(If $C(A,B)$ then B Int $\tau_M^{\alpha}(U)$).

Proof:

1. Suppose $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$, and $C(\underline{A},\underline{B})$.
Then there is a sequence $\langle \underline{A}_1 \ldots \underline{A}_n \rangle$
such that $\underline{A}=\underline{A}_1, \underline{B}=\underline{A}_n$ and $C'(\underline{A}_i,\underline{A}_{i+1})$, for all $1 \leq i < n$.

Hence it suffices to show that if $\underline{A}_i$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$,

$\underline{A}_{i+1}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

Suppose $\underline{P} \in \underline{A}_i$ and $\underline{P} \in \tau_{\underline{M}}^{\alpha}(\underline{U})$. Then:

a. $\underline{P}$ is a literal.
Then $\underline{P} \in \underline{A}_{i+1}$.

Hence $\underline{A}_{i+1}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

b. $\underline{P}$ is $(\underline{Q}\&\underline{R})$.

Then $\underline{Q},\underline{R} \in \tau_{\underline{M}}^{\alpha}(\underline{U})$,
by consistency and completeness,
and either $\underline{Q}$ or $\underline{R}$ is in $\underline{A}_{i+1}$.

Hence $\underline{A}_{i+1}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

c. $\underline{P}$ is $\sim(\underline{Q}\&\underline{R})$.

Then either $\sim\underline{Q}$ or $\sim\underline{R}$ is in $\tau_{\underline{M}}^{\alpha}(\underline{U})$
by consistency and completeness.
Both $\sim\underline{Q}$ and $\sim\underline{R}$ are in $\underline{A}_{i+1}$.

Hence $\underline{A}_{i+1}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

d. $\underline{P}$ is $\sim\sim\underline{Q}$

Then $\underline{Q} \in \tau_{\underline{M}}^{\alpha}(\underline{U})$, by consistency,
and $\underline{Q} \in \underline{A}_{i+1}$.

Hence $\underline{A}_{i+1}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

Consequently $(\underline{B})$(If $C(\underline{A},\underline{B})$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$).

2. Suppose $(\underline{B})$(If $C(\underline{A},\underline{B})$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$).

Suppose, for reductio, $\underline{A} \cap \tau_{\underline{M}}^{\alpha}(\underline{U}) = \Lambda$.
Then there is a sequence $\langle \underline{A}_1 \ldots \underline{A}_n \rangle$ such that $\underline{A}_1 = \underline{A}$, $\underline{A}_n = \underline{B}$ and each member bears C' to its successor. For suppose that there is a sequence $\langle \underline{A}_1 \ldots \underline{A}_n \rangle$ such that $\underline{A}=\underline{A}_1, C'(\underline{A}_i,\underline{A}_{i+1})$ for all $1 < i < n$, and $\underline{A}_n \cap \tau_{\underline{M}}^{\alpha}(\underline{U}) = \Lambda$. I shall show that there is a sequence $\langle \underline{A}_1 \ldots \underline{A}_n, \underline{A}_{n+1} \rangle$ where $C'(\underline{A}_n,\underline{A}_{n+1})$ and $\underline{A}_{n+1} \cap \tau_{\underline{M}}^{\alpha}(\underline{U}) = \Lambda$.

Let $\underline{A}_{n+1}$ be defined as follows:

a. if $\underline{P}$ is a literal in $\underline{A}_n$, $\underline{P} \in \underline{A}_{n+1}$.

b. if $(\underline{P}\&\underline{Q}) \in \underline{A}_n$, then if $\underline{P} \notin \tau_{\underline{M}}^{\alpha}(\underline{U})$, $\underline{P} \in \underline{A}_{n+1}$, otherwise $\underline{Q} \in \underline{A}_{n+1}$.

c. if $\sim(\underline{P}\&\underline{Q}) \in \underline{A}_n$, then $\sim\underline{P}, \sim\underline{Q} \in \underline{A}_{n+1}$.

d. if $\sim\sim\underline{P} \in \underline{A}_n$, then $\underline{P} \in \underline{A}_{n+1}$.

By inspection, $C'(\underline{A}_n, \underline{A}_{n+1})$ and $\underline{A}_{n+1} \cap \tau_{\underline{M}}^{\alpha}(\underline{U}) = \Lambda$, so long as $\tau_{\underline{M}}^{\alpha}(\underline{U})$ is consistent.
Since $\underline{A} \cap \tau_{\underline{M}}^{\alpha}(\underline{U}) = \Lambda$, for some $\underline{B}$ such that $C(\underline{A},\underline{B})$, $\underline{B} \cap \tau_{\underline{M}}^{\alpha}(\underline{U}) = \Lambda$. But this contradicts the hypothesis.
Hence $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

Note: Again, this proof relies on α being a successor ordinal.

Theorem 23: If α is a non-successor ordinal, then for all $m<n$, $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+n}(\underline{U})$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$).

Proof: By induction on m.

1. Base Clause.
   For m=0, $\underline{B} \in \underline{R}(\{\underline{A}\},m)$ iff $C(\underline{A},\underline{B})$.
   Hence by Lemma 22, α set at $\alpha+n$,
   $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+n}(\underline{U})$ iff
   $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$).

2. Induction Clause.
   Hypothesis: for p=m-1
   $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+n}(\underline{U})$ iff $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},p)$
   then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-p}(\underline{U})$).

   a. Assume $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+n}(\underline{U})$. Then
      $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{A}\},p)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-p}(\underline{U})$).
      Suppose $\underline{C} \in \underline{R}(\{\underline{A}\},m)$.
      Then, for some $\underline{D},\underline{E}$
      $C(\underline{D},\underline{C})$ and $D(\underline{E},\underline{D})$.
      By definition of $\underline{R}$, $\underline{E} \in \underline{R}(\{\underline{A}\},p)$,

hence $\underline{E}$ Int$\tau_{\underline{M}}^{\alpha+n-p}(\underline{U})$>

So by Lemma 21, $\underline{D}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$.

So $\underline{C}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$

So, by universal generalization,

$(\underline{B})$(If $\underline{B}\in\underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$)

b. Assume $(\underline{B})$(If $\underline{B}\in\underline{R}(\{\underline{A}\},m)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$).

Suppose $\underline{C}\in\underline{R}(\{\underline{A}\},p)$.
Then, for some $\underline{D}$, $D(\underline{C},\underline{D})$.
By definition of $\underline{R}$,
$(\underline{E})$(If $C(\underline{D},\underline{E})$ then $\underline{E}\in\underline{R}(\{\underline{A}\},m)$

Hence $(\underline{E})$(If $C(\underline{D},\underline{E})$ then $\underline{E}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$).

By Lemma 22 $\underline{D}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$.

So by Lemma 21, $\underline{C}$ Int $\tau_{\underline{M}}^{\alpha+n-m}(\underline{U})$

Hence $(\underline{B})$(If $\underline{B}\in\underline{R}(\{\underline{A}\},p)$ then $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha+n-p}(\underline{U})$).

So $\underline{A}$ Int $\tau_{\underline{M}}^{\alpha+n}(\underline{U})$.

Lemma 24: If $\underline{R}(\{\{\underline{P}\}\})$ is c-closed, then for some m and any successor ordinal α, $(\underline{B})(\underline{U})$(If $\underline{B}\in\underline{R}(\{\{\underline{P}\}\},m)$, $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$).

Proof: Suppose $\underline{R}(\{\{\underline{P}\}\})$ is c-closed. Then at some level m, every member $\underline{B}$ contains either a sentence $\underline{Q}$ true in the model, or $\underline{Q}$ and its negation $\sim\underline{Q}$. Suppose the former: then $\underline{Q}$ is in $\tau_{\underline{M}}^{\alpha}(\underline{U})$ for any $\underline{U}$ and successor ordinal α. Suppose the latter. Since the sentences true at any level for any starting set must include either $\underline{Q}$ or $\sim\underline{Q}$, either $\underline{Q}\in\tau_{\underline{M}}^{\alpha}(\underline{U})$ or $\sim\underline{Q}\in\tau_{\underline{M}}^{\alpha}(\underline{U})$. Either way, $\underline{B}$ Int $\tau_{\underline{M}}^{\alpha}(\underline{U})$.

Lemma 28: For all n, $\underline{P}$ is locally stably true at limit ordinal α for Γ iff $\underline{conj}(\{\underline{P}\},n)$ is locally stably true at α for Γ.

Proof:

1. Suppose $\underline{P}$ is locally stably true at α in Γ.

By definition $(E\beta)(\gamma)$(If $\beta\leqslant\gamma<\alpha$ then $\underline{P}\in\tau_{\underline{M}}^{\gamma}(\Gamma)$).
By Theorem 23

$(E\beta)(\gamma)$(If $\beta \leqslant \gamma < \alpha$ then $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{P}\},n)$ then $\underline{B}$ $\underline{Int}$ $\tau_{\underline{M}}^{\gamma -n}(\Gamma)$).

But $(\underline{B})$(If $\underline{B} \in \underline{R}(\{\underline{P}\},n)$ then $\underline{B}$ $\underline{Int}$ $\tau_{\underline{M}}^{\gamma -n}(\Gamma)$

is just the condition for $\underline{conj}(\{\underline{P}\},n) \in \tau_{\underline{M}}^{\gamma -n}(\Gamma)$.

Hence $(E\beta)(\gamma)$(If $\beta \leqslant \gamma < \alpha$ then $\underline{conj}(\{\underline{P}\},n) \in \tau_{\underline{M}}^{\gamma -n}(\Gamma)$)

Consequently, $\underline{conj}(\{\underline{P}\},n)$ is locally stably true at α.

2. Suppose $\underline{conj}(\{\underline{P}\},n)$ is locally stably true at α.

Then $(E\beta)(\gamma)$(If $\beta \leqslant \gamma < \alpha$ then $\underline{conj}(\{\underline{P}\},n) \in \tau_{\underline{M}}^{\gamma}(\Gamma)$).

By reversing the preceding argument,

$\underline{P}$ is locally stably true at α.

Conclusion

In "The Semantic Paradoxes" Chihara made several claims. First, his diagnosis of the Liar was that it is caused by the truth principle [Tr], which represents the convention which gives the meaning of 'true'. As a corollary, he claimed that exponents of the "consistency" view could at best be offering an alternative meaning, perhaps with a view to replacing the current one. However, he further claimed that since nearly all the time our current principle of meaning works perfectly well, there is little point in replacing it: the only areas where any serious problems arise are in the study of semantics and formal theories, and the suggested changes just seem too cumbersome to adopt for such recherche studies.

I argued in response that the diagnosis was worse than the disease. Accepting the claim that in fact the Liar gives us little trouble, I pointed out that Chihara's explanation of what is involved would mean that the Liar gave us a mountain of trouble. No truth ascription would be preferred over any other: from [Tr] plus ordinary logic, plus the presence of a Liar, $\underline{T}$'$\underline{A}$', ~$\underline{T}$'$\underline{A}$', $\underline{T}$'~$\underline{A}$' and ~$\underline{T}$'~$\underline{A}$' can all be inferred, no matter what sentence $\underline{A}$ might be. This led to the suggestion that more was involved in using 'true' than the simple principle [Tr], even though [Tr] does seem to represent a large part of the meaning of 'true', and I suggested the analogy with Putnam's idea of a stereotype. This in turn led to the claim that the so-called "consistency" theories could be seen as playing another role, namely that of explaining our use of 'true' rather than its meaning.

The paramount constraint on such a theory would be that it accounted for our use of 'true': it should agree with our intuitions about all sorts of straightforward cases, and show how the Liar, in all its many forms, was different and failed to be assertable, yet did not infect other cases. In their theories of truth, however, Martin and van Fraassen adopted a different methodological starting point, one which might have been inspired by the solution of the pseudo-paradox of the Barber. In that solution, a crucial presupposition is shown to be false, and we no longer feel the need to say either that the Barber does, or does not, shave himself: there simply is no such barber. Thus Martin tries to show that the Liar sentence is not semantically correct, and van Fraassen that it presupposes a contradiction. In Martin's discussion, he even claimed that apparently innocent truth-ascriptions like '"Virtue is triangular" is not true' are also tainted, so that it becomes clear that his theory in fact involves a revision of our intuitions about use rather than just accounting for them. However, I argued that Martin's theory suffers from a double defect: as a theory of semantical correctness it contains implausible elements that seem to be justified only by the need to account for the Liar and as a pure theory of truth it contains other implausible elements which are attributable to the desire to explain semantical correctness. In van Fraassen's theory, these particular problems are less acute: the theory of presupposition, though not without its defects, is at least not distorted by the Liar, and the principles which govern the treatment of truth seem plausible once the Principle of Bivalence is abandoned. Nevertheless, the resultant theory remains unsatisfactory: the non-truth of sentences without truth-value is unassertable, and the treatment of the Truth-Teller, the card variant, and quantified examples are all doubtful.

Furthermore, adopting Kripke's point about the importance of

riskiness in truth-ascription, we realize that the programme adopted by Martin and van Fraassen is misguided. In general, paradoxical sentences do not have any obvious defects as sentences: they just happen to go wrong on certain occasions of use. Thus semantic properties like presupposition and sortal correctness turn out to be inappropriate tools for tackling the whole range of Liar-paradoxical sentences, however plausible they may look when applied to 'This sentence is false'. Consequently Kripke concentrated on how to define the extension of a truth-predicate in a model in which various sentences might fall in the scope of various others, thereby leading to paradox. The result is an inductive definition of a whole structure of fixed points, such that in each the truth-predicate T represents truth. This structure has many felicitous features: it allows an intuitively satisfying definition of the notion of grounding a sentence, it gives an invaluable distinction between merely ungrounded Truth-Tellers and truly paradoxical Liars, it has a very plausible story about learning to use 'true' which reflects the structure of the definition, and of course it handles riskiness perfectly. Unfortunately, it has two disadvantages, which interact in an interesting way. The first is that some highly plausible logical and semantic principles turn out not to be true, and are even, on some occasions, paradoxical, and the second is that it seems difficult to connect all the advantages of the theory with the claim that the extension of 'true' is given by, say, the minimal fixed point in the Strong Kleene construction. I shall return to this problem in a moment, but I first want to discuss Burge's theory.

Burge, too, presents a theory in which riskiness should be representable, and has the advantage of using a two-valued framework, in which a quasi-Tarskian theory of levels is constructed. Unfortunately, the theory has grave defects. We cannot tell exactly what levels are attached to what sentences; the handling of sentences requiring

transfinite indices is inadequate; and the claim that 'true' is indexical is insufficiently justified.

However, I adopt from Burge the insistence on the importance of a two-valued theory, and use straightforward classical motivations to construct the dependence function for a sentence. When we look at possible closure rules for such functions, we find that appropriate choices can model any Kripke fixed point, and that other choices closely approximate a-stability in the Gupta revision process. Using the d-closure rule we can see that any logical law will come out true, but that some Tarski biconditionals, that of the Liar in particular, come out false. Thus I claim for a-closure that it gives a better explanation of our use of 'true' than any fixed point.

Now one response to this might come from the sceptical attitude which Kripke's theory apparently encourages, which is to say that d-closure represents another aspect of our already fragmentary theory of truth, and is interesting enough, but is not to be taken any more seriously than any other proposal. I have admitted that there may be some limit to our intuitions about truth, beyond which certain competing theories would be indeterminate as to which was the theory of truth. What I want to deny is that that limit is at the level of the alternatives presented by the Kripke hierarchy, and part of the reason for this denial comes from the way the dependence function shows that the alternatives in fact have a greater unity than appears at first sight. Thus if we think that d-closure represents our intuitive notion of truth, we can agree that groundedness, for example, is an especially important ingredient in our notion, and point out that it characterizes a special subset of the d-closed dependence functions, rather than a significantly different intuition.

There is another reason beyond our mere intuition of their truth, however, for favouring an approach in which classical laws are given a

more prominent status than they are in the Kripke hierarchy, and this reason returns us to some of the constraints imposed by the paradoxes. The principal constraint that I have been exploring is that a theory of truth must allow us to make truth-ascriptions in the presence of a Liar sentence: that is, it must explain why the Liar does not create too many problems. There is a Charybdis to this Scylla, though, since the theory ought not to make the Liar itself completely unproblematic, otherwise we might doubt that there was any need for the theory!

In a thoroughgoing three-valued theory like Martin's, for instance, if his account were right, and Liars were just category mistakes, then satisfying this constraint would be a serious problem. We should never have been worried by them at all, any more that we are worried by 'Virtue is triangular'. The only way this could be handled would be by an account which explained how we had been misled into thinking that the Liar was not a category mistake at all, and were consequently worried by the contradictions which ensued.

Kripke's account is much less clearly three-valued, and indeed, as I suggested, has strong ties to two-valued theories, but nevertheless the same problem arises. Suppose we return to the person learning to use 'true': this person starts off with a basketful of sentences, and progressively distributes more and more into one or other of two further baskets. Eventually no more sentences can be redistributed and the process stops. Nothing about the process suggests that there would be any surprise about finding some sentences left at the bottom of the first basket: indeed, one might expect it. Of course, among those left it will be true that some, but not others, could be put in one of the other baskets right at the beginning, and the process would still work smoothly, but it still is not clear why the Liar should worry us: it just lacks a truth-value.

Only if we take two-valued assumptions seriously can we see why the

Liar creates problems. Starting with both the desire to preserve classical logic and the intuition that the truth of T'P' depends on that of P, I constructed the dependence function for sentences. The result not only validates logical laws like Fa∨~Fa, it also validates semantic laws like T'Fa'∨T'~Fa'. But the appeal of such laws is strongly rooted in the assumption that predicates are fully defined. It comes as a considerable shock to find that for some sentences truth is not defined, because their dependence functions are not d-closed, and neither are those of their negations.

The problem the Liar presents is that if we maintain the assumption of full definition, we find that when we try to make the Liar true or false, the customary contradiction arises. Thus when we actually try to judge the Liar we have to abandon the assumption, and are unable to come to any verdict. We do abandon the assumption and the infection is contained, but nevertheless the Liar gives us a genuine conflict of intuition between assumptions we base the dependence function on, and the results we obtain.

Furthermore, we discover that any attempt to give a total definition of the truth-predicate will result in unpleasant results elsewhere. Suppose, for instance, we "close-off" by declaring that sentences with d-closed dependence functions are true, and all others false. Then the disjunction of the Liar and its negation is true, but both disjuncts are false, and the Tarski biconditional for the Liar is false though the Liar and 'The Liar is true' will both be false. In other words, we not only discover that the dependence function does not give a total definition, we also discover that we cannot give a total definition.

My insistence on the importance of two-valued logic and semantic claims may seem misplaced when the resultant theory ends up not giving a total definition for 'true'. My point is the following: unless we take

the presupposition of two-valuedness seriously, we can neither explain our intuitions about the truth of various sentences, nor explain why the Liar is worrisome. Thus the theory of truth which uses the d-closed dependence function is not merely one of an interesting collection of alternatives which includes various fixed points: it is, rather, a genuine improvement, which includes the advantages of the fixed points, and offers some additional advantages of its own.

The comparison with the different theories which use various limit rules in the Gupta revision process is more open to debate. I have suggested that the d-closure rule and a-stability are as alike as makes no odds: it may be that the differences between them are indeed beyond the power of any intuitions to settle. This question is, at any rate, one for further research. What is clear is that the dependence function gives an invaluable basis for comparing numerous of both the Kripke and Gupta theories, and provides an alternative theory in its own right.

Bibliography

List of Works Consulted

Bar-Hillel, Yehoshua. "Do Natural Languages Contain Paradoxes?" Studium Generale, 19 (1966), 391-397.

Belnap, Nuel D. "Gupta's Rule of Revision Theory of Truth." Journal of Philosophical Logic, 11 (1982), 103-116.

Burge, Tyler. "Semantical Paradox." Journal of Philosophy, 76 (1979), 169-198.

Chihara, Charles. "A Diagnosis of the Liar and Other Semantical Vicious-Circle Paradoxes." The Bertrand Russell Memorial Volumes. Ed. George Roberts. London: Allen and Unwin, 1979. I, 52-80.

_____. "The Semantic Paradoxes: A Diagnostic Investigation." Philosophical Review, 88 (1979), 590-618.

Davis, Lawrence. "An Alternate Formulation of Kripke's Theory of Truth." Journal of Philosophical Logic, 8 (1979), 289-296.

Donnellan, Keith S. "Categories, Negation, and the Liar Paradox." The Paradox of the Liar. Ed. Robert L. Martin. New Haven: Yale Univ. Press, 1970. 113-120.

Fine, Kit. "Vagueness, Truth and Logic." Synthese, 30 (1975), 265-300.

Grice, H. Paul. "Logic and Conversation." Unpublished lectures.

Grover, Dorothy L. "Inheritors and Paradox." Journal of Philosophy, 74 (1977), 590-604.

Grover, Dorothy L., Joseph L. Camp, and Nuel D. Belnap. "A Prosentential Theory of Truth." Philosophical Studies, 27 (1975), 73-125.

Gupta, Anil. "Truth and Paradox." Journal of Philosophical Logic, 11 (1982), 1-60.

Gupta, Anil and Robert L. Martin. "A Fixed Point Theorem for the Weak Kleene Valuation Scheme." Journal of Philosophical Logic,

forthcoming.

Hazen, Allen. "Davis's Formulation of Kripke's Theory of Truth: A Correction." Journal of Philosophical Logic, 10 (1981), 309-311.

Herzberger, Hans G. "The Logical Consistency of Language." Language and Learning. Ed. J. A. Emig, J. T. Fleming and H. M. Popps. New York: Harcourt Brace, 1966. 250-263.

_____. "The Truth-Conditional Consistency of Natural Languages." Journal of Philosophy, 64 (1967), 29-35.

_____. "Paradoxes of Grounding in Semantics." Journal of Philosophy, 67 (1970), 145-167.

_____. "Truth and Modality in Semantically Closed Languages." The Paradox of the Liar. Ed. Robert L. Martin. New Haven: Yale Univ. Press, 1970. 25-46.

_____. "Canonical Superlanguages." Journal of Philosophical Logic, 4 (1975), 45-65.

_____. "Presuppositional Policies." Language in Focus. Ed. Asa Kasher. Dordrecht: Reidel, 1976. 139-164.

_____. "Supervaluations without Truth-Value Gaps." Unpublished. 1978.

_____. "Naive Semantics and the Liar Paradox." Journal of Philosophy, 79 (1982), 479-497.

_____. "Notes on Naive Semantics." Journal of Philosophical Logic, 11 (1982), 61-102.

Hughes, Patrick and George Brent. Vicious Circles and Infinity. London: Cape, 1975.

Kneale, William. "Propositions and Truth in Natural Languages." Mind, 81 (1972), 225-243.

Kripke, Saul A. "Outline of a Theory of Truth." Journal of Philosophy, 72 (1975), 690-716.

Lightbody, T. P. "An Examination of Two Recent Approaches to the Liar Paradox." Diss. Case Western Reserve University, 1977.

Mackie, John L. Truth, Probability and Paradox. Oxford: Oxford Univ. Press, 1973.

Martin, Robert L. "Towards a Solution to the Liar Paradox." Philosophical Review, 76 (1967), 279-311.

_____. "On Grelling's Paradox." Philosophical Review, 77 (1968), 321-331.

_____. The Paradox of the Liar. New Haven: Yale Univ. Press, 1970.

_____. "A Category Solution to the Liar." The Paradox of the Liar. Ed. Robert L. Martin. New Haven: Yale Univ. Press, 1970. 91-112.

_____. "Relative Truth and Semantic Categories." Journal of Philosophical Logic, 3 (1974), 149-153.

_____. "Sortal Ranges for Complex Predicates." Journal of Philosophical Logic, 3 (1974), 159-167.

Martin, Robert L. and Peter W. Woodruff. "On Representing 'True-in-L' in L." Language in Focus. Ed. Asa Kasher. Dordrecht: Reidel, 1976. 113-117.

Mendelson, Elliot. Introduction to Mathematical Logic. Princeton: Van Nostrand, 1964.

Parsons, Charles. "The Liar Paradox." Journal of Philosophical Logic, 3 (1974), 381-412.

Post, John F. "Shades of the Liar." Journal of Philosophical Logic, 2 (1973), 370-386.

_____. "Shades of Possibility." Journal of Philosophical Logic, 3 (1974), 154-158.

Prior, Arthur N. "On a Family of Paradoxes." Notre Dame Journal of Formal Logic, 2 (1961), 16-32.

_____. "Some Problems of Self-Reference in John Buridan." Proceedings of the British Academy, 48 (1962), 281-296.

Putnam, Hilary. Mind, Language and Reality. Cambridge: Cambridge Univ, Press, 1975.

Quine, Willard Van Orman. Set Theory and its Logic. Cambridge, Mass.: Harvard Univ. Press, 1963.

_____. The Ways of Paradox and Other Essays. New York: Random House, 1966.

Skyrms, Brian. "The Return of the Liar: Three-Valued Logic and the Concept fof Truth." American Philosophical Quarterly, 7 (1970), 153-161.

_____. "Notes on Quantification and Self-Reference." The Paradox of the Liar. Ed. Robert L. Martin. New Haven: Yale Univ. Press, 1970. 67-74.

Stalnaker, Robert. "Presuppositions." Journal of Philosophical Logic, 2 (1973), 447-547.

_____. "Pragmatic Presuppositions." Semantics and Philosophy. Ed. Milton K. Munitz and Peter K. Unger. New York: New York Univ. Press, 1974. 197-213.

Strawson, Peter F. "On Referring." Mind, 59 (1950), 320-344.

_____. Introduction to Logical Theory. London: Methuen, 1952.

Tarski, Alfred. "The Semantic Conception of Truth." Philosophy and Phenomenological Research, 4 (1944), 341-376. Rpt. Semantics and the Philosophy of Language. Ed. Leonard Linsky. Urbana: Univ. of Illinois Press, 1952. 13-47.

_____. "The Concept of Truth in Formalized Languages." Logic, Semantics, and Metamathematics. Oxford: Oxford Univ. Press, 1956. 156-287.

_____. "Truth and Proof." Scientific American, 220 (1969), 63-77.

Thomason, Richmond H. "A Semantic Theory of Sortal Incorrectness." Journal of Philosophical Logic, 1 (1972), 209-258.

_____. "Paradoxes of Intentionality?" Unpublished. 1982.

van Fraassen, Bas C. "Singular Terms, Truth-Value Gaps, and Free Logic." Journal of Philosophy, 63 (1966), 481-495.

_____. "Presupposition, Implication, and Self-Reference." Journal of Philosophy, 65 (1968), 136-152.

_____. "Inference and Self-Reference." Synthese, 21 (1970), 425-438.

_____. "Truth and Paradoxical Consequences." The Paradox of the Liar. Ed. Robert L. Martin. New Haven: Yale Univ. Press, 1970. 13-23.

_____. "Rejoinder: On a Kantian Conception of Language." The Paradox of the Liar. Ed. Robert L. Martin. New Haven: Yale Univ. Press, 1970. 59-66.

_____. "Presuppositions, Supervaluations and Free Logic." The Logical Way of Doing Things. Ed. Karel Lambert. Dordrecht: Reidel, 1976.

Wallace, John. "On the Frame of Reference." Semantics of Natural Language. Ed. Donald Davidson and Gilbert Harman. Dordrecht: Reidel, 1972. 219-252.

Yablo, Steve. "Grounding, Dependence, and Paradox." Journal of Philosophical Logic, 11 (1982), 117-138.