THE KNOWER'S PARADOX AND REPRESENTATIONAL THEORIES OF ATTITUDES

Nicholas M. Asher
Dept. of Philosophy
Center for Cognitive Science
University of Texas
Austin, TX 78712

Johan A. W. Kamp
Dept. of Philosophy
Center for Cognitive Science
University of Texas
Austin, TX 78712

ABSTRACT

This paper is about a well-known problem concerning the treatment of propositional attitudes. Results obtained by Kaplan and Montague in the early sixties imply that certain propositional attitude theories are threatened with inconsistency. How large the variety of such theories really is has been stressed by Thomason (1980). It includes all those attitude accounts that are often referred to as "representational." We show that many non-representational theories avoid those paradoxes only so long as they refrain from incorporating certain further notions which seem as worthy of formalization as those they contain. In view of these artificial limitations that must be imposed to keep the paradoxes out, such non-representational theories offer no genuine advantage over representational alternatives. The Kaplan-Montague results therefore require a different response than has often been thought appropriate. Rather than taking refuge in a non-representational theory one should adapt the representational approach in such a way that the threat of inconsistency disappears. The paper ends with a sketch of how this might be accomplished.

INTRODUCTION

This paper is about a well-known problem concerning the treatment of propositional attitudes. Results obtained by Kaplan and Montague in the early sixties imply that certain propositional attitude theories are threatened with inconsistency. Thomason (1980) has stressed just how large the variety of such theories really is. It includes all those attitude accounts that are often referred to as "representational." The reason these results are so problematic is that representational theories of attitudes seem to enjoy some well-known advantages over non-representational accounts. We ourselves are committed to a particular type of representational theory, and so the Kaplan-Montague-Thomason results have a special urgency for us. We are unwilling to give up the advantages of a representational account, even though that might seem to be the obvious solution to the problem. We will instead make a case for a different response. That case involves two separate considerations. First, we will argue that most non-representational theories themselves do not embody a really satisfactory response to the paradoxes, since they avoid these paradoxes only so long as they refrain from incorporating certain further notions which seem as worthy of formalization as those they contain. Thus, non-representational theories offer no genuine advantage over representational alternatives. Second, we will indicate one way in which a representational account might be altered so that the threat of inconsistency is removed. We will end by showing how this alteration affects the logic of the attitudes and how it permits the study of self-referential and even paradoxical attitudes--yielding an additional advantage for our approach.

THE KNOWER'S PARADOX AND NON REPRESENTATIONAL THEORIES

We begin by reviewing how and why the problems that Kaplan and Montague brought to light arise. The difficulty is most easily exposed -- and in fact it was first noted -- in relation to one type of representational theory, that which takes the objects of the attitudes to be sentences of its own language. That attitudes such as belief and knowledge are to be analyzed as predicates of sentences is a view which has been put forward by a number of eminent philosophers, among them Carnap and Quine. In 1963 Montague reached the surprising conclusion that this proposal has paradoxical consequences.

One of the results Montague obtained was the following. Let T be an extension of Q^{∂} (Robinson arithmetic relativized to the formula ∂ whose only free variable is u), and for any sentence ψ let $\langle\psi\rangle$ be the numeral denoting the goedel number of ψ. Suppose that for any sentences ψ, φ and some one place predicate of expressions K: (K1) $\vdash_T K(\langle\psi\rangle) \rightarrow \psi$, (K2) if ψ is a theorem of logic then $\vdash_T K(\langle\psi\rangle)$, (K3) $\vdash_T (K(\langle\psi\rangle) \,\&\, K(\langle\psi \rightarrow \varphi\rangle)) \rightarrow K(\langle\varphi\rangle)$, (K4) $\vdash_T K(\langle K(\langle\psi\rangle)\rightarrow\psi\rangle)$. Then T is inconsistent. So intuitively valid principles of epistemic logic yield a contradiction when knowledge is represented as a predicate of sentences in a language with sufficient syntactic resources. This surprising result has come to be known as the Knower Paradox.

Thomason (1980) shows that commonly accepted principles of doxastic logic lead to similar paradoxes in theories which contain a sentence predicate representing belief.[1] He also argues that

[1]Let $\langle\varphi\rangle$ be the standard name of φ and let 'B' be a 1 place predicate. Intuitively 'B(x)' means that A (some fixed believer) believes that x. Then the following are commonly accepted principles of doxastic logic: (B1) $B(\langle\varphi \rightarrow \psi\rangle) \,\&\, B(\langle\varphi\rangle) \rightarrow B(\langle\psi\rangle)$, (B2) if φ is a theorem of logic then $B(\langle\varphi\rangle)$, (B3) $B(\langle\varphi\rangle) \rightarrow B(\langle B(\langle\varphi\rangle)\rangle)$, (B4) $B(\langle B(\langle\varphi\rangle) \rightarrow \varphi\rangle)$. If the theory, besides containing (B1) - (B4) also contains enough machinery for talking about its own syntax, then, Thomason shows, the belief of some intuitively harmless tautologies entails the belief of any sentence whatsoever.

Montague's results not only affect those representational theories which identify the objects of the attitudes with sentences of their own language, but that they pose a threat to all representational theories. He reasons as follows: Suppose that a certain attitude, say belief, is treated as a property of "proposition-like" objects -- let us call them 'propositions' -- which are built up from atomic constituents in much the same way that sentences are. Then, Thomason observes, with enough arithmetic at our disposal we can associate in the familiar way a "goedel number" with each such object and we can mimic the relevant structural properties of and relations between such objects by explicitly defined arithmetical predicates of their goedel numbers. This goedelization of propositions can then be exploited, he argues, to derive a contradiction in much the same way as it was obtained by Montague.

Thomason stresses the importance of what he refers to as the "recursive" character of the representational objects -- by which we take him to mean the principle that propositions are built up by certain combinatorial principles from basic constituents. From one perspective, the perspective of the believer reflecting upon the nature of his beliefs, this emphasis seems appropriate. Suppose that a person's beliefs involve representations that he himself sees as built up recursively from constituents in much the same way that sentences are. Further, suppose that he has some means for thinking about the constituent structure of representations in a sufficiently systematic and detailed way. Suppose finally that the inferences he is prepared to acknowledge as valid (and which he is consequently prepared to use in forming new beliefs from beliefs he already has) include the schemata (B1)-(B4) as well as those of classical logic. Then he will be able to go from any apparently harmless belief to an explicitly contradictory one by faultlessly reasoning in a way that parallels the Montague-Thomason argument. At this point, such a person should feel perplexed-- no less so, in fact, than the philosopher who sets out with the idea that belief must be analyzable as a predicate of sentences and that (B1)-(B4) are valid principles for such a predicate but who then, perhaps by reading Thomason, discovers to his surprise that things just cannot be that way.

If we focus on ascriptions of belief to sentient beings from an external prespective, however, the relevance of the recursive character of representations is less obvious. In fact the implications of Montague's original results are even more damaging than Thomason's argument suggests. To derive a paradox along the lines Montague and Kaplan discovered, it suffices to exploit a belief or knowledge predicate of propositions (whatever one takes propositions to be) to define a related predicate, satisfying the same familiar epistemic or doxastic principles, on goedel numbers of sentences. There are a number of different situations in which this is possible. Relevant factors are: (i) what machinery the theory contains for talking about the structure of its propositions and what assumptions about propositional structure it makes; (ii) the precise form in which the theory expresses the problematic epistemic or doxastic principles; and (iii) whether the theory has the means of formally representing the expression relation between sentences and propositions (i.e. the relation which holds between a proposition p and a sentence φ expressing p). Only in rather special cases have we been able to verify that something like Thomason's 'recursiveness' assumption is essential to the argument. On the other hand, there are many situations in which the knower paradox causes trouble independently of any assumption about the recursive or compositional structure of the attitudinal objects.

This last point is related to an observation we wish to make about a familiar non-representational framework for the analysis of propositional attitudes. This is the framework provided by Montague's system of Intensional Logic, or IL, in which propositions are treated as sets of possible worlds. IL is known to be immune against the epistemic and doxastic paradoxes, and for that very reason has been thought preferable as a framework for attitude analysis. But suppose IL is enriched with enough arithmetic to permit goedelization (e.g. we add the axioms of Q to the valid sentences of the theory). Let H be some particular goedelization relation -- i.e. n stands in the relation H to the sentence ψ if n

is the goedel number, according to some chosen goedelization scheme, of ψ. This relation determines a second relation G between numbers and propositions, which holds between n and p if n is the goedel number of a sentence which expresses p. Semantically this relation is completely defined; i.e., its extension is fully determined in each of the models of this extended system of IL. It might therefore seem harmless to add to the given system a binary predicate to represent this relation; and to adopt as new axioms such intuitively valid sentences as a) $G(\underline{n},\ulcorner\psi\urcorner)$, where $\underline{n}$ is the n-th numeral and n the goedel number of ψ, b) $(\forall u)\,(Sen(u) \rightarrow (\exists!p)\,G(u,p))$, where 'Sen' is the arithmetical predicate which is satisfied by just those numbers which are goedel numbers of sentences, and c) $\forall p(G(\underline{n},p) \rightarrow (\check{}p \leftrightarrow \psi))$, where $\underline{n}$ and ψ are as under a). However, this addition renders the system inconsistent; for we can now define a 'truth' predicate T of goedel numbers $(T(u) \equiv (\exists p)\,(G(u,p) \,\&\, \check{}p))$ for which we can easily show that $T(\underline{n}) \leftrightarrow \psi$ is valid whenever n is the goedel number of ψ. The inconsistency then follows in the usual way.

The fact that certain semantically well-defined relations cannot be incorporated into IL shows it to be unsuitable as a general framework for philosophical analysis. Thus, it is unsuitable, in particular, as a framework for an analysis of the attitudes. This conclusion should be especially disturbing to those who favor such a theory, as the advantages it is supposed to have depend crucially on the artificial limitations to which the expressive power of IL is subject.

The impossibility of representing in IL the expression relation between propositions and the goedel numbers of the sentences expressing them has, we saw, nothing to do with the presence or absence of attitudinal predicates but arises independently, because the language contains the sentence forming operator $\check{}$. But in certain weaker systems which lack this operator, it may be precisely the presence of an attitudinal predicate, together with the familiar epistemic or doxastic principles that govern its behavior, which prevents the addition of the predicate G. In these cases G could be used to define a corresponding attitudinal predicate of numbers $(K'(n) \equiv (\exists p)\,(G(n,p) \,\&\, K(p))$; and under suitable conditions one could show that the principles governing K also hold for K'. The contradiction then follows as in Montague (1963) or Thomason (1980).

We have argued that the emphasis on the recursive character of representational structure is appropriate when we consider the attitudinal paradoxes from the perspective of somebody who reasons about his own knowledge or beliefs. But we have also emphasized that, when we focus on certain formal theories of attitude attribution, representational structure need not be very significant, since the paradoxical results will ensue in any case as long as the theory has the means of relating the syntactic structure of its own sentences to the attitudinal objects it posits. There are, however, some theories of attitude attribution in which representational structure is a crucial ingredient in the derivation of the paradox. Consider for instance a theory containing predicates expressing structural properties and relations of the representations it posits as attitudinal objects. Thus, for instance, it might contain a 3-place predicate which holds between representations p, q and r iff r has the structure of a conjunction of p and q. Suppose that the theory contains enough set theory to guarantee the existence of arbitrary finite sequences of whatever objects are included in its universe of discourse, as well as the mathematics needed for arithmetization of syntax. Suppose further that the theory states that each representation is built up recursively from atomic constituents (i.e. that for each representation there is a sequence of representations which gives a decomposition of the representation into its constituents). Then it will be possible to give an explicit definition of the relation J which holds between a number n and a representation r iff n is the goedel number of r. In other words there will be a definable predicate $J(x, y)$ which, in any intended model for the theory, will be satisfied by n and r iff n is the goedel number of r. Using J we can explicitly define K' from K: $(\forall x)\,(K'(x) \leftrightarrow (\exists y)\,(J(x, y) \,\&\, K(y)))$. Again, if the theory already contained the principles (K1) - (K4) it will also contain the corresponding sentences with K' instead of K (this is not entirely trivial,

and in fact it is not even quite correctly stated; but the claim can be substantiated). So inconsistency arises in such a theory in any case.

It should be stressed that a theory in which representational structure is able to cause this kind of havoc must contain a non trivial amount of additional machinery. It is not obvious that every representational theory should come so heavily equipped.

This is not to say of course that the "representationalists" that Thomason takes to task in his article should not be criticized. Probably most of those who have advocated a representational theory of propositional attitudes have been unaware that the seemingly innocuous machinery needed to derive the paradoxes causes the troubles that it does. Some might still happily accept this machinery as a useful component or addition to their representational views. Nevertheless, the possibility of representational theorires that are weak enough to escape the paradox should not be dismissed out of hand. Precisely what scope remains for such theories is a matter that needs further investigation.

These remarks provide a far from complete picture of what the full spectrum of attitudinal theories that succumb to the knower paradox is like. But we hope they indicate that the Knower Paradox is not confined to those representational theories that Thomason seems to have had primarily in mind. It equally affects theories that do not attribute much structure to their attitudinal objects, but which are able to express a good deal about the connection between propositions and the sentences expressing them. Only the familiar systems of epistemic and doxastic logic, in which knowledge and belief are treated as sentential operators, and which do not treat propositions as objects, seem solidly protected from the problems we have touched upon. But those systems are so weak that they can hardly serve as adequate frameworks for analyzing the attitudes. For instance there does not seem to be any plausible way of representing within such a system statements like 'Bill knows everything that Sue knows'.

If the only drawback of intensional systems were that one cannot augment them with certain intuitively well-defined predicates such as G, this by itself might not be a good enough reason for abandoning the intensional framework. However, the framework is unsatisfactory for quite different reasons as well. First, sentences which identify the objects of the attitudes with sets of possible worlds cannot be prevented (in any natural way) from entailing that replacement of the complement of an attitude attribution by a sentence logically equivalent to it always preserves correctness of the attribution. This substitution principle goes counter to some of our most deeply rooted intuitions about such attitudes as belief. Also, the way in which attitudes have thus far been formalized within IL seems fundamentally unsuitable for attitudinal and pseudo-attitudinal notions whose objects are unequivocally sentences. ' x is justified in asserting the sentence s', for instance, is intuitively as clear a concept as ' x knows that s' and as much deserving of analysis. But it can hardly be interpreted as a relation between persons and propositions; its only plausible formalization is as a relation between individuals and sentences. As the notion intuitively satisfies the principles **(K1) - (K4)**, its representation as a sentence predicate will introduce the familiar difficulties, irrespective of how other attitudes are handled. At this point the advantage IL once seemed to have over representational theories-- that of treating knowledge and belief in such a way that the familiar principles of doxastic and epistemic logic can be vindicated -- would be very much reduced if not altogether lost.

REPRESENTATIONAL THEORIES

Representational theories do not suffer from some of the drawbacks of the intensional approach. They do not, for instance, imply that attitudes are invariant under logical equivalence. But they need to find some alternative answer to the problems Montague and Thomason have pointed out.

SEMANTICS

There is, as Montague's work made plain, an intimate connection between the Knower Paradox and the Liar Paradox -- in fact both are instances of some more general pattern. In the light of this it is worth noting that the two paradoxes have led to rather different responses. In principle it is possible to deal with the Liar Paradox by treating truth as a sentential operator, or, alternatively, as a property of sets of possible worlds. Such a treatment would be the natural analogue of the operator approach and the intensional approach to knowledge and belief, but it is completely trivial and consequently has had few if any serious proponents. Instead, the Liar Paradox has led to developments in quite different directions. Tarski proposed as a remedy an infinite hierarchy of increasingly powerful languages, each next one containing a truth predicate for its predecessor, while none contains such a predicate for itself. This move blocks the semantic paradoxes; but it has been felt to be unduly restrictive as it also eliminates the possibility of forming any sentences that speak about their own truth or falsity. A reluctance to throw out all sorts of "semantic" self-reference for the sake of consistency has led more recently to a very different approach, that of Kripke (1975), Gupta (1982), and Herzberger (1982) among others. This approach treats truth as a predicate of sentences and freely permits self reference; but its semantics is partial. The effect is that all the 'good' sentences, including the sound self-referential ones, end up with a definite truth value, while the Liar Sentence and other truly paradoxical sentences do not. This approach has proved fruitful and illuminating in connection with truth. We believe that it also holds considerable promise in relation to the attitudes.

A parallel treatment of the attitudes, however, is considerably more complicated. While truth is an extensional notion -- in the sense that the truth value of 'it is true that φ' is determined by the truth value of φ -- knowledge and belief are not. In fact we have argued these attitudes are not even intensional: φ and ψ may have the same intension while 'x believes that φ' and 'x believes that ψ' differ in truth value. This is one of the reasons for abandoning efforts to analyze the attitudes in strictly intensional terms.

Indeed the kind of analysis we prefer uses the framework of discourse representation theory. It would be free of the inadequacies of intensional semantics.[2] But to develop that analysis here would require far more explanation and justification than we have room for. We have therefore adopted a more traditional framework, familiar since the work of Hintikka (1962), in which knowledge and belief are characterized in terms of possible worlds. In this approach the knowledge (beliefs) of a person a at a world w is (are) represented by a set $W_{K,a}(w)$ ($W_{B,a}(w)$) of possible worlds, the set of all worlds compatible with the totality of a's knowledge (beliefs) in w. In the extant versions of this analysis, the sets $W_{K,a}(w)$ and $W_{B,a}(w)$ determine the truth values of knowledge and belief reports at w in a way familiar from modal logic; for instance, 'a knows that φ' is true at w iff φ is true in all worlds in $W_{K,a}(w)$.

Once the object language countenances self-referential reports, however, the formula for determining the truth values of attitude reports ceases to be self-evident. Just as there is a problem about the truth value of the liar sentence even when all the relevant facts are established, so there remains a problem about the truth values of some knowledge and belief reports, even after all facts, including those about the subject's knowledge and beliefs, have been determined. So, if we think of

---

[2]For a discussion of some of these issues, see Kamp (1985), Asher (forthcoming).

$W_{B,a}(w)$ as determining all the facts about a's beliefs at w, there will still be a problem about which self-referential belief reports about a are true at w. Our problem, then, will be to determine, for any possible world structure $\mathcal{W}$ with alternativeness relations for knowledge and belief, what in each world of $\mathcal{W}$ are the extensions for the knowledge and belief predicates, K and B.

As in similar analyses of the concept of truth that have inspired our work, we must expect the extensions of K and B to be essentially partial. In particular, the truly paradoxical attitude reports, such as the "knower sentence" which says of itself that its negation is known, should come out as neither definitely true nor definitely false. A judicious analysis, however, will succeed in assigning many other sentences, including some that contain elements of self-reference a truth value. There exist two quite different ways for arriving at such partial extensions: the first due to Kripke (1975), the second due to Herzberger (1982) and Gupta (1982). We will follow here the Herzberger-Gupta method. This method uses only classical, bivalent extensions but incorporates a process of repeated revision. Only the elements which from some point in the sequence of revisions onwards remain inside the predicate's extension count as definitely in the extension of the predicate. We have no absolutely compelling argument for our choice of the Herzberger Gupta strategy, but we think it has a number of advantages, some already discussed in Gupta and some others which we will detail below.

From these informal remarks, it ought to be fairly clear what our semantics for knowledge and belief will be like. So the formal definitions below will hold few surprises. We follow the familiar practice of representing the worlds compatible with all of a's knowledge in w by means of an "alternativeness" relation $R_{K,a}$; $w R_{K,a} w'$ iff w' is compatible with the totality of a's knowledge in w. We shall write '[wR]' to denote the set of all w' such that wRw'. Our object language will be a first order language L which contains two two-place predicates K and B. K(x,y) is to be read as 'x knows that y' and B(x,y) as 'x believes that y'. Our models for L are of the form $\langle W, D, \mathbb{I}, \{R_{K,a}\}_{a \in A}, \{R_{B,a}\}_{a \in A} \rangle$, where: (i) W is a set of worlds; (ii) D is a function that assigns to each $w \in W$ a non empty set; D_w is called the *universe* of w; (iii) $\mathfrak{M}$ has a *fixed universe*; i.e., for all $w, w' \in W$, $D_w = D_{w'}$; (iv) $A \subseteq D_w$; (v) $\mathbb{I}$ is a function which assigns to each non logical constant of L a classical extension at each world; thus if c is an individual constant of L, $\mathbb{I}c\mathbb{I}_w$ is a member of D_w, and if Q is an n-ary predicate $\mathbb{I}Q\mathbb{I}_w \subseteq D_w^n$; (vi) each individual constant c is *rigid* in $\mathfrak{M}$; i.e., for all $w, w' \in W$ in $\mathfrak{M}$, $\mathbb{I}c\mathbb{I}_w = \mathbb{I}c\mathbb{I}_{w'}$; (vii) $\mathfrak{M}$ is *sentence complete*; i.e., every sentence of L is included in the fixed universe of $\mathfrak{M}$.

Such models appear to provide two different means for determining, at any world w, the truth value of sentences of the forms $K(a, \varphi)$ and $B(a, \varphi)$. On the one hand, the model theory for predicate logic implies that $B(a, \varphi)$ is true in $\mathfrak{M}$ at w iff $\langle a, \varphi \rangle \in \mathbb{I}B\mathbb{I}_{\mathfrak{M},w}$. On the other hand, given that $B(a, \varphi)$ is intended as 'a believes that φ', the sentence should be true just in case φ is true at all $w' \in [wR_{B,a}]$. Ideally it should not matter which means we choose; the extension of B should correctly reflect the beliefs a has at w, as determined via the relations $R_{B,a}$. So it ought to be that

(1) For every sentence ψ and world w $\langle a, \psi \rangle \in \mathbb{I}B\mathbb{I}_{\mathfrak{M},w}$ iff ψ is true at all $w' \in [wR_{\mathfrak{M},B,a}]$.

We shall call L-models *doxastically coherent* iff (1) holds. Similarly, we shall say that a model $\mathfrak{M}$ is *epistemically coherent* iff (2) holds in M:

(2) For every sentence ψ and world w $\langle a, \psi \rangle \in \mathbb{I}K\mathbb{I}_{\mathfrak{M},w}$ iff ψ is true at all $w' \in [wR_{\mathfrak{M},K,a}]$.

In general, coherence is more than we can hope for. There are many models in which we find worlds w such that $\mathbb{I}B\mathbb{I}_w$ and $\mathbb{I}K\mathbb{I}_w$ conflict with what is true at the members of $[wR_{K,a}]$ and $[wR_{B,a}]$.

For instance, it can happen that for some sentence φ of L $\langle a,\varphi\rangle \in [\![B]\!]_w$ but that for some w' such that $wR_{B,a}w'$, φ is false at w'. What are we to say in this situation about the truth or falsity at w of the report 'a believes that φ'?

Before we discuss this question, let us simplify matters by considering the special case where the set A consists of a single agent a. We assume that in each model $\mathfrak{M}$ we consider, a is named by the constant $\underline{a}$ ($a = [\![\underline{a}]\!]$ in $\mathfrak{M}$), and we shall abbreviate '$K(\underline{a}, \varphi)$' and '$B(\underline{a},\varphi)$' to '$K(\varphi)$' and '$B(\varphi)$'.[3] Moreover, we will confine our attention in what follows to the predicate B and ignore K and the corresponding alternativeness relation $R_{K,a}$. We will denote the alternativeness relation $R_{B,a}$ simply by 'R'. All that we will say from here on about belief also applies to knowledge.

To return to the situation just described, it evidently does not involve a doxastically coherent model, and there is no one unequivocally right answer to the question we have asked. The answer we shall give stems in part from the motivation we gave for our models $\mathfrak{M}$. We already adopted the view that the alternativeness relations determine whatever facts there are about a's beliefs. So if there is a conflict between $[\![B]\!]$ and R, it is the former we should regard as misrepresenting the true state of affairs and thus in need of adjustment. The obvious formula for this is,

(3) $[\![B]\!]_w = \{\varphi: (\forall w' \in [wR])\ \varphi$ is true in $w'\}$.

But (3) only brings us back to the original question: if φ is of the form 'B(c)', what is it for φ to be true at w'? The seemingly sensible suggestion that the R alternatives of w' provide the answer leads to an infinite regress for precisely the sentences that we are most interested in here. Suppose for instance that the constant b denotes in $\mathfrak{M}$ the sentence $\neg B(b)$; b says that a does not believe it. Let us call such a sentence the "believer sentence." Should $b \in [\![B]\!]_w$? According to R, that will be so just in case b-- that is $\neg B(b)$-- holds in every $w' \in [wR]$. But whether b is true in w' reduces to the question of whether b fails in some $w'' \in [w'R]$, and so on. Evidently, this strategy for evaluation leads nowhere in such cases. The policy we will adopt instead, a direct analogue of that followed by Herzberger and Gupta, is to evaluate sentences at a world using the extension of the belief predicate.

The effect of this decision may be, of course, that the adjustment given by (3) will not be definitive. For instance, it may alter the extension of B, and with it the truth value of φ, at worlds $w' \in [wR]$; consequently, the extension of B at w may be out of sync once again. One might hope that further adjustments will lead eventually to coherence. But, as with truth, there are situations in which such harmony is never achieved. As we will see shortly, this is so in particular for truly paradoxical sentences such as the sentence b above. Another, weaker hope one might have is that those sentences that get "settled," i.e., which do not move in and out of the extension of B any more once a certain number of adjustments have occurred, get settled already after a finite number of adjustments. This would make our task easier, since we would not have to contemplate transfinite sequences of corrections. But, again as with the parallel theories of truth, there are ways of carrying the adjustment procedure past limit ordinals, and when those are added one finds that certain sentences only get settled at some transfinite stage.

Unfortunately, there are different ways of carrying the adjustment process past limit ordinals which lead to different continuations at subsequent ordinals but between which it is difficult to choose. We have adopted a clause that minimizes the positive extension of B at limit ordinals; it is in essence the intensional analogue of the clause adopted by Herzberger (1982). This clause will prevent all

[3] We will almost always ignore the distinction between objects and the constants denoting them. In the last sentence above this would have meant using the metalinguistic symbol 'a' to refer to the agent a and to the constant $\underline{a}$ of L that denotes a. No confusion should, we hope, arise from this practice.

paradoxical sentences, such as the believer sentence, from being stably true (i.e. true at a world w in $\mathfrak{M}$ for all γ in excess of some ordinal β).[4]

We thus arrive at the following definition. Given any model $\mathfrak{M}$ we define for each ordinal α the model $\mathfrak{M}^\alpha$, where $\mathfrak{M}^\alpha = \langle W_{\mathfrak{M}}, D_{\mathfrak{M}}, R_{\mathfrak{M}}, \mathbb{I}^\alpha \rangle$, $\mathbb{I}Q\mathbb{I}^\alpha = \mathbb{I}Q\mathbb{I}_{\mathfrak{M}}$ for all nonlogical constants Q other than B, and $\mathbb{I}B\mathbb{I}^\alpha$ is defined as follows:

i) $\mathbb{I}B\mathbb{I}^0_w = \mathbb{I}B\mathbb{I}_w$

ii) $\mathbb{I}B\mathbb{I}_w^{\alpha+1} = \{\varphi: \forall w'(wRw' \rightarrow \mathbb{I}\varphi\mathbb{I}_{\mathfrak{M}^\alpha, w'} = 1)\}$

iii) For limit ordinal α, $\mathbb{I}B\mathbb{I}_w^\alpha = \{\varphi: (\exists\beta<\alpha)(\forall\gamma)(\beta\leq\gamma<\alpha \rightarrow \varphi \in \mathbb{I}B\mathbb{I}_w^\gamma)\}$

The adjustment procedure defined above for $\mathbb{I}B\mathbb{I}_w$ reflects the idea that the (initial) extensions of B should be seen as secondary. From this perspective it is natural to consider, besides models for L, what we shall call *model structures*. Model structures are like models except that they do not assign extensions to the predicate B. Thus, a model structure $\mathcal{M}$ can be turned into a model by extending $\mathbb{I}\,\mathbb{I}_{\mathcal{M}}$ so that it interprets B as well. In general there is more than one way of turning a model structure into a model. We say that a model structure is *essentially incoherent* if every model that can be obtained from it is incoherent.

So far we have not given any of our reasons for adopting the revision method of Herzberger and Gupta rather than the substantially different strategy of Kripke (1975). Some of these have to do with formal advantages that we see in the Herzberger-Gupta approach, w;hich we do not yet have the technical tools to describe. But there is also a conceptual motive that underlies our choice, and this seems a good place to explain what it is.

Until now we have spoken of the problem how the truth values of self-referential sentences should be determined from what might be called an external perspective. We assumed that there was a determinate set of facts concerning the subject's beliefs and asked what, in the light of those facts, could be said about the truth values of certain self-referential belief to that subject. But besides this external point of view there is also an internal perspective on the issue, and it is from that perspective, we feel, that some of the puzzling features ofparadoxical self-reference are most clearly visible. The internal perspective is that of a subject who wonders whether he should regard a certain self-referential sentence as true or should regard himself as knowing or believing it. In reflecting upon such a question, the subject is easily led to engage in hypothetical reasoning of the form: 'suppose φ were true. Then that would mean...' Sometimes the outcome of such a deduction is a conclusion that contradicts the assumption from which it starts and this conclusion can then serve as the point of departure for a similar bout of reasoning that produces a new conclusion that contradicts the first and so on. In this way the subject finds himself driven from one answer to the question he posed himself to the opposite answer, and back again. The conclusion that he is likely to draw from all this-- that

[4]A perhaps more plausible alternative, given the lack of arguments pointing towards one of these possibilities for revision at limit ordinals, is to allow all of them. Each of these schemes is available at each limit ordinal. In this way we get, starting from a given model $\mathfrak{M}$, not a linear hierarchy of models $\mathfrak{M}^\alpha$ but a branching structure. The definitely believed sentences, according to such a model, would then be those which settle into the extension of B along every branch of the structure, and the definitely not believed sentences those that fall outside the extension of B along every branch. We will, however, refrain from working out that alternative here.

there really is no definite answer to be found-- derives from his awareness that every answer one might want to give would lead to its contradictory and thus be inherently unstable.

We see this process of rationally driven revision as a crucial feature of paradoxical self-reference. For this reason we prefer the method of Gupta and Herzberger, which we believe captures some of the essential features of this process. Kripke's method, in contrast, offers no explication of this aspect of self-referential sentences at all. At best it can be said to offer a plausible account of the way in which we settle the truth values of what Kripke himself has called the *grounded* sentences, sentences which do contain occurrences of the relevant predicate (for Kripke this is of course the truth predicate) but which are not self-referential that their evaluation never leads us back to the question of their own truth or falsity.

TYPES OF MODEL COHERENCE AND SELF-REFERENCE

While some models become coherent after one or more revisions, others do not. For the remainder of this paper, we will look into some of the questions relating to coherence. Which models can be turned into coherent ones? How many iterations are necessary before coherence is reached? Which sentences get settled and which do not? And finally what are the "logics" of knowledge and belief that coherent and incoherent models determine?

There are three distinct factors that determine whether a model $\mathfrak{M}$ becomes coherent after revision (i.e., whether $\mathfrak{M}^{\alpha}$ is coherent for some ordinal α): (i) the forms of self-reference that are realized in $\mathfrak{M}$, (ii) the constraints on the alternativeness relation $R_{\mathfrak{M}}$ (i.e., whether $R_{\mathfrak{M}}$ is transitive, etc.), (iii) the initial intension $[\![B]\!]^0$. The role of each of these factors will become clear as we look in detail at the effects of the revision in some particular cases.

But first a general remark about forms of self-reference. There are essentially two semantic mechanisms by means of which self-reference can arise, naming and quantification. Quantification always produces self-reference in our models. For every quantifier includes in its range the set of all sentences, and thus in particular the sentence in which it occurs. Consequently, there exist model structures that are essentially incoherent. For we can always select certain predicates of L to play the role of those syntactic predicates that are sufficient for the construction along Goedelian lines of sentences, which on the intended interpretation of their predicates can be paradoxical. In this way, we could, for instance, formulate a version φ of the believer sentence. If M is a model structure in which the syntactic predicates get their intended interpretations and in which some reasonable conditions are placed on the alternativeness relation, then in no model $\mathfrak{M}$ obtainable from M will φ ever settle at any world w-- i.e., for every ordinal α there are $\beta, \gamma > \alpha$ such that $\varphi \in [\![B]\!]_w^{\beta}$ iff $\varphi \notin [\![B]\!]_w^{\gamma}$. Thus, $\mathfrak{M}$ is essentially incoherent.

Quantificational self-reference is a subject about which we have little of importance to say in this paper. It is an extremely important topic but much too complex to handle here. We do want to look closely into the other variety of self-reference which arises through naming. But to study this other kind of self-reference, we must eliminate all possible interference from self-reference of the quantificational sort. There are several ways in which this can be done. Given our aims, it is immaterial which we choose. The one we have adopted is to restrict attention to those sentences of L in which all quantifiers are restricted by the formula $\neg S(u)$, and to those model structures and models in which for all w $[\![S]\!]_w$ is the set of L sentences and $[\![B]\!]_w \subseteq [\![S]\!]_w$.

As Kripke (1975) was the first to make fully explicit, self-reference may arise because a certain name designates a sentence in which it itself occurs. The model theoretic counterpart of this situation is that where a certain constant c denotes in a given model $\mathfrak{M}$ a sentence that contains c. We will refer to this type of reference as *designative* self-reference, and we will concentrate for most of the remainder of this section on models in which such self-reference arises. We will be largely preoccupied, moreover, with looking at one particular instance of designative self-reference, that of the believer sentence, in the form in which it was first given on p. 9 above. Although the results we will obtain for this sentence depend to some extent on special properties that it has, we hope they will give the reader some idea of what may be expected in connection wih other cases of designative self-reference.

Two very simple examples of designative self-reference in a model $\mathfrak{M}$ are exhibited by the following assignments to the constants b and c: (i) $[\![b]\!]_{\mathfrak{M}} = \neg B(b)$ (i.e., b denotes the believer sentence), (ii) $[\![c]\!]_{\mathfrak{M}} = B(c)$. To get an impression of how such sentences fare under iterated revision, let us consider models in which they constitute the only cases of designative self-reference. We shall first concentrate just on the believer sentence. Let $\mathcal{M}$ be a model structure such that (a) (i) holds, (b) for every individual constant $d \neq b$, $[\![d]\!]_{\mathcal{M}}$ is not a sentence of L. Given what b denotes in $\mathcal{M}$ we might expect that $\mathcal{M}$ cannot be turned into a coherent model. Propositions (1) - (3) show that this is generally, though not invariably, true.

Proposition 1: Suppose $R_{\mathcal{M}}$ is transitive and (ii) $(\exists w \in W_{\mathcal{M}})$ $([wR_{\mathcal{M}}] \neq \emptyset$ & $(\forall w' \in [wR_{\mathcal{M}}])$ $[w'R_{\mathcal{M}}] \neq \emptyset$. Then $\mathcal{M}$ is essentially incoherent.

The proof, though simple, is instructive in that one can see how b behaves under revision. Suppose $\mathfrak{M}$ is any model obtained from $\mathcal{M}$, and suppose that $\mathfrak{M}$ is coherent. Let w be a world such that $([wR_{\mathcal{M}}] \neq \emptyset$ & $(\forall w' \in [wR_{\mathcal{M}}])$ $[w'R_{\mathcal{M}}] \neq \emptyset$. There are two possibilities. a) $b \in [\![B]\!]_{\mathfrak{M},w}$. Then $\forall w' \in [wR]$ $\mathfrak{M} \Vdash_{w'} b$. So since b is the sentence $\neg B(b)$, $\forall w' \in [wR]$ $b \notin [\![B]\!]_{w'}$. $[wR] \neq \emptyset$. So let $w' \in [wR]$. Since $b \notin [\![B]\!]_{w'}$ there is a $w'' \in [w'R]$ such that it is not the case that $\mathfrak{M} \Vdash_{w''} b$. So $b \in [\![B]\!]_{w''}$. Since R is transitive, $w'' \in [wR]$, which contradicts that $b \in [\![B]\!]_{w}$. b) Now suppose that $b \notin [\![B]\!]_{w}$. Then there is a $w' \in [wR]$ such that it is not the case that $\mathfrak{M} \Vdash_{w'} b$. So $b \in [\![B]\!]_{w'}$. So $(\forall w'' \in [w'R])$ $\mathfrak{M} \Vdash_{w''} b$. $[w'R] \neq \emptyset$, so let $w'' \in [w'R]$. Then $\mathfrak{M} \Vdash_{w''} b$, and so $b \notin [\![B]\!]_{w''}$. So there is a $w''' \in [w''R]$ such that it is not the case that $\mathfrak{M} \Vdash_{w'''} b$. But since R is transitive, $w''' \in [w'R]$ and so $\mathfrak{M} \Vdash_{w'''} b$, which is a contradiction.

Proposition 2: Suppose $\mathcal{M}$ is essentially incoherent. Then condition (ii) of proposition 1 holds.

Since the transitivity of the alternativeness relation is standardly assumed in semantics for doxastic logic, there is a point to combining propositions 1 and 2:[5]

Proposition 3: Suppose R is transitive. Then $\mathcal{M}$ is essentially incoherent iff $\mathcal{M}$ satisfies the condition (ii) of proposition 1.

We need just a little more machinery to state some results concerning how the believer sentence and other paradoxical sentences drift in and out of the extensions of B. These results are interesting, at least in part because they do not depend on the initial extensions of B. Define the *b-profile at a world w in* $\mathfrak{M}$ to be the set of ordinals Γ such that for each $\alpha \in \Gamma$, $b \in [\![B]\!]^{\alpha}_{\mathfrak{M},w}$. Similarly, for any set of ordinals β, the *b-profile at w in* $\mathfrak{M}$ *on* β is the intersection of the b-profile at w in $\mathfrak{M}$ and β. We also need to define more precisely certain forms of designative self-reference. Let $<_{\mathfrak{M}}$ be the transitive closure of the relation that holds between two constants c and d iff c names in $\mathfrak{M}$ a sentence containing d. We will define $\mathfrak{M}$ to be *non-self-referential* iff $<_{\mathfrak{M}}$ is well founded. Otherwise, $\mathfrak{M}$ is *self-referential*. Moreover if C is a set of individual constants and $<_{\mathfrak{M}}$ on C is not well founded, then we say C is *self-referential in* $\mathfrak{M}$. A set of sentences S is *self-referential in* $\mathfrak{M}$ iff there is a set of individual constants C such that: (i) C is self-referential in $\mathfrak{M}$, (ii) each member of C denotes in $\mathfrak{M}$ a member of S, and (iii) C contains a designator for each sentence in S.

We can predict the character of some models for L without reference to either the initial intension of B or the alternativeness relation: we can extend the main proof in Gupta (1982) to show that every non-self-referential model $\mathfrak{M}$ is coherent. This is true regardless of the type of alternativeness relation or initial intension for B in $\mathfrak{M}$.

For self-reflexive models even of a very simple kind, however, we need to have at least some information concerning the character of the alternativeness relation to make some concrete predictions.

Proposition 4: Suppose that $\mathfrak{M}$ is a model such that: (i) $[\![b]\!]_{\mathfrak{M}} = \neg B(b)$; (ii) no constant other than b denotes a sentence in $\mathfrak{M}$; (iii) $R_{\mathfrak{M}}$ is transitive. Then for any $w \in W_{\mathfrak{M}}$ the b-profile at w in $\mathfrak{M}$ on ω is one of the following: (i) $\varnothing$, (ii) the even natural numbers, (iii) the odd natural numbers, and (iv) ω. Moreover, (iv) arises only if $[wR] = \varnothing$.

---

[5] We have not at this point been able to supply a condition on R which is both necessary and sufficient for the essential incoherence of $\mathcal{M}$. A necessary condition for essential incoherence that is relevant to cases where R is not transitive as well as to cases where it is, is that the inverse of $R_{\mathcal{M}}$ be not well-founded. But this condition is not sufficient as the following model structure in which R is indeed not-well-founded shows. Let $\mathcal{M}$ be a model structure with the properties (a) and (b) and which is such that $W_{\mathcal{M}} = \{w_0, w_1\}$ and $R_{\mathcal{M}}$ is the relation $\{\langle w_0, w_1\rangle, \langle w_1, w_0\rangle\}$. Let $\mathfrak{M}$ be a model obtained from $\mathcal{M}$ by adding $[\![B]\!]_{w_0}$ and $[\![B]\!]_{w_1}$ such that $b \in [\![B]\!]_{w_0}{}^{0}$ and $b \notin [\![B]\!]_{w_1}{}^{0}$ or $b \notin [\![B]\!]_{w_0}{}^{0}$ and $b \in [\![B]\!]_{w_1}{}^{0}$. Then $\mathfrak{M}$ will be coherent and thus $\mathcal{M}$ not essentially incoherent. When we generalize from this last example, we come to appreciate that a condition on R that is both necessary and sufficient for essential incoherence could not be very simple. To appreciate this observe that while the model structure we just defined is not essentially incoherent, a similar model structure with three worlds w_1, w_2, w_3 and an alternativeness relation consisting of the pairs $\langle w_1, w_2\rangle, \langle w_2, w_3\rangle, \langle w_3, w_1\rangle$ is essentially incoherent. This oddity generalizes to all odd and even loops: if $R_{\mathcal{M}}$ contains a loop with an odd number of elements then it is essentially incoherent; on the other hand, if $R_{\mathcal{M}}$ consists only of loops with even number of elements then $\mathcal{M}$ is not.

Proof: Assume first that R is serial-- i.e., $\forall w\ [wR] \neq \varnothing$. Suppose $w \notin W$. We distinguish the following cases.

(a) $(\forall w_1 \in [wR])(\exists w_2 \in [w_1R])\ b \in [\![B]\!]_{w_2}^{0}$.

Then $b \notin [\![B]\!]_{w'}^{1}$ for all $w' \in [wR] \cup \{w\}$. From this and the seriality of R, it follows that for $n \geq 2$ and $w' \in [wR]$ $b \in [\![B]\!]_{w'}^{n}$ iff n is even. So in particular the b-profile at w on ω is the set of even numbers.

(b) $(\exists w_1 \in [wR])(\forall w_2 \in [w_1R])\ b \notin [\![B]\!]_{w_2}^{0}$.

Let w_1 be such a member of $[wR]$. Then for each $w' \in [w_1R] \cup \{w_1\}$, $b \in [\![B]\!]_{w'}^{1}$. By an argument similar to that in (a), $b \in [\![B]\!]_{w'}^{n}$ iff n is odd. Since $w_1 \in [wR]$, this implies that if n is even, $b \notin [\![B]\!]_{w}^{n}$. To make further progress, we divide (b) into two subcases:

(b.1) $(\exists w_1 \in [wR])(\forall w_2 \in [w_1R])\ b \in [\![B]\!]_{w_2}^{0}$.

Let w_1 be as assumed. Then $(\forall w_2 \in [w_1R] \cup \{w_1\})\ (b \in [\![B]\!]_{w_2}^{n} \leftrightarrow$ n is even). Again because $w_1 \in [wR]$, $b \notin [\![B]\!]_{w}^{n}$, if n is odd. By what we have already seen under (b), this implies that the b-profile on ω at w in $\mathfrak{M}$ is $\varnothing$.

(b.2) $(\forall w_1 \in [wR])(\exists w_2 \in [w_1R])\ b \notin [\![B]\!]_{w_2}^{0}$.

This case requires yet another bifurcation.

(b.2.i) $(\exists w_1 \in [wR])(\forall w_2 \in [w_1R])\exists w_3 \in [w_2R])\ b \in [\![B]\!]_{w_3}^{0}$.

Then if w_1 is as assumed, we conclude as under (a) that for each $w' \in [w_1R] \cup \{w_1\}$ $(b \in [\![B]\!]_{w'}^{n}$ iff n is even). This holds in particular for w_1, and as $w_1 \in [wR]$, we conclude that b cannot belong to $[\![B]\!]_{w}^{n}$ when n is odd. So in this case the b-profile at w on ω is again $\varnothing$.

(b.2.ii) $(\forall w_1 \in [wR])(\exists w_2 \in [w_1R])(\forall w_3 \in [w_2R])\ b \notin [\![B]\!]_{w_3}^{0}$.

Then for each $w_1 \in [wR]$ there is a w_1' such that for all $w'' \in [w_1'R] \cup \{w_1'\}$ $(b \in [\![B]\!]_{w''}^{n}$ iff n is odd). Consequently, for every $w_1 \in [wR]$ $b \notin [\![B]\!]_{w_1}^{n}$ for even n, and so $b \in [\![B]\!]_{w}^{n}$ when n is odd. So the b-profile at w on ω consists of all the odd natural numbers.

If we drop the assumption that R is serial on W, we must also consider w such that $[wR] = \varnothing$ and w such that $(\exists w' \in [wR])\ w'R = \varnothing$. These give us $\varnothing$ and ω as b-profiles. This completes the proof.

It is not difficult to extend the result of proposition 4 so that it covers the full b-profile. Because of our clause for limit ordinals b (and any other paradoxical sentence) is never in $[\![B]\!]_{w}^{\lambda}$ for λ a limit ordinal and for all $w \in W$. So we immediately conclude that for any limit ordinal λ, the b-profile at w in $\mathfrak{M}$ on $(\lambda + \omega)\backslash\lambda$ is one of the sets: $\varnothing$, $(\lambda + \omega)\backslash\lambda$, $\{\lambda + 2n + 1 : n \in \omega\}$. Indeed, we have:

Proposition 5: Suppose that $\mathfrak{M}$ is as in proposition 4. Then for any limit ordinal λ and natural number n:

(i) if $[wR] = \varnothing$, then $b \in [\![B]\!]_{w}^{\lambda + n}$

(ii) if $(\exists w_1 \in [wR])\ [w_1R] = \varnothing$, then $b \notin [\![B]\!]_{w}^{\lambda + n}$

(iii) if $[wR] \neq \varnothing$ and $\neg(\exists w_1 \in [wR])\ [w_1R] = \varnothing$, then $b \in [\![B]\!]_{w}^{\lambda + n}$ iff n is even.[6]

[6] When R is not transitive, the b-profiles cannot be described in nearly such simple terms.

It is instructive to compare the behavior of the paradoxical believer sentence with the harmlessly self-referential c. Suppose that $\mathcal{M}$ is a model structure in which $[\![c]\!]_{\mathcal{M}} = B(c)$ and all other constants denote non-sentences in $\mathcal{M}$. Then if R is transitive, any model $\mathfrak{M}$ obtained from $\mathcal{M}$ will be coherent after one revision-- i.e., the coherent model will be $\mathfrak{M}^1$. If R is not transitive, there is no guarantee that coherence will be achieved that quickly; but it will be reached eventually. It should be noted that although c is not paradoxical, it is not a *grounded* sentence either in the sense of Kripke (1975). This implies that c's truth value cannot be determined without reference to the initial extensions of B. Indeed we find that in all but a few marginal cases, the model structure $\mathcal{M}$ does not determine the truth value of c: we can always turn $\mathcal{M}$ into two different models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ for each w, so that in $\mathfrak{M}_1$ c is true at w while in $\mathfrak{M}_2$ it is false at w.

The results obtained in this section so far are all quite easily established. It appears to be much more difficult to arrive at an equally detailed understanding of the behavior under revision of more complicated cases of designative self-reference. The only general result we are in a position to state as a theorem here uses rather strong constraints on the alternativeness relation. Before we can state this result, however, we must introduce a few more concepts. Suppose that $\mathfrak{M}$ is a model and that the set C of constants is self-referential in $\mathfrak{M}$. We say that C *is simply self-referential* in $\mathfrak{M}$ iff each $c \in C$ denotes in $\mathfrak{M}$ a boolean combination of sentences each one of which either (a) is of the form B(d) with $d \in C$ or (b) does not contain B. For any set of constants C of L, model $\mathfrak{M}$ and $w \in W_{\mathfrak{M}}$, the *C-characteristic of* w *in* $\mathfrak{M}$ is the function $f: C \rightarrow \{0,1\}$ such that for $c \in C$ $f(c) = 1$ iff $c \in [\![B]\!]_{\mathfrak{M}, w}$. By the C-*profile at* w *in* $\mathfrak{M}$ we understand the function defined on the class of all ordinals which maps each ordinal α onto the C-characteristic of w in $\mathfrak{M}^{\alpha}$. Similarly, if β is a set of ordinals then the *C-profile at* w *on* β *in* $\mathfrak{M}$ is the restriction to β of the C-profile at w in $\mathfrak{M}$.

> **Proposition 6**: Suppose that $\mathfrak{M}$ is a model, C a finite set of constants that is simply referential in $\mathfrak{M}$ and that $R_{\mathfrak{M}}$ is transitive, serial and euclidean (i.e., $(\forall w_1, w_2, w_3)((w_1Rw_2 \ \& \ w_1Rw_3) \rightarrow w_2Rw_3)$. Then there are natural numbers n and m such that for each $w \in W_{\mathfrak{M}}$ the C-profile at w on ω in $\mathfrak{M}$ is *cyclical after* n *with period* m; that is, if $r \geq n$ and $s = k.m + r$ then the C-characteristic at w in $\mathfrak{M}^s$ equals the C-characteristic at w in $\mathfrak{M}^r$.

It is straightfoward to extend this result to a similar one about full C-profiles.

The constraints we have imposed on R in proposition 6 are such as to make the proof almost trivial. But they are also quite strong; in particular the euclidean property can hardly be justified on the strength of our intuitions about belief. One might conjecture that the conclusion of proposition 6 also follows when the serial and the euclidean constraint are dropped. But a proof for this claim would be much more difficult.

We conclude this section with a result that is, like proposition 4, a special case of the conjecture we have juist made, and which concerns an instance of self-reference that has been discussed elsewhere in the literature on this topic (see e.g. Herzberger (1982)).

> **Proposition 7**: Suppose $\mathfrak{M}$ is a model such that (i) $[\![b]\!]_{\mathfrak{M}} = \neg B(c)$, (ii) $[\![c]\!]_{\mathfrak{M}} = B(b)$, (iii) all other constants of L do not denote sentences (iv) $R_{\mathfrak{M}}$ is transitive. Then for each $w \in W_{\mathfrak{M}}$ the {b,c}-profile at w on ω in $\mathfrak{M}$ is cyclical after 4 with period 4.

LOGIC OF THE ATTITUDES

The paradox Montague and Kaplan discovered was that languages capable of expressing enough about their own syntax cannot contain sentence predicates for concepts like belief that satisfy the intuitively valid and commonly accepted logical principles ascribed to them. The point of this paper has been to explore an as yet scarcely investigated way out of this difficulty, which sacrifices as little as possible from the totality of logical and semantical intuitions that their work shows to be incompatible. Of course, in a straightforward sense their results are definitive: any consistent, formal theory which treats belief as a predicate of sentences must give up something. In approaches of the sort we have advocated what has to go is part of the concept specific logical principles (such as, e.g., (B1) - (B4) for a predicate of belief) by which we would like to see these predicates governed, but which they cannot obey without exception. In this respect, the familiar and prima facie desirable doxastic logics are in the same position as is the Tarski T-schema in the work on truth on which we have built here. In particular, at least one of the axioms (B1) - (B4) will have to give up its absolute validity, if the semantic analysis we have offered is viable at all.[8]

It should not be surprising that (B4) is prominent among the principles that will be so affected. Propositions 10 and 11 below make this explicit. But it should be noted that it is only in the presence of truly paradoxical sentences that there is a need for giving up anything at all, as is evident from propositions 8 and 9.

Proposition 8: Suppose $\mathcal{M}$ is a designatively non-self-referential model structure (i.e. $<_{\mathcal{M}}$ is well founded). Then (a) there is an intension $[\![B]\!]$ such that the model $\mathfrak{M}$ obtained by adding $[\![B]\!]$ to $\mathcal{M}$ is coherent; (b) for any model $\mathfrak{M}$ obtained from $\mathcal{M}$ by adding intensions for B there is an α such that $\mathfrak{M}^{\alpha}$ is coherent.

Proposition 9: Suppose $\mathfrak{M}$ is coherent and that (i)R is transitive and (ii) reflexive on its range (i.e., $(\forall w\, w' \in W_{\mathfrak{M}})(wRw' \rightarrow w'Rw')$). Then all instances of (B1)-(B4) are true in $\mathfrak{M}$ at all worlds.

These two propositions suggest that it is legitimate to take the logic of belief sentences that are free of self-reference to contain all these axioms. In fact it may be appropriate to extend this claim to a somewhat larger domain, which also includes some self-referential sentences. As we have seen, there are self-referential sentences which admit of coherent models. For instance, every model structure $\mathcal{M}$, in which the only instance of self-reference is the sentence c, where $[\![c]\!]_{\mathcal{M}} = B(c)$, can be turned into a coherent model with the appropriate choice for $[\![B]\!]$. Moreover, if $R_{\mathcal{M}}$ is transitive and reflexive on its range, then <u>any</u> model $\mathfrak{M}$ obtained from $\mathcal{M}$ will coherent after one revision (i.e. $\mathfrak{M}^1$ is coherent).

What are we to make of the suggestions made in propositions 8 and 9 and the ensuing remarks? Evidently to turn them into arguments, we need an antecedent account of what would qualify as logically valid, given the kind of model theory we have developed. To get such an account we cannot simply extrapolate from the well understood and straightforward relationship between logic and semantics that is found in classical logic, exemplified by the familiar syntax and semanics of first order predicate logic. For like many other alternatives to the classical case, the model theory of the

---

[8]We again in this section continue as we have talking solely of belief. The parallel remarks to be made about knowledge are, we hope, an easy extrapolation.

previous section offers a number of different options for defining an associated logic, between which it is quite difficult if not impossible to make a well motivated choice. Suppose for instance we want to define what it is for a sentence to be logically valid. Presumably we want to say that the valid sentences are those that are invariably true. But how should we interpret this in the context of the model theory of this paper? What is it for a sentence to be "invariably true"? True at which worlds in which models? Here we face a number of different options, no one of which stands out unequivocally as the right one. Should we , for instance, include only coherent models, or should the truth values in non coherent models also count? If we go for the second option, should we consider all of them or only some distinguished subclass? These are only some of the questions that an account of validity must answer. Others, familiar from the literature on modal logic pertain to the constraints that should be imposed on the alternativeness relation. As we shall see below, there are further questions as well. Not until all these questions are settled will it be possible to assess the tentative claims about the logic of the non-self-referential and of the "harmlessly" self-referential sentences of L.

It is with similar caution that the reader should interpret the next two propositionswhich concern the logic of arbitrary designative self-refernce. When self-reference is not of the harmless variety exemplified by c, models as a rule start out incoherent and cannot be made coherent upon revision. Even such models, however, reach a certain kind of stability after enough revisions.

Proposition 10 : For each L model $\mathfrak{M}$ there is a least ordinal α_ϕ, such that: (i) for each $w \in W_{\mathfrak{M}}$ and each sentence φ $\varphi \in [\![B]\!]_{\mathfrak{M}, w}^{\alpha_\phi}$ iff $(\forall \beta \geq \alpha)\ \varphi \in [\![B]\!]_{\mathfrak{M}, w}^{\beta}$, (ii) after α_ϕ the revision process goes through a fixed cycle-- i.e. there is an ordinal γ such that for any $\beta_1, \beta_2 > \alpha_\phi$ if there is a $\delta < \gamma$ such that $\beta_1 = \gamma\pi_1 + \delta$ and $\beta_2 = \gamma\pi_2 + \delta$, then for all $w \in W_{\mathfrak{M}}$ $[\![B]\!]_w^{\beta_1} = [\![B]\!]_w^{\beta_2}$.

We shall call such an ordinal as α_ϕ a *minimization ordinal* for $\mathfrak{M}$ and $\mathfrak{M}^{\alpha_\phi}$ a *metastable model*. It follows from our definitions that if $\mathfrak{M}$ is coherent then the minimization ordinal for $\mathfrak{M}$ is 0 and the smallest γ satisfying (ii) in proposition 10 is 1.

One way of defining the logic of designative self-refernce would be to identify the valid sentences as those which come out true throughout all metastable models. On this assumption and given the appropriate choice of constraints for R, the schemata (B1) - (B3) will still come out valid, in the sense that all their instances are valid. But (B4) will now have false instances and so lose its validity as a schema.

Proposition 11 : Suppose $\mathfrak{M}^\alpha$ is a metastable model and suppose R is transitive in $\mathfrak{M}^\alpha$. Then every instance of (B1) -(B3) is true at every world in $W_{\mathfrak{M}^\alpha}$. Further, as long as R satisfies condition (ii) of Proposition 1, $\mathfrak{M}^\alpha$ will yield counterinstances to (B4) at some world.

Of course, this is not surprising; something had to give. However, within our framework, even (B4) retains a weaker kind of validity. To explain this, we should note that when $\mathfrak{M}^\alpha$ is metastable, the sentences that are not in $[\![B]\!]_{\mathfrak{M}, w}^{\alpha}$ fall into two natural classes-- those that remain outside $[\![B]\!]_{\mathfrak{M}, w}^{\beta}$ for all $\beta \geq \alpha$ and those that continue to move in and out of $[\![B]\!]_w$. The first might be naturally regarded as the *antiextension* of B at w-- the set of sentences that are definitely not believed at w. The second consists of sentences whose status as beliefs is forever in doubt. We can use the extension and antiextension of B in $\mathfrak{M}^\alpha$ to construct a partial model $\mathfrak{M}$. In any of the familiar valuation schemes for partial models-- the Kleene valuations or supervaluation schemes, certain sentences will not get a truth value. With reference to (B4), we now find that, on any of these valuation schemes, none of its instances come out false at any world of $\mathfrak{M}$. So if we were to identify

the valid sentences as those which never come out false in any of the partial models associated with metastable models, the schema (B4) gets reinstated as valid.

One must not forget, however, that this rehabilitation of (B4) is something of a sham. For example, the instance of (B4) which we get by replacing φ in it with the paradoxical believer sentence will lack a truth value at each partial model, and so it will never actually come out false. But it certainly will never come out true either. Thus, it would be unwise to rely upon such instances of the principle when engaging in doxastic reasoning.

This last observation leads us back into not only questions about choices between valuation schemes but also fundamental issues of logic and its relation to semantics. Since we have already said we cannot deal adequately with these questions here, better to stop now and save a fuller treatment of these issues for another occasion.

References

N. Asher: 1985, 'Belief in Discourse Representation Theory,' forthcoming in *Journal of Philosophical Logic*.

A. Gupta: 1982, 'Truth and Paradox,' *Journal of Philosophical Logic* 11, pp. 1-60.

H. Herzberger: 1982, 'Notes on Naive Semantics,' *Journal of Philosophical Logic* 11, pp. 61-102.

H. Herzberger: 1982, 'Naive Semantics and the Liar Paradox,' *Journal of Philosophy* 79, pp. 479-497.

H. Kamp: 1985, 'Context Thought and Communication,' *Proceedings of the Aristotelian Society*, 1984-85.

D. Kaplan & R. Montague: 1960, 'A Paradox Regained,' *Notre Dame Journal of Formal Logic* 1, pp. 79-90.

S. Kripke: 1975, 'Outline of a New Theory of Truth,' *Journal of Philosophy* 72, pp. 690-715.

R. Montague: 1963, 'Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability,' *Acta Philosophica Fennica* 16, pp. 153-167.

R. Thomason: 1980, 'A Note on Syntactical Treatments of Modality,' *Synthese* 44, pp. 391-395.