NICHOLAS ASHER AND HANS KAMP

SELF-REFERENCE, ATTITUDES AND PARADOX

I. MOTIVATIONS

I.1.

There is a sense, Gödel taught us, in which elementary arithmetic and other theories of comparable power can express and prove facts about their own syntax. Gödel showed also how that capacity can be exploited in the construction of certain "self-referential" sentences by a method of what might be called syntactic diagonalization. The example of self-reference which figures most prominently in Gödel's own paper is the sentence which appears to say of itself that it is not provable. But that this is only one out of an infinity of self-reflective statements is indicated by the following lemma[1] which we repeat here for general reference.

LEMMA 1. *Suppose T is an extension of Q^β* (Robinson's arithmetic relativized to β). *Let φ be a formula whose only free variable is v_0. Then there is a sentence ψ such that $\vdash_T \psi \leftrightarrow \varphi(\boldsymbol{\psi}/v_0)$, where, if n is the Gödel number of ψ, $\boldsymbol{\psi}$ is the n-th numeral.*

The Gödel sentence, which asserts its own unprovability, is obtained from Lemma 1 by taking φ to be the negation of the formula which expresses provability from the axioms of T. Another instance of the lemma gives us Tarski's theorem of the "undefinability of truth", which asserts that in extensions of Q^β and similar theories there can be no formula $\varphi(v_0)$ which "expresses truth", in the sense that all instances of the Tarski T-schema',

$$\psi \leftrightarrow \varphi(\boldsymbol{\psi}),$$

are provable in the theory.

Actually it is somewhat misleading to refer to Tarski's Theorem as the "undefinability" of truth, for the result is really stronger than that term suggests. It is not just that in a theory such as Q^β, in which no

Gennaro Chierchia, Barbara H. Partee, and Raymond Turner (eds.), Properties, Types and Meaning, I, 85—158.

predicate has been set aside specifically as a truth predicate, no formula can be found that satisfies all instances of the T-schema; the same argument which proves the non-existence of such a formula also establishes the perhaps even more surprising fact that a truth predicate cannot be *added* to such theories. If we extend the language of, e.g., Q^β with a new 1-place predicate T, and adopt all instances of the corresponding version of Tarski's schema,

$$\psi \leftrightarrow T(\boldsymbol{\psi}),$$

as axioms, then the resulting theory is inconsistent. Both this last observation and the undefinability theorem properly so-called are easy corollaries of Lemma 1. To obtain the first, take ψ to be the formula $\neg T(v_0)$. This yields a particular sentence φ such that both $\psi \leftrightarrow T(\boldsymbol{\psi})$ and $\psi \leftrightarrow \neg T(\boldsymbol{\psi})$ are theorems of the new theory, which shows the theory to be inconsistent. The undefinability of truth is proved by the same reductive argument.

The stronger version of Tarski's Theorem may be viewed as an *incompatibility* result: two theories, each of which appears to speak truly about its intended domain, nevertheless cannot be combined into a consistent whole. In the case covered by Lemma 1, the first theory is Q^β, while the second theory is given by the set of all instances of Tarski's Schema for the language of Q^β extended with the predicate T. As Montague discovered, there are many more results of this general type: Two intuitively true theories, the first comprehending, in one form or another, the theory of its own syntax, and the second embodying principles that seem to capture all or part of the "logic" of concepts such as necessity, knowledge or belief, turn out to be jointly inconsistent. Montague (1963) shows how such results can be proved, in much the same way as Tarski's theorem about truth, by judicious applications of Lemma 1. The majority of the axiom systems which turn out to be incompatible with theories like Q^β are strictly weaker than the set of instances of Tarski's Schema. Examples are the modal systems S_4 and S_5, and — of particular relevance to the present paper — certain doxastic and epistemic logics, each consisting of a number of intuitively valid principles concerning 'x knows that' or 'x believes that'.[2] Montague (1963) exhibits a quite weak epistemic system consisting of all instances of the schemata $(K1)$ $K(\boldsymbol{\varphi}) \rightarrow \varphi$; $(K2)$ $K(\boldsymbol{\varphi})$, if φ is an axiom of 1st order logic; $(K3)$ $K(\boldsymbol{\varphi \rightarrow \psi}) \rightarrow (K(\boldsymbol{\varphi}) \rightarrow K(\boldsymbol{\psi}))$; and $(K4)$ $K(\boldsymbol{K(\varphi)} \rightarrow \boldsymbol{\varphi})$. These principles all have a good deal of intuitive plausibility when K is interpreted as 'x knows that'. When the

predicate is taken to stand for belief rather than knowledge the first is no longer justified. However, Thomason (1980) shows that if $(K1)$ is replaced by the reasonable doxastic principle $(B1)$ $B(\varphi) \rightarrow B(\boldsymbol{B}(\varphi))$ while $(K2)$—$(K4)$ are retained (with, of course, B instead of K) then, the resulting system, although not formally inconsistent, is clearly unacceptable. We repeat this revised set explicitly for future reference:

$(B1)$ $B(\varphi) \rightarrow B(\boldsymbol{B}(\varphi))$;
$(B2)$ $B(\varphi)$, if φ is an axiom of first order logic;
$(B3)$ $B(\varphi \rightarrow \psi) \rightarrow (B(\varphi) \rightarrow B(\psi))$;
$(B4)$ $B(\boldsymbol{B}(\varphi) \rightarrow \varphi)$.[3]

I.2.

This paper is concerned with the epistemic and doxastic incompatibility theorems. Our active interest in the problems they raise was kindled by Thomason (1980). Thomason argues that the results of Montague (1963) apply not only to theories in which attitudinal concepts, such as knowledge and belief, are treated as predicates of sentences, but also to "representational" theories of the attitudes, which analyze these concepts as relations to, or operations on (mental) representations. Such representational treatments of the attitudes have found many advocates; and it is probably true that some of their proponents have not been sufficiently alert to the pitfalls of self-reference even after those had been so clearly exposed in Montague (1963) and its predecessor, Kaplan & Montague (1962). To such happy-go-lucky representationalists, Thomason (1980) is a stern warning of the obstacles that a precise elaboration of their proposals would encounter.

Thomason's argument is, at least on the face of it, straightforward. He reasons as follows: Suppose that a certain attitude, say belief, is treated as a property of "proposition-like" objects — let us call them 'representations' — which are built up from atomic constituents in much the way that sentences are. Then, with enough arithmetic at our disposal, we can associate a Gödel 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 Gödel numbers. This Gödelization of representations can then be exploited to derive a contradiction in ways familiar from the work of Gödel, Tarski and Montague.

We ourselves have been developing a theory of propositional atti-

tudes with strong representational implications,[4] and thus seem to have made ourselves vulnerable to Thomason's critique. As a matter of fact the general approach we have taken does not commit one irrevocably to the kind of representationalism that Thomason recognized as troublesome. However, we are persuaded that a comprehensive theory of the attitudes will have to take account of representational structure in such a way that it lays itself open to paradox. (We will give some of our reasons for this conviction presently.) It is our view therefore that Thomason's challenge should be met, not by eliminating the problematic forms of representationalism, but by developing a coherent framework in which these forms are possible.

This conclusion differs from the one that many philosophers have drawn from Montague's incompatibility theorems. They have often been interpreted as showing that notions such as knowledge, belief or necessity cannot be treated as predicates of sentences or similarly structured representational objects. This conclusion seemed reasonable because, already at the time when the Montague results became known, an alternative framework for studying the logic and semantics of modal and attitudinal notions had been established.

I.3.

This alternative framework is based on the semantical foundations of modal logic that were provided by Kripke and others in the late fifties and early sixties. Necessity is represented, following C. I. Lewis, as a one-place sentential operator $\Box$, and $\Box\varphi$ is analyzed as true in a possible world w iff

(*) φ is true in all worlds that are possible alternatives to w.

Hintikka (1962) exploited what is in essence the same conception in an analysis of knowledge and belief. Knowledge and belief are represented as sentential operators K and B, and the truth conditions for sentences '$K\varphi$' and '$B\varphi$' are given in the general form exemplified by (*).

The possible worlds approach has been immensely fruitful, not only because it has given us insight into the logic of the modalities proper, but also because it has offered a unified treatment of a large variety of notions belonging to intuitively distinct conceptual domains. Moreover, it has the apparent advantage of being immune against the paradoxes that Kaplan and Montague discovered. To be precise, let T be a theory

such that: (1) the language of T is a first order modal language, i.e. a language of first order predicate logic with an additional 1-place sentential operator $\Box$; (2) the axioms of T are (a) all instances of the schemata of S_5,[5] and (b) all necessitations of instances of the axioms and schemata of Q^{β};[6] then T is consistent. Moreover, T has Kripke models which incorporate a standard model of arithmetic at each of their worlds; that is, there are models $M = \langle W, R, D, [\![\ ]\!] \rangle$ such that at some world $w_0 \in W$ all sentences of T are true in M at w_0, and for each world $w \in W$, D_w (the universe at w) includes the set N of natural numbers, and succ, $+$, and $\cdot$ have at w their standard interpretations within N.[7]

An analogous treatment of knowledge and belief is equally immune to the paradoxes that cause trouble for the sentential and representational treatments. This follows from the claim just made about $\Box$, together with the fact that all the familiar (and plausible) epistemic and doxastic logics are subsystems of S_5. Since modal treatments are capable of validating the familiar epistemic and doxastic logics, whereas their sentential and representational competitors do not, the former seem preferable.

But they also have serious drawbacks. Perhaps the most obvious one is that, in the form in which we have just described the modal approach, it does not permit *quantification* over beliefs and similar entities. There are many things that we are quite ready to say in ordinary discourse, and which appear to have a well-defined meaning, but which can only be formalized if such quantification is available; for instance

(1) Bill believes everything that Joe believes.

(2) Some of the things Joe believes are not true.

and so on.

This, however, is a shortcoming that can be corrected without giving up the aforementioned advantage. It was one of the accomplishments of Montague's Intensional Logic[8] that it managed to introduce quantification over propositions in a theory which treats necessity, knowledge, and belief as predicates of propositions without thereby lapsing into the inconsistencies that, as he himself had noticed a few years earlier, threaten the sentential treatments of knowledge and necessity, and this without having to abandon any of the intuitively desirable principles of modal, epistemic or doxastic logic. (For instance, there are models for Intensional Logic in which the arithmetical predicates have their

intended interpretations at each world and in which a necessity predicate has an interpretation which validates all of S_5.) Within the framework of Intensional Logic sentences such as (1) and (2) are straightforwardly formalizable; e.g. (2) can be rendered by a formula asserting that not every proposition Joe believes is a true proposition.

However, there are other problems with the possible worlds approach that are more serious. One, which arises equally for expressively weak theories like Hintikka's and for more powerful systems such as that provided by Montague's Intensional Logic, is that in all of them exchange of necessarily equivalent sentences within the scope of attitudinal verbs cannot alter the truth value of the whole. For any two sentences that are necessarily equivalent will, according to these theories, identify one and the same attitudinal object. But this does not seem right. As has been noted by many critics of the possible worlds approach towards intentionality, belief and other propositional attitudes do not seem to obey this substitutivity principle. It often appears true to say that someone believes that S_1 but does not believe that S_2, even though S_1 and S_2 are in fact necessarily equivalent.

Advocates of the possible worlds approach will reply that in such cases it isn't strictly speaking true that the subject does not believe that S_2. Rather, he has the belief, but does not recognize it under the description which S_2 gives of it. For some of the cases that have been put forward as counterexamples to the interchangability of necessary equivalents in belief contexts this reply is more plausible than for others. Examples in relation to which it is not particularly plausible are those involving mathematical beliefs. Anyone, we suspect, who has done serious mathematical work will at some point have lacked belief in a hypothesis he subsequently discovered to be a theorem (either by proving the statement himself, or by finding it was already proved by someone else). The possible worlds theorist should, if he wishes to remain consistent with the reply we have just put in his mouth, maintain that, of course, such a mathematician did believe the proposition the hypothesis expresses — which according to standard possible worlds theory is the one and only necessary proposition — all along; it is just that he did not realize that that proposition could be expressed by the words in which the hypothesis happened to be presented.

According to this explanation many things we quite naturally say — e.g. 'I don't believe that', in response to a mathematical statement that happens to be a theorem — are (trivially) false. This seems counterintui-

tive, and an indication that the account is artificial at best. More importantly, even if we were prepared to overlook this, the difficulty that the proposal is meant to overcome would still be with us. For we would now have to acknowledge that among the beliefs people normally hold there are in particular those to the effect that a certain sentence S expresses a given proposition p.[9] Whatever the possible worlds theorist may wish to say about such beliefs, it is clear that they involve sentences essentially: There is no way to construe their contents as involving only propositions, because it is precisely the point of these beliefs that they are *not* preserved under necessary equivalence. So a theory in which there is room for them will necessarily transcend the bounds of possible worlds semantics.

A different response to the criticism is to point out that the possible worlds analysis of belief is concerned only with *implicit*, not with *explicit* belief. Here implicit belief is to be understood roughly as follows: Someone implicitly believes that φ if φ is entailed by the totality of his doxastic commitments, even if he himself cannot recognize φ as expressing a proposition to which he subscribes. This is a belief concept which the possible worlds analysis captures accurately enough, and arguably we do use the verb *believe* at least some of the time to refer to this kind of belief. As implied by the last paragraph, however, it certainly isn't the only notion of belief that enters into ordinary discourse and thought. Nor is it the only kind of belief that is important for logic or philosophy. So, the most that the possible worlds approach can claim is that it gives a satisfactory analysis of *some* of the attitudinal concepts. For the remaining ones it is simply not suited.

There is a multitude of such concepts. Beliefs that relate sentences to propositions they are thought to express are among them, but they do not constitute by any means the only cases. Consider for example the relation which holds between an agent K and a (declarative) sentence S if 'K is justified in asserting S'. This is a relation not unlike knowledge; in particular it would appear to verify the schemata $(K1)$—$(K4)$ we cited in Section 1. At the same time it is a relation whose second arguments are sentences, not propositions, and one that is not invariant under necessary equivalence. To be 'justified in asserting' S — at least on one interpretation of this phrase — you not only need to have a justified belief in the proposition that S; you must also be aware that S expresses that proposition. But you may be aware that the proposition is expressed by a sentence S_1, and yet not be aware that it is expressed

by some other, necessarily equivalent sentence S_2. So you may be justified in asserting S_1 without being justified in asserting S_2, even though S_2 is necessarily equivalent to S_1.

I.4.

These are only some of a number of interconnected reasons why the possible worlds approach becomes untenable when attitude theories are modified so that they reflect ordinary intuitions about attitudinal notions more faithfully and directly, or are expanded to incorporate a larger variety of such notions. To conclude that the possible worlds approach is not the answer to the paradoxical results of Kaplan and Montague no further arguments are, we think, required. Even so we want to mention yet another difficulty, one which does not arise for the expressively weak theories exemplified by Hintikka's, but does affect those which are as powerful as Montague's treatment within the framework of IL. This difficulty is related to an interesting aspect of the general problem that the paradoxes present and that, perhaps, Thomason's words of caution did not, for all their persuasiveness, bring clearly enough into focus.

Thomason emphasizes 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 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 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 consequently feels he may use to arrive at 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 argument of Thomason (1980). 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.[10]

This is a description of the problem which looks at knowledge and belief from what might be called an *internal* perspective, one that focuses on the subject's own reflections about his knowledge and beliefs. We can also look at the issue from an external perspective which concentrates on the practice of knowledge and belief *attribution*, and on the logic and meaning of those sentences of natural language (or of some regimented substitute for it) which serve to make such attributions. The most important difference between these two perspectives relates to the distinct formal frameworks that they suggest for an analysis of the problem which confronts us. As we see it, a formalization suited to the internal perspective ought to be much like that adopted in the cited papers of Kaplan and Montague, one in which the attitudes are represented as predicates that apply to the very expressions in which they themselves occur.[11] The external perspective suggests a different framework, one more like what, if we are not mistaken, Thomason, and many of those for whom his caveats were intended, conceive a representational theory to be like. In such a theory the objects of the attitudes — let us refer to them once more as 'propositions' — are distinct from the sentences of the language in which the theory is stated. But of course propositions and sentences are systematically related. The sentences *express* propositions. Indeed, it is only by analyzing the expression relation that the theory will be in a position to explain what attitude attributions are made by the sentences which are most commonly used for this purpose — those in which an attitudinal verb, noun or adjective (e.g. *believe, belief* or *credible*) is followed by a complement sentence serving to identify the content of the attributed attitude.

It would seem reasonable to demand of such a theory that it be capable of asserting certain intuitively true facts concerning the relation of expression. If it is able to say too much about that relation, however, it will fall prey to the very inconsistencies which the semantic and attitudinal paradoxes produce in theories that treat the relevant notions as sentential or representational predicates. The reason is that paradoxical sentences can be constructed not only when the objects of the attitudes have enough propositional structure intrinsically. It suffices to have a mechanism for correlating the attitudinal objects with the sentences by which they are expressed, and thereby "transferring", as it

were, the syntactic structure of the sentence to the object it expresses. Under certain conditions — viz. when the correspondence is representable within the theory itself — this will lead to paradox in the familiar ways.

To illustrate the point, consider Montague's Intensional Logic, which, as we noted, is immune to the paradoxes so long as it represents the attitudes as relations between individuals and sets of possible worlds. But suppose IL is enriched with enough arithmetic to permit Gödelization (e.g. we add the axioms of Q to the valid sentences of the theory). Let H be some particular Gödelization relation — i.e. n stands in the relation H to the sentence ψ if n is the Gödel number, according to some chosen Gödelization scheme, of ψ. This relation determines a second relation G between numbers and propositions, which holds between n and p if n is the Gödel 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 this system a binary predicate E to represent this relation, and to adopt as new axioms such intuitively valid sentences as: (a) $E(\boldsymbol{n}, \hat{}\psi)$, where $\boldsymbol{n}$ is the numeral denoting the Gödel number n of ψ; (b) $(\forall u)(\mathrm{Sen}(u) \leftrightarrow (\exists!p)E(u, p))$, where 'Sen' is the arithmetical predicate which is satisfied by just those numbers which are Gödel numbers of sentences; and (c) $\forall p(E(\boldsymbol{n}, p) \rightarrow (\check{}p \leftrightarrow \psi))$, where $\boldsymbol{n}$ and ψ are as under (a). However, this addition renders the system inconsistent. For we can now define a 'truth' predicate T of Gödel numbers, viz. by $T(u) \equiv_{\mathrm{df}} (\exists p)(E(u, p) \,\&\, \check{}p))$, and show that $T(\boldsymbol{\psi}) \leftrightarrow \psi$ is provable for arbitrary sentence ψ. The inconsistency then follows as usual.[12]

Let us summarize the conclusions we have reached so far. Thomason's warning must be taken to heart by anyone advocating representational theories of the propositional attitudes. But in fact the perils he observed are even more pervasive than his paper makes clear. They equally affect theories that do not attribute much structure to their attitudinal objects, but which are able to express a good deal about the connection between these objects 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 of reference and quantification, seem solidly protected from this difficulty. But those systems are so weak that they can hardly serve as adequate frameworks for analyzing attitudinal

concepts. Once a framework has the expressive power which a comprehensive account of attitudinal expressions and constructions requires, it will succomb to paradox unless the attitudinal logics it countenances are substantially weaker than the familiar systems of epistemic and doxastic logic.

I.5.

In the two preceding sections we gave some of the considerations that have convinced us that a viable account of the propositional attitudes must acknowledge the syntactic or representational structure of attitudinal objects; and thus that such a theory, if it is to remain consistent within a setting that allows for self-reference, must achieve this by placing fairly rigid limits on the attitudinal logics it endorses.[13] But what are these limits? There are two sides to this question. First, we may see it as a request for a specification, say, in axiomatic terms, of which sets of doxastic, epistemic or other attitudinal principles lead to inconsistency and which do not. Secondly, we may look upon it as the demand for a new conceptual *foundation* of attitudinal logics, one from which such logics will emerge naturally, and in such a way that their compatibility with classical logic and self-reference is warranted by the general method that is used to define them. It is with this second concern in mind that we undertook the formal investigations reported in Part II, and, in particular, adopted the model theory of II.1.2.

Fortunately, we did not have to start from scratch. Over the last fifteen years much progress has been made on the closely related problems that arise in relation to the liar paradox. It is from this work, much of which was initiated by Kripke's seminal *Outline of a Theory of Truth*, that we have taken our inspiration. In fact, Kripke's paper suggests an analysis of necessity along the same lines as the account of truth which he explicitly develops. The model theory we will present here follows this suggestion quite closely. But there is one important difference. Rather than remaining with Kripke's own partial-valued method we have opted for the bivalent theory developed by Herzberger and Gupta.[14]

We have adopted the Herzberger-Gupta approach because it captures certain aspects of the process of belief revision that the reflection on epistemically or doxastically paradoxical statements tends to set in motion. Belief revision is an important aspect of the truth paradoxes as

well, a point that many papers on the liar paradox have stressed and that seems to have motivated Herzberger and Gupta to develop their alternative to Kripke's original idea. But it seems to us that in connection with the epistemic and doxastic paradoxes the revision aspect is especially important. This is illustrated by the hangman paradox, the study of which led Kaplan and Montague to the discoveries mentioned in Section 1.

The story of the hangman has many versions, but they all come to essentially the same thing. Here is one. A judge decrees on Sunday noon that a prisoner, K, is to be hanged at 6.00 a.m. on one of the next seven days, but that K will not know until 30 minutes before the time he will be hanged that the execution will take place on that particular day. By a series of superficially plausible inferences K concludes that the judge's decree cannot be fulfilled. In this way he comes to believe that he cannot be executed, a euphoric condition which is ended abruptly when, at 5.30 a.m. on Wednesday morning, the hangman arrives to lead him away to the gallows. By 6.30 a.m., when death has been officially pronounced, the decree has, K's deductions notwithstanding, been carried out to the letter.

The reasoning K goes through becomes easier to discuss if we simplify the story so that it involves only two days. (As has been observed before, this does not alter or eliminate any of the important features of the paradox.) In this simpler version the decree states that K will be executed at 6.00 on either Monday or Tuesday and that he will not know when he will be executed until 30 minutes before it happens. For convenient reference, let us represent the decree as

$$(1) \qquad (M \,\&\, \neg(\exists t < t_M) K_t M) \vee (T \,\&\, \neg(\exists t < t_T) K_t T),$$

where 'M' ('T') abbreviates 'K will be hanged at 6.00 a.m. on Monday (Tuesday)'; '$K_t\varphi$' abbreviates 'K knows at t that φ'; and 't_M' ('t_T') abbreviates '5.30 a.m., Monday (Tuesday) morning'. K starts out in the belief that (1) is true, accepting it on the strength of the judge's authority. He then reasons as follows: Can they execute me on Tuesday? No. For then, any time after 6.00 a.m. Monday I would know that I have not been executed on Monday, i.e. $\neg M$. This would, in combination with my knowing (1), entail that I know that T at some time before t_T, which contradicts (1). Therefore, since (1) entails $\neg T$, and I know now that (1), I also know now, and thus at some time previous to t_M, that M. But that contradicts (1). So I cannot know (1) after all.

In the story as we have told it, *K* considers this to establish that the decree cannot be fulfilled. It has been claimed that this is a simple fallacy. All that *K* is entitled to infer at this point is that he does not know that (1), not that (1) is false.[15] We do not think however that this bland observation does complete justice to the situation. The inference might be defended as follows. At the outset *K* took (1) to be something he could know because that is the norm for judges' decrees; as a rule we are entitled to take verdicts issued by a sitting magistrate as propositions we know to be true. But after *K* has inferred from this plausible starting assumption that it is after all impossible for him to know that the decree will be fulfilled, it seems not unreasonable that he should come to see the decree as defective, and to reject it.[16]

However, *K*'s subsequent reaction, that of relaxing in the comforting conviction that the execution cannot take place, has no justification. He would have been well-advised to reflect further on the conclusion reached, and to reason as follows. "But if I do not know that the decree is true now, then, presumably, I will not know this either if and when, one of these next two days, they will come and hang me. So it would after all be possible for the decree to be carried out." Having thus concluded that his grounds for rejecting the decree were no good, he might then have returned to his original assumption that the decree is something that he can take himself to know. At this point he would be back where he started, poised for another deductive loop.[17]

As a matter of fact *K* is guilty of a fallacy well before he makes the alleged mistake of concluding that the decree cannot be fulfilled. This fallacy occurs when he infers from the premise that he will know on Monday after 6.00 a.m. that he has not been executed on Monday the conclusion that he will then also know that he will be executed on Tuesday. That inference depends on the further premise that he will know at that point that the decree is true. It is tempting to infer this premise from the standing assumption that the decree is known to him as true at the earlier time when he makes the inference, on the principle that knowledge once gained is to be had forever. But it is precisely in the context of reflections such as those occasioned by the hangman paradox that this principle becomes questionable. For in the course of his reasoning *K* abandons beliefs he previously held. Thus, if any of those beliefs qualified as knowledge while he had it, then — on the reasonable assumption that one cannot be said to know what one does not believe — he no longer has that knowledge once he has given up the belief.

This fallacy illustrates what in recent years has come to be recognized as one of the main sources of *non-monotonicity* in human reasoning.[18] When we reason about our own knowledge or beliefs we often alter our beliefs in the process. As a consequence some of the very inferences we have been drawing may become subsequently unsound, since they were based on premises that are no longer true. This possibility arises already when the subject confines his attention to the present: assumptions about his beliefs or knowledge that were true at the outset may become invalidated by the changes of mind that his reflections have brought about in the meantime. But the non-monotonicity we encounter in the deliberations of K has an added dimension, since they involve projections of what his beliefs will be like in the future. As K's case makes evident, such deliberations can erode the very basis on which these projections were made long before the beliefs on which they are based are in a position to confirm them.

This brief gloss of the hangman puzzle is reminiscent of certain points that have been made in relation to the liar paradox. In particular, the inference which K seems entitled to draw after he has rejected the decree, and in reflecting on this has realized that the decree could be fulfilled after all, recalls a familiar move in reasoning about the liar sentence "I am false": after the sentence has been established as neither true nor false, one notices that this is a state of affairs which contradicts what the sentence says, so that the sentence is false after all. The intuitive justification for this inference is that at the point when it is made the actual state of affairs appears to be not as the sentence says it is. However, we expressly used the word 'appears' here, for whether the actual state of affairs *really* is this way is precisely the point at issue; in fact, only a few additional steps of the argument will make it seem to be just the opposite.

K's inference resembles this move in that it too yields a new assessment of a certain sentence by having another look at a state of affairs that is relevant to its truth.[19] But there is nevertheless an important difference between the two inferences. The person who infers that the liar sentence is false after all does so because at that point the relevant state of affairs *appears* to be contrary to what the sentence says. In contrast, when K concludes that the decree is after all capable of fulfillment, he does so because the state of affairs in question — which, in his case, includes his own beliefs — has undergone a real change.

Of course, by continuing to reason along the same lines K may soon

find himself once again in a state in which he does believe that the decree can be fulfilled. After that we may expect him to reach an uncomfortable equilibrium, in which he simply does not know any more whether or not he should believe the decree. Yet, there would seem to be at least an initial period during which his belief states undergo real changes as his reflections progress. It is for this reason that we see revision as an especially significant aspect of the attitudinal paradoxes, and that we have tried to find a framework which, even if it ignores other important factors — such as, e.g., that of time, or the relation between knowledge and belief — captures something of the doxastic oscillations which these paradoxes tend to set in motion.

I.6.

The framework we have adopted for the formal investigation reported in Part II combines the iterative method developed in the cited papers of Kripke, Herzberger and Gupta with Hintikka's possible worlds analysis of knowledge and belief. In the light of what we have said about the possible worlds approach and our stated preferences for theories that allow attitudes a certain sensitivity to representational form, this is probably surprising. So here is a brief explanation of this choice.

On the one hand our choice is a practical one: most recent technical work on the logic and semantics of the attitudes presupposes the possible worlds analysis. By adopting this same analysis of intensionality, we have, we believe, made it easier to appreciate those respects in which the present treatment differs from extant theories, in particular our use of a semi-inductive revision scheme in an intensional setting, the exploration of which has been our primary concern in the investigations reported in Part II. A framework reflecting our own theoretical prejudices more accurately would certainly have made the comparison with existing proposals more tenuous.

Our choice, however, is not solely one of convenience. Moreover, it is not as much in conflict with our theoretical convictions as it might seem. As we have argued elsewhere,[20] there is not just one doxastic or epistemic "logic", but a variety of them, differing either because they are associated with distinct notions of knowledge or belief, or because they are designed to capture distinct concepts of logical consequence. Many of those alternatives, however, turn out to be substantially weaker than

the systems we have mentioned in Section 1, and often they are weak enough to be compatible with theories which can speak about their own syntax. The familiar logics of implicit knowledge and belief are among the few that need to be revised if these attitudes are treated as predicates of sentences in such theories. Most importantly, the investigations we have begun here chart the limits within which doxastic and epistemic logics must remain if they are to be compatible with these theories. In this way they establish admissibility criteria that any compatible attitudinal logics, including those based on representational conceptions of knowledge and belief, must satisfy.[21]

While these reasons provide a certain justification for the framework we have adopted, we do not regard it as definitive. In future work we hope to reconsider the issues we address in Part II within a more explicitly representational setting.

II. FORMAL DEVELOPMENTS

II.1. Semantics

II.1.1.

The results we mentioned in Section I.1 pertain explicitly to theories which include certain fragments of arithmetic. Such theories can mimic talk about their own syntax by identifying syntactic objects with natural numbers and representing their syntactic properties and relations by means of arithmetical predicates. This offers one way of constructing "self-referential" sentences, sentences that behave as if they were referring to themselves but which, strictly speaking, talk about their own codes. But not all self-reference is by proxy. There are theories that we interpret as referring literally to their own expressions, and as thereby permitting the construction of sentences which are self-referential in the strict sense of the word.

For many purposes it is immaterial whether the theory one studies can speak about its own syntax directly or only vicariously. We have found, however, that for some of our purposes theories that allow for literal self-reference are somewhat more natural. In theories that contain arithmetic, or which have equivalent means for representing syntax, there is what might be called a *global* potential for expressing self-reference. As Lemma 1 of I.1 asserts, such theories have, for each

expressible property *P*, a sentence which is equivalent to the statement that it itself has *P*. However, many actual cases of self-reference do not arise in this way, but rather because a singular term (e.g. a definite description or a demonstrative) picks out, often for contingent reasons, a sentence in which it itself occurs. Such examples have often been studied in abstraction from other instances of self-reference.

While pursuing some of the same issues, we too have found it convenient to investigate the relevant examples in settings where all other instances of non-trivial self-reference are absent. The method by which we secure this involves two separate devices. The first is to consider models in which individual constants denote sentences containing those constants. The second is to make sure that there are no other self-referential sentences, by stipulating (a) that no constants besides the given ones denote sentences in which they occur, and (b) that the syntactic predicates needed to construct self-referential sentences in the first way are not definable. Condition (b) can be imposed by assigning the primitive predicates of the language interpretations that draw no (or only rudimentary) distinctions between sentences.[22] It is in this connection that a framework permitting literal self-reference seems more natural than one in which self-reference arises through arithmetical simulation. In order to limit self-reference in an arithmetical language to just those instances that are the subject of investigation, we would have to give its arithmetical predicates interpretations that are extremely non-standard. While this is technically feasible, the result is an interpreted language that is arithmetic in name only. The distinction between a theory that talks about its own expressions by referring to correlated "numbers" and one which refers to its own expressions directly has at this point become academic. For this reason we will restrict our attention to literal self-reference.

II.1.2.

Kripke (1975), Herzberger (1982), and Gupta (1982) treat truth as a predicate *T* belonging to some language *L* and applicable to the sentences of *L*. They develop methods for approximating the usually unattainable ideal of finding an interpretation for *L* in which the extension of *T* coincides with the set of all *L* sentences true on that interpretation. These methods all involve repeated adjustment of the extension of *T*, where each adjustment consists in making the extension

of T equal to the set of sentences that have just been determined as true. As a rule the process has to be repeated, since a change in the extension of T is likely to alter the set of true sentences as well. The present study follows this work in two ways: the first in that it treats belief — which is the only propositional attitude with which we deal in this formal part of the paper — as a predicate of L which applies to the sentences of L, and the second in that it employs the same techniques to approximate the ideal of an interpretation in which the extensions of the belief predicate coincide with the sets of sentences which, on the interpretation, are in fact believed.

But what is it for a sentence to be "in fact believed"? It is here that we have decided to rely on the possible worlds approach: the sentences that are believed are those that are true in all the doxastically possible worlds. As in Hintikka's theory which we briefly described in Section I.2, the natural formalization of this principle makes belief itself relative to a world: the belief predicate B has an extension at each world w, and at each world w its extension should ideally coincide with the set of sentences true in each of w's doxastic alternatives. When the two do not coincide, an adjustment is called for. We adjust by setting the extension of B at w equal to the set of sentences true in each of w's doxastic alternatives.

We thus arrive at the following semantic framework. Let L be a language of first order logic with identity, which has a denumerable infinity of individual constants, but no function constants of one or more places; we assume that L has two distinguished 1-place predicates: S, a predicate to be thought of as true of all and only the sentences of L; and B, a predicate that is to be thought of as true of all and only those sentences which are believed by some fixed subject K. We will refer to the set of constants of L as C_L.

A *model* for L is a quadruple $\mathcal{M} = \langle W, R, D, [\![\ ]\!] \rangle$ such that:

(i) W is a set (of possible worlds);

(ii) R is a binary relation on W (wRw' means that w' is a doxastic alternative for K in w; we will usually denote the set of alternatives to w as $[wR]$);

(iii) D is a function that assigns to each $w \in W$ a non-empty set D_w (the domain of individuals at w);

(iv) $[\![\ ]\!]$ is a function which assigns to each non-logical constant

of L at each world a suitable extension: if c is an individual constant of L, $[\![c]\!]_w \in \bigcup_{w \in W}(D_w)$; and if Q is an n-ary predicate of L, $[\![Q]\!]_w \subseteq (\bigcup_{w \in W}(D_w))^n$;

(v) for each $w \in W$, $[\![S]\!]_w$ is the set of sentences of L;

(vi) for each $w \in W$, $[\![S]\!]_w \subseteq D_w$;

(vii) for each $w \in W$, $[\![B]\!]_w \subseteq [\![S]\!]_w$.

For simplicity, we assume that models satisfy the following additional conditions:

(viii) the domain of individuals is constant — i.e., for all w, $w' \in W$, $D_w = D_{w'} = D$;

(ix) each individual constant c is a *rigid designator*, i.e., for all w, $w' \in W$, $[\![c]\!]_w = [\![c]\!]_{w'}$.

(x) the model is *referentially complete*, i.e. for each $d \in D$ and $w \in W$, there is a constant c of L such that $[\![c]\!]_w = d$.[23]

Besides the notion of a model we also need that of a *model structure*. Model structures are like models except that they do not assign extensions to the predicate B. In other words they are models for the language L-$\{B\}$ rather than for L. In general there is more than one way of expanding a model structure into a model. Model structures will be indicated by ordinary capitals, e.g. M, while we use script letters, e.g. $\mathscr{M}$, to refer to models.

Any model $\mathscr{M}$ appears to provide two distinct means of determining whether K believes that φ in a given world w. The first is to check whether φ is true in all worlds w' such that wRw'. The second is to check whether $\varphi \in [\![B]\!]_{\mathscr{M}, w}$. Ideally it should not matter which means we choose; the extension of B should correctly reflect the beliefs K has at w, as determined by the set of worlds that stand in the relation R to w. So the following analogue of Tarski's famous requirement ought to hold:

Convention B: $\psi \in [\![B]\!]_{\mathscr{M}, w}$ iff ψ is true in $\mathscr{M}$ at all $w' \in [wR_{\mathscr{M}}]$.

We call $\mathscr{M}$ (*doxastically*) *coherent* iff Convention B is satisfied for each sentence ψ and each world $w \in W_{\mathscr{M}}$.

Many models fail to meet Convention B. In an incoherent model $\mathscr{M}$,

$[B]_{\mathcal{M}}$ is a faulty record of what the alternativeness relation R has to say about K's beliefs. To correct the faults, $[B]_{\mathcal{M}}$ must be adjusted, so that for all $w \in W_{\mathcal{M}}$ $[B]_{\mathcal{M},w}$ equals the set of sentences true at all $w' \in [wR_{\mathcal{M}}]$.[24] However, since adjustments in the extension of B at $w' \in [wR_{\mathcal{M}}]$ can cause certain sentences to change their truth values at w', the new extension of B at w may still be out of synch with what is true in all of w's alternatives. In that case the new model is still incoherent, and a further round of adjustments is called for. Often the adjustments will have to be made again and again before coherence is reached. Sometimes coherence will remain unattainable no matter how many times the procedure is repeated. In these respects our theory resembles the semi-inductive truth theories of Herzberger (1982) and Gupta (1982).

Another point of similarity with these theories is that even when the ideal of full coherence is unattainable, repeated adjustments may lead to closer approximations of that ideal, in which there are fewer and fewer exceptions to Convention B. It may happen, moreover, that any finite number of adjustments yields a model that can further be improved by another revision. In such situations, it is desirable to collect the successive improvements gained through an ω-sequence of revisions in a single model. Gupta and Herzberger propose different rules for doing this. Their rules handle the intuitively unequivocal cases in the same way but differ on the cases in which intuition does not provide firm guidelines. In fact, they are only two of an open ended set of alternative rules for revision, all of which coincide in the cases where intuitions are clear. We have chosen Herzberger's rule of revision, in which the extension of B at w is to consist of all and only those sentences which have been in $[B]_w$ uninterruptedly from some stage onwards. Thus, the new extension records φ as believed (at w) if and only if this has been on record unchallenged throughout some end segment of the infinite sequence of revisions. Although we have no conclusive arguments favoring this rule over all its competitors, we will show some of the implications of this choice for the logic of belief in Section II.2.[25]

As with the mentioned theories of truth, the model which results from collecting all revisions achieved at finite stages of the adjustment procedure may still be capable of further improvements. Again, it may be that no finite sequence of further improvements produces an ap-

proximation that is optimal. In that case we should once more collect the improvements that have been achieved at all those stages. In general we want to be able to collect the benefits of any unbounded sequence of revisions (at successor stages) and collections (at limit stages) into a single model.

Following Herzberger, then, we arrive at the following general definition of model revision. Given any model $\mathcal{M}$ we define for each ordinal α the model $\mathcal{M}^\alpha$, where $\mathcal{M}^\alpha = \langle W_{\mathcal{M}}, D_{\mathcal{M}}, R_{\mathcal{M}}, [\ ]^\alpha \rangle$, $[\theta]^\alpha = [\theta]_{\mathcal{M}}$ for all nonlogical constants θ other than B, and $[B]^\alpha$ is defined as follows:

(i) $[B]^0_w = [B]_w$

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

(iii) For limit ordinal α,
$[B]^\alpha_w = \{\varphi : (\exists \beta < \alpha)(\forall \gamma)(\beta \leqslant \gamma < \alpha \rightarrow \varphi \in [B]^\gamma_w)\}$

We say that a sentence φ is *stably true (false) in* a model $\mathcal{M}$ *at* a world w iff $[\varphi]_{\mathcal{M}^\beta, w} = 1 ([\varphi]_{\mathcal{M}^\beta, w} = 0)$ for all β. φ is *positively (negatively) stable in* a model $\mathcal{M}$ *at* a world w iff $\varphi \in (\notin) [B]^\beta_{\mathcal{M}, w}$ for all β. Evidently, if $[c]_{\mathcal{M}} = \varphi$ then φ is positively (negatively) stable in $\mathcal{M}$ at w iff $B(c)$ is stably true (false) in $\mathcal{M}$ at w. φ is called *stable in* $\mathcal{M}$ *at* w iff φ is either stably true or stably false in $\mathcal{M}$ at w, and *stable in* $\mathcal{M}$ iff it is stable in $\mathcal{M}$ at all $w \in W_{\mathcal{M}}$. Furthermore, φ *stabilizes at* an ordinal α *in* a model $\mathcal{M}$ (*at* a world w) iff α is the first ordinal β such that φ is positively or negatively stable (at w) in $\mathcal{M}^\beta$. α is a *stabilization ordinal for* $\mathcal{M}$ (*at* w) iff every φ that stabilizes in $\mathcal{M}$ (*at* w) stabilizes at some ordinal $\leqslant \alpha$ in $\mathcal{M}$ (*at* w). Along the lines of Herzberger (1982), it can be shown that for every model $\mathcal{M}$ there is a γ such that (i) for every sentence φ and any world w if φ stabilizes in $\mathcal{M}$ at w then φ is positively or negatively stable in $\mathcal{M}^\gamma$ at w, and (ii) for each w and $\beta \geqslant \gamma$ $[B]^\gamma_{\mathcal{M}, w} \subseteq [B]^\beta_{\mathcal{M}, w}$. We call γ a *perfect stabilization ordinal for* $\mathcal{M}$, and call $\mathcal{M}^\gamma$ *semistable*. If β is any ordinal greater or equal to the first perfect stabilization ordinal for $\mathcal{M}$, the model $\mathcal{M}^\beta$ is called a *metastable* model.

A model structure M which assigns viciously paradoxical interpretations to certain sentences cannot be turned into a coherent model in any way whatever. We will call such a model structure essentially incoherent: M is *essentially incoherent* iff every model that is an expansion of M is incoherent.

II.1.3.

Whether a model $\mathcal{M}$ can be made coherent through revision depends on a number of factors. Three of these factors will be of special importance in what follows. They are: (i) the properties of the alternativeness relation $R_{\mathcal{M}}$ (i.e., whether $R_{\mathcal{M}}$ is transitive, etc.); (ii) the initial extension $[\![B]\!]^0_{\mathcal{M},w}$ and (iii) the kinds of self-reference that are realized in $\mathcal{M}$. What is meant by (i) and (ii) should be clear. But (iii) requires explanation.

There are essentially two semantic mechanisms by means of which self-reference can arise, naming and quantification. Self-reference through naming, which we shall call *designative* self-reference, arises when a constant refers to a sentence which contains that constant as a constituent. Some instances of designative self-reference — for instance that of a constant c which denotes the sentence $B(c)$ — are relatively innocuous. An example of troublesome designative self-reference is that of a constant b which denotes the sentence $\neg B(b)$. Indeed, we will see in Section II.1.5 that a model structure M such that $[\![b]\!]_M = \neg B(b)$ is essentially incoherent unless its alternativeness relation has some quite counterintuitive property (see Proposition 7 below).

As has been known at least since medieval times, designative self-reference need not involve just one constant. For instance, it may happen that b denotes the sentence $\neg B(c)$ while c denotes the sentence $B(b)$. A model structure in which this is so will be essentially incoherent under the same conditions which entail essential incoherence for a structure in which b denotes $\neg B(b)$. In general this kind of "self"-reference may involve any finite number of constants. So the relevant general definition of designative self-reference is that there are constants $c_1, \ldots, c_n$, such that for all i with $1 \leqslant i \leqslant n-1$, c_i denotes a sentence containing c_{i+1}, and c_n denotes a sentence containing c_1. When the set $\{c_1, \ldots, c_n\}$ satisfies these denotation conditions in the model structure M we call it a *self-referential set in M*. We say that M *has designative self-reference* iff there is some set of constants $\{c_1, \ldots, c_n\}$ which is self-referential in M.

The other kind of self-reference arises whenever a quantifier in a sentence φ ranges over a set to which φ itself belongs. In our models, all of whose universes include the set of sentences of L, such self-reference is literally ubiquitous. However, there are situations in which this kind of self-reference is nonetheless harmless. In the next section we will take a closer look at the conditions under which this is so.[26]

II.1.4.

Since the subject of the present section is quantificational self-reference, we will restrict our attention to model structures in which there are no instances of designative self-reference. The following condition guarantees this. Given any model structure M, let $<_M$ be the transitive closure of the relation which holds between two constants c_1 and c_2 iff $[\![c_2]\!]_M$ is a sentence of L containing c_1 as a constituent. Evidently M has designative self-reference iff $<_M$ has a loop. So the model structures to be considered here are those for which $<_M$ is loop-free. As a matter of fact we will impose a slightly stronger constraint, viz. that $<_M$ be well-founded.[27]

Through Gödel's work we know that in certain contexts, exemplified by theories of first order arithmetic, the power of quantificational self-reference is unlimited. Lemma 1 gives a concise statement of this fact. But for this to be so it is not enough that the sentences of L belong to the domain of quantification. Gupta (1982) discusses this matter at length. He addresses in particular the question precisely of how much we may allow a theory to say about its own syntax before quantificational self-reference leads to paradox. The results of this section largely follow his discussion. But Gupta is concerned with truth, not belief, and therefore he does not have to contend with the complexities of intensionality. To be precise, the models he and other truth theorists have studied can be identified with those model structures for which the set W is a singleton $\{w\}$ and the relation R consists of the single pair $\langle w, w\rangle$.[28] We will call such model structures, as well as the models that expand them, *extensional.*

Gupta notices that one type of situation in which quantification over the set of sentences is harmless arises when no predicate other than the one under scrutiny — for him the truth predicate T, for us the belief predicate B — is capable of making any distinctions between different sentences. Gupta calls a predicate that lacks this capacity *sentence-neutral.* In our terminology, an n-place predicate Q of L is *sentence-neutral in* a model structure M iff for each $w \in W_M$, each i such that $1 \leqslant i \leqslant n$ and all $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n \in D_M$ and $s, s' \in [\![S]\!]_M$, $\langle a_1, \ldots, a_{i-1}, s, a_{i+1}, \ldots, a_n\rangle \in [\![Q]\!]_{M,w}$ iff $\langle a_1, \ldots, a_{i-1}, s', a_{i+1}, \ldots, a_n\rangle \in [\![Q]\!]_{M,w}$. We say that M is *sentence-neutral* iff every predicate of L other than B is sentence-neutral in M.

One of Gupta's statements is to the effect that every (extensional)

sentence-neutral model structure can be expanded to a coherent model. This result generalizes to non-extensional model-structures.[29] Our proofs of this and related results follow a somewhat different method from Gupta's; we make essential use of a normal form lemma, which we state here without proof.[30] This lemma relies on the following notion. Let $\mathcal{M}$ be a model and let φ and ψ be formulae of L whose free variables are among $x_1, \ldots, x_n$. We say that φ and ψ are *$\mathcal{M}$-equivalent* iff for any $w \in \mathbf{W}_{\mathcal{M}}$ and for any $a_1, \ldots, a_n \in D$, $[\![\varphi(a_1, \ldots, a_n)]\!]_{\mathcal{M}, w} = [\![\psi(a_1, \ldots, a_n)]\!]_{\mathcal{M}, w}$.

LEMMA 2. *Let M be a sentence-neutral model structure and let $\mathcal{M}$ be a model which can be obtained by revision of some expansion $\mathcal{M}'$ of M (i.e. $\mathcal{M} = \mathcal{M}'^{\alpha}$ for some $\alpha \geqslant 1$). Then each formula $\varphi(\vec{y})$ of L is $\mathcal{M}$-equivalent to a Boolean combination $nf(\varphi)$ of formulae of the three following forms*: (i) *$B(x)$ for some variable x occurring in the list $\vec{y}$*, (ii) *$B(c)$ for some constant c occurring in φ*, (iii) *formulae not containing B*, (iv) *$B(c_0)$ where $[\![c_0]\!]_M = (\exists x)\ x \neq x\ \&\ \neg(\exists x)\ x \neq x$. Moreover, if $\forall w \in W_M[wR] \neq 0$ then $nf(\varphi)$ can be taken to be a Boolean combination of formulae of the forms* (i), (ii) *and* (iii) *only.*

PROPOSITION 1. *Let M be any sentence-neutral model structure such that $<_M$ is well-founded and let $\mathcal{M}$ be a model expanding M. Then for some α, $\mathcal{M}^{\alpha}$ is coherent.*

It is possible to extend this result further by weakening the assumption of sentence neutrality. One way in which Proposition 1 can be strengthened is the following. Define for any set A of sentences of L a model structure M to be *A-neutral* just in case for any non-logical n-ary predicate Q other than B, all $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n \in D_{\mathcal{M}}$ and $s, s' \in [\![A]\!]_{\mathcal{M}}$, $\langle a_1, \ldots, a_{i-1}, s, a_{i+1}, \ldots, a_n\rangle \in [\![Q]\!]_{\mathcal{M}, w}$ iff $\langle a_1, \ldots, a_{i-1}, s', a_{i+1}, \ldots, a_n\rangle \in [\![Q]\!]_{\mathcal{M}, w}$. Gupta (1982), who considers a number of such extensions, argues that whenever M is extensional and A is the set of sentences ungrounded in M according to any one of the valuation schemes mentioned in Kripke (1975), M can be expanded to a coherent model. To generalize this result to arbitrary model structures, however, we need an intensional equivalent of the notion of groundedness.

To prove that a model structure M can be made into a coherent model provided it is neutral with respect to the V-ungrounded sentences for any of the valuation schemes V mentioned in Kripke (1975),

we make use of a result in Gupta (1982) and Herzberger (1982): for any such valuation scheme V, if φ is a sentence that is V-grounded in M, then φ stabilizes in the *primary* model expansion of M. Using this lemma we obtain the desired result from the following proposition.

PROPOSITION 2. *Suppose* (i) *M is a model structure for L*; (ii) *$<_M$ is well-founded*; (iii) *A is some set of sentences of L*; (iv) *M is A-neutral*; (v) *$\mathscr{M}$ is an expansion of M*; (vi) *the set U of sentences of L which do not stabilize in $\mathscr{M}$ is included in A. Then for some α, $\mathscr{M}^\alpha$ is coherent.*

Proposition 1 can also be strengthened in a different direction. Gupta notes that for extensional model structures coherence is still attainable when certain syntactic relations between sentences are expressible in L. For instance, if M is a model structure in which a 2-place predicate Neg and a 3-place predicate Con of L are interpreted as the relations 'x is the negation of y' and 'x is the conjunction of y and z', respectively, while all other predicates are sentence-neutral and $<_M$ is well-founded, then it is still true that every expansion of M becomes coherent upon repeated revision. We do not know whether this result generalizes to all model structures. However, the following partial result covers a substantial number of cases:

PROPOSITION 3. *Let M be a model structure such that* (i) *$<_M$ is well-founded*; (ii) *for all $w \in W_M$ $[\![\text{Neg}]\!]_{M,w}$ is the set of all pairs $\langle \varphi, \psi \rangle$ of sentences of L such that φ is the negation of ψ, and $[\![\text{Con}]\!]_{M,w}$ is the set of all triples $\langle \varphi, \psi_1, \psi_2 \rangle$ of sentences of L such that φ is the conjunction of ψ_1 and ψ_2*; (iii) *all predicates of L other than* Neg, Con *and B are sentence-neutral in M. Moreover let $\mathscr{M}$ be an expansion of M such that* (iv) *the sentence*

$$(1) \qquad (\exists x)(\exists y)(\text{Neg}(x, y) \,\&\, \neg B(x) \,\&\, \neg B(y))$$

stabilizes in $\mathscr{M}$ at all $w \in W_M$. Then there is an α such that $\mathscr{M}^\alpha$ is coherent.

There are a number of different conditions on M which guarantee that every expansion of M satisfies (iv). One of these is:

(v) For all $w \in W_M$ if $|[wR_M]| \geqslant 2$ then there are w_1 and $w_2 \in [wR_M]$ and a B-free sentence φ of L such that $[\![\varphi]\!]_{M,w_1} \neq [\![\varphi]\!]_{M,w_2}$.

It is easily seen that if $\mathcal{M}$ satisfies (v) then for any $w \in W_{\mathcal{M}}$ and $\alpha \geqslant 1$ (1) is true in $\mathcal{M}^{\alpha}$ at w iff $|[wR_{\mathcal{M}}]| \geqslant 2$. (v) is entailed by the simpler and, it seems to us, equally plausible condition of M being differentiated: we say that M is *differentiated* iff for any two distinct members w_1 and w_2 of W_M there is a B-free sentence φ of L such that $[\![\varphi]\!]_{M, w_1} \neq [\![\varphi]\!]_{M, w_2}$.

Another condition entailing (iv) is the transitivity of R_M. For any model $\mathcal{M}$ and $w \in W_{\mathcal{M}}$ let $T_{\mathcal{M}, w}$ be the set of sentences true in $\mathcal{M}$ at w, $\{\varphi : [\![\varphi]\!]_{\mathcal{M}, w} = 1\}$, and let for any subset $W' \subseteq W_{\mathcal{M}}$ $T_{\mathcal{M}, W'} = \bigcap\{T_{\mathcal{M}, w} : w \in W'\}$. Evidently for every $\mathcal{M}$, $w \in W_{\mathcal{M}}$ and ordinal α $[\![B]\!]_{\mathcal{M}^{\alpha+1}, w} = T_{\mathcal{M}^{\alpha}, [wR]}$. It follows that (1) is true at w in $\mathcal{M}^{\alpha+1}$ iff $T_{\mathcal{M}^{\alpha}, [wR]}$ is an incomplete theory. So if for some α either for all $\beta \geqslant \alpha$ $T_{\mathcal{M}^{\beta}, [wR]}$ is complete or else for all $\beta \geqslant \alpha$ $T_{\mathcal{M}^{\beta}, [wR]}$ is incomplete, we are done. Suppose then that $T_{\mathcal{M}^{\beta}, [wR]}$ is complete. Then clearly $T_{\mathcal{M}^{\beta}, [w'R]}$ is complete for all $w' \in [wR]$, and $T_{\mathcal{M}^{\beta}, w_1} = T_{\mathcal{M}^{\beta}, w_2}$ for all $w_1, w_2 \in [wR]$. Note that this entails in particular that for $w_1, w_2 \in [wR]$ $[\![Q]\!]_{w_1} = [\![Q]\!]_{w_2}$ for all predicates Q of L other than B. Because R is transitive, we have for all $w' \in [wR]$ that $[w'R] \subseteq [wR]$. So we will have for each such w' that $[\![B]\!]_{\mathcal{M}^{\beta+1}, w'}$ is a complete theory. Also for $w_1, w_2 \in [wR]$, $T_{\mathcal{M}^{\beta}, [w_1R]} = T_{\mathcal{M}^{\beta}, [w_2R]}$ and $[\![B]\!]_{\mathcal{M}^{\beta+1}, w_1} = [\![B]\!]_{\mathcal{M}^{\beta+1}, w_2}$. We distinguish between two cases. (i) The sets $[\![B]\!]_{\mathcal{M}^{\beta+1}, w'}$ are all equal to the set of all sentences. This means that for each of the $w' \in [wR]$, $[w'R] = \emptyset$. It is easily seen that then for $\alpha \geqslant 1$ $T_{\mathcal{M}^{\alpha}, w'}$ is constant and the same for all $w' \in [wR]$. So $[\![B]\!]_{\mathcal{M}^{\alpha}, w}$ is complete for all $\alpha \geqslant 2$. Second, suppose that the sets $[\![B]\!]_{\mathcal{M}^{\beta+1}, w'}$ are equal to a complete consistent theory. Then for each $w' \in [wR][w'R] \neq \emptyset$. This implies that for each β $T_{\mathcal{M}^{\alpha+\beta}, w'} = T_{\mathcal{M}'^{\beta}, w'}$ where $\mathcal{M}'$ is the model whose only world is w', w' is R-related to itself, and $[\![\ ]\!]_{\mathcal{M}', w'} = [\![\ ]\!]_{\mathcal{M}^{\alpha}, w'}$. However, $\mathcal{M}'$ satisfies condition (v) and so, by Proposition 3 there will be an ordinal γ such that $\mathcal{M}'^{\gamma}$ is coherent. So for all $\beta \geqslant \alpha + \gamma$ $T_{\mathcal{M}'^{\beta}, w'}$ will be the same, and moreover this set will be the same for all $w' \in [wR]$. So for all $\beta \geqslant \alpha + \gamma$ $T_{\mathcal{M}^{\beta}, [wR]}$ will be complete. So (1) stabilizes to falsity in $\mathcal{M}$ at w. □

We know from extant work on the liar paradox that when certain syntactic relations become expressible, coherence is no longer attainable. In this connection it is best to return briefly to the setting which we abandoned at the beginning of Part II, that in which self-reference

arises through arithmetization. Adopt a Gödelization scheme G. Where δ is any syntactic object let $\mathrm{gn}(\delta)$ be the number assigned to δ by G. Assume that L has predicate constants O, $'$, $+$ and $\cdot$, of 1, 2, 3 and 3 places, respectively, and consider model structures M which

(a.1) incorporate the standard model of arithmetic at each of their worlds; i.e., D includes the set of natural numbers and at each $w \in W_M$ $[\![\ ]\!]_M$ assigns to O, $'$, $+$, $\cdot$ their standard interpretations. (Thus $[\![O]\!]_{M,w}$ consists of the number zero only, $[\![']\!]_{M,w}$ is the successor relation, etc.).

Assume further that

(a.2) for all $w \in W_M$ $[\![S]\!]_{M,w}$ is the set of all numbers $\mathrm{gn}(\varphi)$ where φ is a sentence of L,

and

(a.3) for any expansion $\mathscr{M}$ of M and ordinal α $[\![B]\!]^{\alpha+1}_{\mathscr{M},w}$ is to be the set of all $\mathrm{gn}(\varphi)$ such that $(\forall w' \in [wR_M])[\![\varphi]\!]_{\mathscr{M}^\alpha,w'} = 1$.

In models of this kind, Lemma 1 applies in that for any formula $\psi(v_0)$ of L there is a sentence φ such that $\varphi \leftrightarrow \psi(\boldsymbol{\varphi})$ is true in $\mathscr{M}$ at all $w \in W$.[31]

As such model structures provide a general licence for self-reference, one would expect them to be essentially incoherent. However, this is so only if the alternativeness relation satisfies certain constraints. For instance, if

$(C1)$ $\quad \forall w([wR] = \varnothing \vee \forall w'(w' \in [wR] \rightarrow [w'R] = \varnothing))$

then M can be expanded to a coherent model. The reason for this is obvious: If w is a world such that $[wR_M] = \varnothing$ then in any expansion $\mathscr{M}$ of M all sentences will be stable at w after one revision, and if $[wR_M] \neq \varnothing$ but $(\forall w' \in [wR])[w'R] = \varnothing$ then every sentence becomes stable at w after at most two revisions.

Although $(C1)$ does not have much *a priori* plausibility as a constraint on relations of doxastic accessibility, we should perhaps not exclude the possibility that K's actual doxastic situation is reflected by a model structure verifying $(C1)$. K could be convinced that his beliefs are inconsistent, even though as a matter of fact they are not. In that case all the alternatives to the actual world would be worlds w' such that $[w'R] = \varnothing$. But even if we regard $(C1)$ as possible, we certainly

would not want to impose it as a general constraint. For it should also be possible for K to believe that his beliefs are consistent and to be right about this, in which case the actual world would constitute a counterexample to the universal claim that $(C1)$ makes.

On the other hand there are many model structures satisfying (a.1) and (a.2) which are essentially incoherent. In particular, this is the case if R_M is transitive and satisfies the following condition $(C2)$:

$$(C2) \quad (\exists w \in W_M)([wR_M] \neq \emptyset \,\&\, (\forall w' \in [wR_M])[w'R_M] \neq \emptyset.$$

PROPOSITION 4. *Let M be a model structure satisfying* (a.1), (a.2) *and* $(C2)$ *and in which R is transitive. Then M is essentially incoherent.*

For the proof of this proposition we refer the reader to that of Proposition 5 in Section II.1.5 below. The only additional observation needed to turn that proof into a demonstration of Proposition 5 is the familiar fact that with the means of arithmetic we can construct a sentence $(\forall x)(\psi(x) \rightarrow \neg B(x))$ such that at every $w \in W_M$ the only object satisfying ψ in M at w is that sentence itself.

We are uncertain whether the conclusion of Proposition 4 can be established without the assumption that R_M is transitive. How much importance one wishes to attach to this problem depends in part on the plausibility of transitivity as a general constraint on R. Someone who is convinced that any reasonable semantics for belief should verify the general principle according to which whatever is believed is believed to be believed will perceive at best a technical interest in the issue whether transitivity can be eliminated from the assumptions of Proposition 4. But if one feels that this principle is not part of the logic of belief, one is likely to see the problem as having more than a merely formal significance.

We will have more to say about constraints on the alternativeness relation in the next section.

II.1.5.

In this section we prove a few results concerning designative self-reference. We begin by having a closer look at a particular instance of this phenomenon, that of a sentence which says of itself that it is not believed. While the behavior of this sentence is in certain ways tied to

its particular meaning, it also illustrates some quite general features of designative self-reference. Later results in this section will make clearer what is special about the sentence and what is not.

Evidently there is a close affinity between sentences that say of themselves that they are not believed and liar sentences, which assert their own falsehood. In fact, within the semantics developed in this paper the former reduce to the latter at any world $w \in W_M$ such that $[wR_M] = \{w\}$, and thus in particular in all extensional model structures. In view of this similarity, one would expect model structures in which this kind of self-reference is realized to be essentially incoherent. As we noted at the end of the last section, however, this is not invariably true; incoherence depends additionally on the properties of R. The next three propositions cover much the same ground as the closing paragraphs of II.1.4. In particular, Proposition 5 is, but for the fact that we deal here with designative self-reference and the precise conditions imposed upon R, analogous to Proposition 4. Partly for future reference let $(C3)$ be the following condition on M:

$(C3)$ $\quad [\![b]\!]_M = \neg B(b)$.

From now on we will make use of the following convention. If c is any constant of L such that in the model structure M under discussion $[\![c]\!]_M$ is a sentence, then we write $\boldsymbol{c}$ instead of $[\![c]\!]_M$.

PROPOSITION 5. *Suppose M is a model structure such that* (i) R_M *is transitive*; (ii) M *satisfies* $(C2)$; *and* (iii) M *satisfies* $(C3)$. *Then M is essentially incoherent.*

Proof. The proof, though simple, is instructive in that it shows in some detail how $\boldsymbol{b}$ behaves under revision. Let $\mathscr{M}$ be any model obtained from M, and suppose that $\mathscr{M}$ is coherent. $(C2)$ tells us that there is some world w such that $([wR_M] \neq \emptyset \ \& \ (\forall w' \in [wR_M])[w'R_M] \neq \emptyset)$. There are two possibilities.

(a) $\boldsymbol{b} \in [\![B]\!]_{\mathscr{M},w}$. Then $\forall w' \in [wR]\, \mathscr{M} \models_{w'} \boldsymbol{b}$. So since $\boldsymbol{b}$ is the sentence $\neg B(b)$, $\forall w' \in [wR]\ \boldsymbol{b} \notin [\![B]\!]_{w'}$. By assumption $[wR] \neq \emptyset$. So let $w' \in [wR]$. Since $\boldsymbol{b} \notin [\![B]\!]_{w'}$, there is a $w'' \in [w'R]$ such that it is not the case that $\mathscr{M} \models_{w''} \boldsymbol{b}$. So $\boldsymbol{b} \in [\![B]\!]_{w''}$. Since R is transitive, $w'' \in [wR]$, which contradicts that $\boldsymbol{b} \in [\![B]\!]_w$.

(b) $\boldsymbol{b} \notin [\![B]\!]_w$. Then there is a $w' \in [wR]$ such that it is not the case that $\mathscr{M} \models_{w'} \boldsymbol{b}$. So $\boldsymbol{b} \in [\![B]\!]_{w'}$. So $(\forall w'' \in [w'R])\mathscr{M} \models_{w''} \boldsymbol{b}$. $[w'R] \neq \emptyset$; so let $w'' \in [w'R]$. Then $\mathscr{M} \models_{w''} \boldsymbol{b}$, and so $\boldsymbol{b} \notin [\![B]\!]_{w''}$. So

there is a $w''' \in [w''R]$ such that it is not the case that $\mathcal{M} \models_{w'''} \boldsymbol{b}$. But since R is transitive, $w''' \in [w'R]$ and so $\mathcal{M} \models_{w'''} \boldsymbol{b}$: contradiction. □

The next proposition shows that if the self-referential sentence $\boldsymbol{b}$ of Proposition 5 is the only instance of designative self-reference in M, M is sentence-neutral and R_M is transitive, then $(C2)$ is a necessary (as well as a sufficient) condition for the essential incoherence of M.

PROPOSITION 6. *Suppose M is a model structure such that* (i) *M is sentence-neutral*; (ii) *R_M is transitive*; (iii) *M satisfies* $(C3)$; *and* (iv) $<_M - \{\langle b, b\rangle\}$) *is well-founded. Then if M is essentially incoherent, it satisfies* $(C2)$.

Proof. The proof of Proposition 6 makes use of Lemma 2 of Section II.1.4. Suppose M is a sentence-neutral model structure with transitive alternativeness relation, which satisfies $(C3)$ and for which $<_M - \{\langle b, b\rangle\}$) is well-founded. Suppose further $(C2)$ does not hold in M. Let $\mathcal{M}$ be a metastable expansion of M. We show that $\mathcal{M}$ is coherent. We proceed by induction on the rank of a sentence *relative* to the set of constants in L distinct from b, which we define as follows: If φ is a sentence of L, then $\mathrm{rk}(\varphi) = 0$ iff φ contains no constants other than b, and if c is a constant of L which denotes such a sentence, or denotes an element of the domain which is not a sentence, then $\mathrm{rk}(c) = 0$. If φ contains constants $c_1, \ldots, c_n$ other than b and if c is a constant denoting φ, then $\mathrm{rk}(\varphi) = \mathrm{rk}(c) = \max(\{\mathrm{rk}(c_1), \ldots, \mathrm{rk}(c_n)\}) + 1$. To prove that every sentence of L is stable in $\mathcal{M}$ at every $w \in W_M$ we proceed as follows. Since $\mathcal{M}$ is metastable, $\mathcal{M} = \mathcal{M}'^{\alpha}$ for some expansion $\mathcal{M}'$ of M and for $\alpha \geqslant 1$. So Lemma 2 applies, giving for each φ an $\mathcal{M}$-equivalent formula $\mathrm{nf}(\varphi)$. Evidently, φ is stable at $w \in W_M$ iff $\mathrm{nf}(\varphi)$ is.

First, we show that the sentence $B(b)$ is stable in $\mathcal{M}$ at all worlds of M. Let w be any world in W_M. From the fact that M does not satisfy $(C2)$ we infer that either $[wR] = \emptyset$ or $(\exists w' \in [wR])[w'R] = \emptyset$. First, suppose $[wR] = \emptyset$. Then for all β, $B(b)$ is true at w in $\mathcal{M}^{\beta}$. So $B(b)$ is stably true at w in $\mathcal{M}$. Now suppose $(\exists w' \in [wR])[w'R] = \emptyset$. Since for all β, $B(b)$ is true at w' in $\mathcal{M}^{\beta}$ and $[\![b]\!]_M = \neg B(b)$, $[\![\boldsymbol{b}]\!]_M$ is false in $\mathcal{M}^{\beta}$ at some world that is an alternative to w. So $B(b)$ is false in $\mathcal{M}^{\beta}$ at w. Since this holds for all $\beta \geqslant 1$ and $\mathcal{M}$ is metastable, $\neg B(b)$ is stably true in $\mathcal{M}$ at w.

To show that every sentence φ is stable in $\mathscr{M}$ at every $w \in W_M$, we proceed by induction on the rank of φ. (i) $\mathrm{rk}(\varphi) = 0$. Then, since φ is a sentence, $\mathrm{nf}(\varphi)$ is a Boolean combination of $B(b)$ and sentences not containing B. Each of the latter is of course stable at all w. So since $B(b)$ is also stable in $\mathscr{M}$ at all $w \in W_M$, it follows that φ is stable in $\mathscr{M}$ at all $w \in W_M$. (ii) Now suppose that for all ψ, where $\mathrm{rk}(\psi) \leqslant n$, ψ is stable at all $w \in W_M$ and assume that $\mathrm{rk}(\varphi) = n + 1$. Then $\mathrm{nf}(\varphi)$ is a Boolean combination of sentences which are either like those mentioned under (i) or else of the form $B(c)$ with $\mathrm{rk}(c) \leqslant n$. If $\mathrm{rk}(c) \leqslant n$, either $[\![c]\!]_M$ is not a sentence, or $[\![c]\!]_M$ is a sentence ψ with rank $\leqslant n$. In the first case $B(c)$ is stably false at all w in any expansion of M. In the second case, ψ is stable in $\mathscr{M}$ at all $w \in W_M$ by the inductive hypothesis. So if w is any world in W_M, ψ is stable in $\mathscr{M}$ at all $w' \in [wR_{\mathscr{M}}]$. This implies that $B(c)$ is stable at w in $\mathscr{M}$. It follows that $\mathrm{nf}(\varphi)$ is a Boolean combination of sentences that are stable at each w and so is itself stable at each w. So φ is stable in $\mathscr{M}$ at all $w \in W_M$. □

Combining Propositions 5 and 6 we obtain

PROPOSITION 7. *Suppose that M is a model structure such that* (i) *M is sentence-neutral,* (ii) *M satisfies* (*C*3), (iii) $<_M - (\{\langle b, b\rangle\})$ *is well founded, and* (iv) R_M *is transitive. Then M is essentially incoherent iff it satisfies* (*C*2).

The truth of Proposition 7 depends on the specific properties of the self-referential sentence $\neg B(b)$. This becomes evident when we compare the model structures which satisfy the hypotheses of Proposition 7 with those discussed towards the end of Section II.1.4, which contain a standard model of arithmetic at each of their worlds. For the latter structures, the conclusion of Proposition 7 fails. For instance, let M' be a model structure such that $W_{M'} = \{\langle w_0, w_0\rangle, \langle w_0, w_1\rangle\}$. M' satisfies (*C*1) but not (*C*2). Suppose also that (*C*3) holds in M'. Then in any expansion of M' the sentence $\neg B(b)$ is stably true in w_0 and stably false in w_1. But nevertheless M' is essentially incoherent. To see this note that by Lemma 1 there is a sentence φ such that $\varphi \leftrightarrow (\neg B(\varphi) \,\&\, B(\neg\varphi))$. φ is easily verified as not stabilizing in any model $\mathscr{M}$ that expands M'.[32]

We now discuss some results concerning the patterns which the sentence ***b*** and other paradoxical sentences follow as they drift in and out of the extensions of B. To this end we introduce a few additional notions. Let $\mathcal{M}$ be a model and $w \in W_{\mathcal{M}}$. For any sentence φ, we understand by the *φ-profile in $\mathcal{M}$ at w* the class of ordinals Π such that $(\forall\alpha)(\alpha \in \Pi \leftrightarrow \varphi \in [\![B]\!]^{\alpha}_{\mathcal{M},w})$. Similarly, for any class of ordinals Γ, the *φ-profile in $\mathcal{M}$ at w on Γ* is to be the intersection of Γ and the φ-profile in $\mathcal{M}$ at w. When $[\![c]\!]_{\mathcal{M}} = \varphi$ we also refer to the φ-profile in $\mathcal{M}$ (at w) as the ***c*-profile in $\mathcal{M}$** *(at w)*. These notions can be straightforwardly generalized to sets of sentences and constants: For any set of sentences Θ of L, the *Θ-characteristic of w in $\mathcal{M}$* is the function $f : \Theta \to \{0, 1\}$ such that for $\varphi \in \Theta$, $f(\varphi) = 1$ iff $\varphi \in [\![B]\!]_{\mathcal{M},w}$. By the *$\Theta$-profile in $\mathcal{M}$ at w*, we understand the function which is defined on the class of all ordinals and maps each ordinal α onto the Θ-characteristic of w in $\mathcal{M}^{\alpha}$. Similarly, if Γ is a class of ordinals, then the *Θ-profile in $\mathcal{M}$ at w on Γ* is the restriction to Γ of the Θ-profile at w in $\mathcal{M}$. Finally, if C is a set of individual constants such that for each $c \in C$ $[\![c]\!]_{\mathcal{M}}$ is a sentence of L and $\Theta = \{[\![c]\!]_{\mathcal{M}} : c \in C\}$, then the *$C$-profile in $\mathcal{M}$ at w (on Γ)* is the Θ-profile in $\mathcal{M}$ at w (on Γ).

PROPOSITION 8. *Suppose that M is a model structure such that R_M is transitive and $[\![b]\!]_M = \neg B(b)$. Then for any $w \in W_{\mathcal{M}}$, the **b**-profile at w in $\mathcal{M}$ on $\omega - \{0\}$ is one of the following*: (i) $\emptyset$, (ii) *the even positive integers*, (iii) *the odd positive integers, and* (iv) $\omega - \{0\}$. *Moreover*, (iv) *arises if and only if* $(\exists w' \in [wR])[w'R] = \emptyset$.

Proof. Assume first that R is serial — i.e., $(\forall w)[wR] \neq \emptyset$. Suppose $w \in W$. We distinguish the following cases.

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

Then $\boldsymbol{b} \notin [\![B]\!]^1_{w'}$ for all $w' \in [wR] \cup \{w\}$. From this and the seriality of R it follows that for $n \geq 2$ and $w' \in [wR]$ $\boldsymbol{b} \in [\![B]\!]^n_{w'}$ iff n is even. So in particular the ***b***-profile at w on $\omega - \{0\}$ is the set of even positive integers.

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

Let w_1 be such a member of $[wR]$. Then for each $w' \in [w_1 R] \cup \{w_1\}$, $\boldsymbol{b} \in [\![B]\!]^1_{w'}$. By an argument similar to that in (a), $\boldsymbol{b} \in [\![B]\!]^n_{w'}$ iff n

is odd. Since $w_1 \in [wR]$, this implies that if n is even, $\boldsymbol{b} \notin [\![B]\!]^n_w$. To make further progress, we divide (b) into two subcases:

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

Let w_1 be as assumed. Then $(\forall w_2 \in [w_1R] \cup \{w_1\})(\boldsymbol{b} \in [\![B]\!]^n_{w_2} \leftrightarrow n$ is even). Again because $w_1 \in [wR]$, $\boldsymbol{b} \notin [\![B]\!]^n_w$, if n is odd. By what we have already seen under (b), this implies that the $\boldsymbol{b}$-profile on $\omega - \{0\}$ at w in $\mathscr{M}$ is $\emptyset$.

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

This case requires yet another bifurcation.

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

Then if w_1 is as assumed, we conclude as under (a) that for each $w' \in [w_1R] \cup \{w_1\}$ ($\boldsymbol{b} \in [\![B]\!]^n_{w'}$ iff n is even). This holds in particular for w_1, and as $w_1 \in [wR]$, we conclude that $\boldsymbol{b}$ cannot belong to $[\![B]\!]^n_w$ when n is odd. So in this case the b-profile at w on $\omega - \{0\}$ is again $\emptyset$.

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

Then for each $w_1 \in [wR]$ there is a w_1' such that for all $w'' \in [w_1'R] \cup \{w_1'\}$ ($\boldsymbol{b} \in [\![B]\!]^n_{w''}$ iff n is odd). Consequently, for every $w_1 \in [wR]\boldsymbol{b} \notin [\![B]\!]^n_{w_1}$ for even n, and so $\boldsymbol{b} \in [\![B]\!]^n_w$ when n is odd. So the $\boldsymbol{b}$-profile at w on $\omega - \{0\}$ consists of all the odd positive integers.

If we drop the assumption that R is serial on W, we must also consider w such that $[wR] = \emptyset$ and w such that $(\exists w' \in [wR])[w'R] = \emptyset$. It is easily verified, however, that these give us $\emptyset$ and $\omega - \{0\}$ as $\boldsymbol{b}$-profiles. This completes the proof. □

It is not difficult to extend the result of Proposition 8 so that it covers also the transfinite parts of the profiles of $\boldsymbol{b}$. Our rule for revision at limit ordinals λ has the effect that $\boldsymbol{b}$ is never in $[\![B]\!]^\lambda_w$. An inspection of the proof of Proposition 8 shows that, in the light of this fact, for any limit ordinal λ the $\boldsymbol{b}$-profile at w in $\mathscr{M}$ on $(\lambda + \omega) - \lambda$ is one of the sets: $\emptyset$, $(\lambda + \omega) - \lambda$, $\{\lambda + 2n + 1 : n \in \omega\}$. Indeed, we have:

PROPOSITION 9. *Suppose that $\mathscr{M}$ is as in Proposition 8. Then for any limit ordinal λ and natural number n:*

(i) *if* $[wR] = \emptyset$, *then* $\boldsymbol{b} \in [\![B]\!]_w^{\lambda+n}$

(ii) *if* $(\exists w_1 \in [wR])[w_1 R] = \emptyset$, *then* $\boldsymbol{b} \notin [\![B]\!]_w^{\lambda+n}$

(iii) *if* $[wR] \neq \emptyset$ *and* $\neg(\exists w_1 \in [wR])[w_1 R] = \emptyset$, then $\boldsymbol{b} \in [\![B]\!]_w^{\lambda+n}$ iff n is odd.[33]

It is instructive to compare the behavior of the paradoxical sentence $\boldsymbol{b}$ with the harmlessly self-referential sentence which says of itself that it is believed. Suppose that M is a model structure in which $[\![c]\!]_M = B(c)$ and this is the only non-trivial instance of self-reference (i.e. M is sentence-neutral and $<_M - \{\langle c, c\rangle\}$ is well-founded). Then, if R_M is transitive, any model $\mathcal{M}$ obtained from M will be coherent after one revision. 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 $\boldsymbol{c}$ is not paradoxical, neither is it *V-grounded* for any of the valuations V considered in Kripke (1975). As noted there $\boldsymbol{c}$'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 M does not determine the truth value of $\boldsymbol{c}$: we can always expand M, for each $w \in W_M$, to two different models $\mathcal{M}_1$ and $\mathcal{M}_2$ so that $\boldsymbol{c}$ is true at w in $\mathcal{M}_1$ while in $\mathcal{M}_2$ it is false at w.

The following proposition, whose proof is somewhat long and tedious, presents a self-referential set consisting of two elements.[34] The profiles of this set become, like those of the simpler sets examined so far, cyclical on ω after a finite number of revisions have filtered out the arbitrariness of the initial assignments. More precisely,

PROPOSITION 10. *Suppose M is a model structure such that* (i) $[\![b]\!]_M = \neg B(c)$; (ii) $[\![c]\!]_M = B(b)$; (iii) *M is sentence-neutral*; (iv) $<_M - \{b, c\}^2$ *is well-founded*; (v) R_M *is transitive. Let $\mathcal{M}$ be an expansion of M. Then for each $w \in W_{\mathcal{M}}$, the $\{b, c\}$-profile at w on ω in $\mathcal{M}$ is cyclical after* 4 *with period* $\leqslant 4$.

So far the results presented in this section are all quite easily established. However, as the syntactic connections between the members of self-referential sets get more involved, it becomes increasingly difficult to establish what the profiles of these sets are like. It might seem a natural conjecture that even though the profiles of more complicated self-referential sets follow increasingly intricate patterns, they nevertheless share with the profiles of the sets $\{\boldsymbol{b}\}$ and $\{\boldsymbol{b}, \boldsymbol{c}\}$ of Propositions 5—10 the property of becoming cyclical after some finite

number of revision steps. If this were so, it would mark a contrast between designative and quantificational self-reference. For it is easily seen that the profiles which arise through quantification can have a high degree of complexity.[35]

As it turns out, designative self-reference does not lead to periodical profiles invariably. Only when the alternativeness relation R satisfies certain fairly strong constraints can we be certain that every designatively self-referential set has an ω-profile of finite periodicity. The next proposition establishes a result to this effect.

PROPOSITION 11. *Suppose that M is a sentence-neutral model, that C is a finite set of constants that is self-referential in M, that $<_M - C^2$ is well-founded, that R_M is transitive and Euclidean — i.e., $(\forall w_1, w_2, w_3)((w_1Rw_2 \ \& \ w_1Rw_3) \rightarrow w_2Rw_3)$ — and that $\mathcal{M}$ is an expansion of M. Then there are natural numbers n and m such that for each $w \in M_M$, the C-profile at w on ω in $\mathcal{M}$ is cyclical after n with period m; that is, if $r \geqslant n$ and $s = k \cdot m + r$ then the C-characteristic at w in $\mathcal{M}^s$ equals the C-characteristic at w in $\mathcal{M}^r$.*

Proof. The proof of this proposition is based on the following idea. If M is as described, then $\mathcal{M}$ will, after only one revision, become locally *homogeneous*, in that, for any $w \in W_M$, the extensions of B will be the same at all worlds in the set $\{w\} \cup [wR]$. This follows from the simple observation that R has the specified properties iff it is an equivalence relation on its range and links each element that belongs to its domain but not to its range with exactly one of the equivalence classes generated by that equivalence relation. The homogeneity, moreover, will persist after further revisions. To show that the profile of C becomes cyclical in these worlds after a finite number of revisions, we argue as follows. We define, as in the proof of Proposition 6, a rank on sentences and on the constants denoting them, but this time only on those sentences φ such that for no $c \in C$ and c' occurring in φ $c <_M c'$. An argument by induction on rank similar to the one given in the proof of Proposition 6 shows that if the rank of φ is n, then φ stabilizes at every $w \in W_M$ after n revisions.

By König's Lemma, since C is finite, so is $<_M^{-1}[C]$. Let $C' = <_M^{-1}[C] - C$. Clearly, if $c \in C'$, then c has a rank. Let n_0 be the maximum of all the ranks of constants $c \in C'$. We now use the normal form established in Lemma 2. Let $c_i \in C$. Then if φ_i is a normal form of c_i and $B(c)$ is a constituent of φ_i, $c \in C' \cup C \cup \{c_0\}$, where c_0 is as in Lemma 2. Let n, m be arbitrary numbers $> n_0$ and let

$w' \in \{w\} \cup [wR]$. If $c \in C' \cup \{c_0\}$, then $[B(c)]_{\mathcal{M}^n, w'} = [B(c)]_{\mathcal{M}^m, w'}$. Similarly, if ψ is a B-free constituent of φ_i then $[\psi]_{\mathcal{M}^n, w'} = [\psi]_{\mathcal{M}^m, w'}$. So if the truth values of φ_i at w' in $\mathcal{M}^n$ and $\mathcal{M}^m$ are different this must be because of a difference in the truth values at w' in $\mathcal{M}^n$ and $\mathcal{M}^m$ of one or more constituents $B(c)$ of φ_i with $c \in C$. In other words, $[\varphi_i]_{\mathcal{M}^n, w'}$ is, for each $n > n_0$, determined by the C-characteristic at w' in $\mathcal{M}^n$. We have already noted that sentences of the form $B(c)$ have, for $n \geqslant 1$, the same truth values in $\mathcal{M}^n$ at all $w' \in \{w\} \cup [wR]$, and thus that the C-characteristic in $\mathcal{M}^n$ at w' is the same for each $w' \in \{w\} \cup [wR]$. Let us refer to this C-characteristic for simplicity as the *C-characteristic at level n*. Since the C-characteristic at level n determines for each $c_i \in C$ the truth value of φ_i and thus also that of $[c_i]_M$ at each $w' \in \{w\} \cup [wR]$, it follows that the C-characteristic at level n completely determines the C-characteristic at level $n + 1$. Since there are only 2^k distinct C-characteristics, where k is the cardinality of C, it is the case for each $w' \in \{w\} \cup [wR]$, and so for w, that the C-profile at w' in $\mathcal{M}$ will be cyclical after n_0 with a periodicity $\leqslant 2^k$. □

It is straightforward to extend Proposition 11 to a similar result concerning the full profiles of C.

We do not know if Proposition 11 can be strengthened in any interesting way by weakening the assumptions on R. Certain constraints, however, must be retained, as is evident from the following, somewhat surprising result. There are finite self-referential sets C of constants containing a "distinguished" constant c_1 which have the following property: for each set E of natural numbers $\geqslant 3$ there is a model $\mathcal{M}$ in which R is transitive and serial and a world $w_0 \in W_{\mathcal{M}}$ such that $\{n \geqslant 3 : c_1 \in [B]^n_{\mathcal{M}, w_0}\} = E$. Thus not only does the c_1-profile on ω fail to be cyclical; it can be just about any set whatever. One example is provided by the set $C = \{b, c, d, e, f\}$ with f as distinguished constant and a model $\mathcal{M}$ which verifies the denotation relations:

(1)
$$[b]_{\mathcal{M}} = \neg B(b) \,\&\, \neg B(c) \,\&\, B(d)$$
$$[c]_{\mathcal{M}} = B(b) \vee B(c) \vee \neg B(d)$$
$$[d]_{\mathcal{M}} = B(d)$$
$$[e]_{\mathcal{M}} = \neg B(c) \vee B(d)$$
$$[f]_{\mathcal{M}} = \neg(B(c) \,\&\, B(e) \,\&\, \neg B(d)),$$

as well as some further conditions to be detailed below.

The construction of $\mathcal{M}$ rests on the following observation.

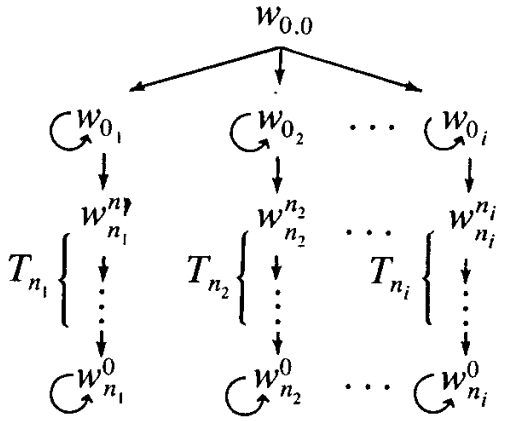
LEMMA 3. *Let T_n be a model such that W_{T_n} is a transitive, linear chain of worlds $w_nRw_{n-1}Rw_{n-2}\ldots Rw_1Rw_0Rw_0$ — in other words, $W_n = \{w_0, \ldots, w_n\}$ and w_iRw_j iff $i > j$ or $(i = 0$ and $j = 0)$ — and such that $[\![b]\!]_{T_n} = (\neg B(b)\ \&\ \neg B(c))$ and $[\![c]\!]_{T_n} = (B(b) \vee B(c))$. Moreover, assume that $\boldsymbol{b} \notin [\![B]\!]_{T_n, w_0}$ and $\boldsymbol{c} \in [\![B]\!]_{T_n, w_0}$ and for $1 \leqslant i \leqslant n$, $\boldsymbol{b} \notin [\![B]\!]_{T_n, w_i}$ and $\boldsymbol{c} \notin [\![B]\!]_{T_n, w_i}$. Then for all $k \geqslant 1$ and $i \leqslant n$,*

$$(\boldsymbol{b} \notin [\![B]\!]^k_{T_n, w_i}\ \&\ \boldsymbol{c} \in [\![B]\!]^k_{T_n, w_i}) \leftrightarrow k \geqslant i.$$

In particular, $(\forall w \in W_n)(\boldsymbol{b} \notin [\![B]\!]^k_{T_n, w}\ \&\ \boldsymbol{c} \in [\![B]\!]^k_{T_n, w})$ iff $k \geqslant n$.

We can exploit this lemma to show that the set C with the denotational equations (1) has the property claimed.

PROPOSITION 12. *Let E be any set of natural numbers $\geqslant 3$. Let $D = \{n - 2 : n \notin E\}$. Let $\mathscr{M}(E)$ be the model which* (a) *satisfies the designation relations in* (1) *and* (b) *has the following world structure:*



where $\{n_1, n_2, \ldots\}$ is an enumeration of D and the W_{Tn_i} are pairwise disjoint as well as disjoint from the set $\{w_{00}, w_{n_1}, w_{n_2}, \ldots, w_{n_i}, \ldots\}$. Moreover, assume that (c) *the extensions of B in $\mathscr{M}(E)$ satisfy the following conditions:*

(i) *if $w \in T_{n_i}$ then $\boldsymbol{b} \notin [\![B]\!]^0_{\mathscr{M}(E), w}$ and $\boldsymbol{c} \in [\![B]\!]^0_{\mathscr{M}(E), w}$ if $|[wR]| = 1$, and $\boldsymbol{b} \notin [\![B]\!]^0_{\mathscr{M}(E), w}$ and $\boldsymbol{c} \notin [\![B]\!]^0_{\mathscr{M}(E), w}$ if $|[wR]| > 1$*

(ii) *$\boldsymbol{d} \in [\![B]\!]^0_{\mathscr{M}(E), w}$ iff $w \in \bigcup_{i \in \omega} W_{Tn_i}$.*

Then for $k \geqslant 3$, $\boldsymbol{f} \notin [\![B]\!]^k_{\mathscr{M}(E), w_{00}}$ iff $k = n_i + 2$ for some $n_i \in D$. So $\{k \geqslant 3 : f \in [\![B]\!]^k_{\mathscr{M}(E), w_{00}}\} = E$.

The proofs of Lemma 3 and Proposition 12 are straightforward. It may be that there are even smaller self-referential sets than C, or sets whose members are connected via simpler denotation relations and which are also capable of generating all sets of natural numbers $\geqslant n_0$ (for some small n_0, e.g. 3 or 2) as the profiles on $\omega - n_0$ of their distinguished members. But this is something we have not bothered to look into. Another question to which we do not have an answer is whether there exist conditions under which the family of sets of natural numbers representable as profiles is somewhat restricted, but not to the point of containing only sets that become cyclical with finite periodicity after a finite number of steps. The technical problems in this domain appear to be quite difficult and almost all of the terrain remains to be explored.

Both Proposition 11 and Proposition 12 raise the type of question we have encountered earlier in this section: can these results be strengthened by relaxing the assumptions they make about the relation R? This is a question that comes to mind with particular force in connection with Proposition 12. For the result established there depends crucially on the world structure of the model $\mathcal{M}(E)$ described in that proposition. If there were some intuitively natural constraints on doxastic alternativeness relations which such models violate, Proposition 12 would lose much of its interest. Similarly, the import of Proposition 11 depends in part on the plausibility of the constraints that it imposes on R. Although from an intuitive perspective these constraints might well appear too strong, we have not found any conceptually natural but weaker conditions on R under which the conclusion of Proposition 11 still holds.

Whether the constraints of Proposition 11 are intuitively unacceptable is a matter open to dispute. But it seems to us that there is a certain perspective from which they are defensible. Recall that R is transitive and Euclidean iff (i) it is an equivalence relation on its range, and (ii) it relates each element of its domain to exactly one of the equivalence classes into which it partitions its range. An accessibility relation of this sort mirrors the informal notion of a believer who has full knowledge of his beliefs, i.e. who, for any sentence φ, believes that he believes that φ if he does believe that φ and believes that he does not believe that φ if he does not believe that φ. We would be reluctant to endorse this principle as an intrinsic part of the meaning of the term 'believe'. Nevertheless, it seems to us that there exists a conception of belief, which is prevalent in every day language and thought, and which

conforms to this principle pretty well. In other words, while the principle does not appear valid for every one of the spectrum of notions that go with the term there are some for which it seems valid.

Here we run into a complication which we already noted in Part I. It appears that the word 'believe' and its cognates do not denote a single concept with its fixed meaning and corresponding logic, but a family of related notions. If by the logic of belief we want to understand that which all members of the family share, then the principle should be excluded. But in so far as it is possible to detach the particular common sense notion we alluded to from the fabric of connections that hold the family together we may accept the principle as part of the logic of that particular notion. The task of separating out the different strands that are woven together in the meaning of 'believe' is one we will not undertake here. All the same we should not lose sight of the plurality that lies hidden behind the apparent unity suggested by a single word. This is especially important in connection with the logical questions which we will consider in the remaining three sections.

II.2. Logic

II.2.1.

The paradox Montague and Kaplan discovered was that languages capable of expressing enough about their own syntax cannot contain sentence predicates which satisfy all the logical principles commonly ascribed to such concepts as knowledge or belief. As we noted in Part I, there are two principal strategies for dealing with this problem. The first is to opt for an analysis which prevents these concepts from being construed, directly or indirectly, as relations to sentences. The second is to treat them openly as such relations, while being prepared to give up some of the principles that Kaplan, Montague, Thomason and others have shown to be jointly inconsistent. As we made clear in Section I, we think there are compelling reasons to pursue the second strategy. Our semantics in Section II.1 develops that strategy into the beginnings of an alternative to the familiar intensional theories in which belief is treated as a property of sets of possible worlds. One aim in developing that semantics was that it should serve as a basis for the logical reassessment that the results of Kaplan, Montague and Thomason show to be necessary. There are however a number of different ways in

which the model theory of II.1 can be used to define a consistent logic of belief. This is a common feature of semantic theories that deviate, in one way or another, from the paradigm set by the classical model theory for the predicate calculus. Classical model theory offers a single, unequivocal explication of the concepts of logical truth and consequence. For most of the alternatives which for a variety of philosophical and linguistic reasons have been developed in recent years this is not so. These theories offer as a rule a number of definitions for those concepts that appear all equally plausible. Consequently, to determine the logic generated by such a theory tends to be a task fraught with conceptual as well as technical difficulties. In connection with the semantics we have developed here these difficulties appear to be particularly severe.

Before we investigate these problems there is a more fundamental question that we must clarify. What are the criteria that a good doxastic logic should satisfy? In Part I we have spoken at several points as if the logical task facing someone who wishes to develop a sentential or representational theory of belief were that of discovering the most parsimonious ways in which the incompatible combinations of general logic, elementary syntax, and special attitudinal logics can be restored to consistency. If this is our only goal, our model theory offers countless ways of pursuing it: with each model $\mathcal{M}$ in which paradoxical self-reference is realized — e.g. one of those discussed at the end of Section II.1.4 — and world $w \in W_{\mathcal{M}}$, we can associate the set of those general principles all of whose instances are true in $\mathcal{M}$ at w. Any such set of principles is clearly consistent with classical logic and the means of self-reference realized in $\mathcal{M}$. If $\mathcal{M}$ is of the sort described in II.1.4 and thus encompasses the full spectrum of self-referential possibilities, the resulting "logic" will be compatible with self-reference generally.

The systems that are obtained in this way satisfy the desideratum, mentioned in Section I.5, of providing safe upper limits for doxastic logics, including those that future investigations may yet turn up. But there is little reason to think of them as having much significance beyond that. Formal consistency is not the only condition that a good logic should meet. For one, a system of logic should not just be free of internal contradiction, it should also be compatible with any compossible set of sentences — i.e. any set of contingent sentences that could be jointly true. It isn't very clear how this constraint can be made precise. For what is it for a set Γ of sentences to be compossible? Indeed, the one analysis of this notion that comes readily to mind is

that Γ is compossible iff it is compatible with the logic governing those concepts of which the sentences in Γ make use; but in the present context that would lead us straight back to the original question, viz. what the logic of belief is. Even so, this criterion provides some informal guide; and, whatever its precise content, the sets of principles that are verified by particular combinations of $\mathscr{M}$ and w are likely to be in violation of it.

Another, connected consideration points the search for doxastic logics in the same general direction. For someone persuaded that a theory of the attitudes must include an account of attitudinal relations to sentences, the moral of the Kaplan-Montague results is not just that some of the familiar attitudinal logics cannot be upheld in a classical setting which allows for paradoxical self-reference. They also carry the deeper message that the intuitions supporting the familiar logics are themselves flawed, and in need of revision. One should not simply look for compatible logics which preserve as much of the old intuitions as possible; rather the search ought to be for a new conceptual foundation on which such logics can be built. It has been with this aim in mind, no less than that of restoring consistency, that the semantics of II.1 was developed.

Both these desiderata, that of formulating a logic which is compatible with any intuitively compossible set and that of resting it on a proper conceptual foundation, suggest that the logically valid sentences should be those that are true for *all* relevant combinations of models and worlds, and not just for one. As implied above, our semantics suggests a number of such definitions between which it seems difficult to make a reasoned choice. At the same time, however, it also seems to exclude some of the avenues along which a consistent belief logic might be constructed. So, lest we be at risk of searching in the wrong place altogether, let us briefly consider whether, or to what extent, we are right to ignore the options which our approach eliminates.

All definitions which characterize validity in terms of some class of pairs $\langle \mathscr{M}, w \rangle$, where $\mathscr{M}$ is a model in our sense and $w \in W_{\mathscr{M}}$, have in common that they preserve the whole of the classical predicate calculus. Thus by confining our quest for a consistent belief logic to an investigation of definitions of this form, we seem to have committed ourselves to classical logic. We should recall in this connection that the incompatibility results which gave the initial impetus to the present study show the joint inconsistency of three distinct components: (i) the underlying general logic — in this case, the classical predicate calculus;

(ii) devices and postulates needed to express paradoxically self-referential sentences and to establish that these have their intended meanings; and (iii) the logic of the particular concept or concepts in question. To restore consistency one could tamper with any one of these components, or with any combination of them.[36]

Tampering with the underlying logic is among those options. Is it among those we ought to explore? Perhaps we should not dismiss it out of hand. In fact, in the light of the disturbing evidence that will emerge in II.2.3, the case for a consistent system which sacrifices some of the background logic as well as of the prima facie desirable doxastic principles may deserve more serious attention than it appears to merit at first view. Nevertheless, we side with those who have, when they confronted this question in relation to the truth predicate, felt that to abandon classical logic is to pay so dear a price that it should be considered only as a last resort. Accordingly, in the next two sections, which address the problem of defining doxastic logics in some more detail, we will concentrate on the problem of finding a satisfactory logic that is compatible with classical logic.

The incompatibility results, as we have just reminded ourselves, involve three components, general logic, the logic of some special concept or concepts, and the devices responsible for self-reference. Evidently changes that delimit or eliminate the self-referential devices are among the surest and simplest ways to restore consistency. But of course to adopt any of those ways would be to abandon the very task that we have set ourselves. Nevertheless we will begin the next section by presenting a result that pertains to coherent models, i.e. to models in which self-reference is either non-existent or harmless. We have included that result not so much for its own intrinsic interest but rather to put the difference between the logics of coherent and those of incoherent models into sharper focus.

Nevertheless, that is an issue of secondary importance. Our principal interest in the next section concerns doxastic logics that are consistent with a combination of classical logic and the full spectrum of self-referential possibilities.

II.2.2.

In the last section we proposed that validity be defined as truth in all relevant models at all relevant worlds. But which models and which

worlds are "relevant"? Here we seem to be facing a number of possible choices. To simplify matters somewhat we will assume that there is no distinction of relevance to be made between worlds; i.e., if $\mathscr{M}$ is a relevant model then every member of $W_{\mathscr{M}}$ is a relevant world.[37] This reduces our problem to that of determining which are the relevant models. There is one aspect to this more limited question that is familiar from the possible worlds analysis of modality: the modal schemata that come out valid correlate, to a very high degree, with the properties that are assumed for the alternativeness relation R.[38] This is equally true of the semantics developed here, except that model incoherence somewhat complicates the picture. The similarity between the modal approach and our own is most clearly visible when we focus on coherent models. The first result of this section, which is concerned with coherent models only, is meant to illustrate this. The result has an obvious counterpart within the modal treatment of belief, and like its modal counterpart it is one of an indefinite number of similar theorems, which differ from each other in the varying assumptions they make about the properties of R. We have chosen the particular theorem we present because it deals, as closely as possible for theorems of its type, with the familiar schemata $(B1)$—$(B4)$ which we introduced in Section I.1.

It is a well-known fact of modal logic that the schemata $B\varphi \rightarrow BB\varphi$ and $B(B\varphi \rightarrow \varphi)$, i.e. the modal analogues of our $(B1)$ and $(B4)$, correspond to the conditions that the alternativeness relation is transitive, and reflexive on its range. Moreover, possible worlds semantics verifies, as a matter of course, the schema $B(\varphi \rightarrow \psi) \rightarrow (B\varphi \rightarrow B\psi)$, the equivalent of our $(B3)$; and, finally, it conforms to the principle that $B\varphi$ is valid whenever φ is, which strengthens the modal counterpart of $(B2)$. (We will refer to the corresponding stronger principle for our language L as $(B2')$.) In modal logic this correspondence between schemata and conditions on R can be made precise in the form of completeness theorems, e.g. the theorem that a sentence of modal propositional logic is true at all worlds in all possible worlds models satisfying the mentioned conditions on R iff it is derivable from (the modal counterparts of) the principles $(B1)$, $(B2')$, $(B3)$ and $(B4)$. Proposition 14 establishes a similar result within our framework. It too has the form of a completeness theorem: a sentence of L is provable from a certain theory T iff it is true in every one of a certain class $\mathscr{B}$ of models (at each world of that model).

The class $\mathscr{B}$ we want to consider consists of coherent models only. We have chosen the simplest means of securing the coherence of the members of $\mathscr{B}$, viz. by insisting that they be expansions of sentence-neutral model structures (that satisfy the specified conditions on R). Since the theory T must "match" $\mathscr{B}$, i.e. must be strong enough to guarantee that any sentence consistent with it is true at some world in some member of $\mathscr{B}$, T must contain the information that the predicates of L other than B are sentence-neutral. This entails, however, that T cannot express each of the schemata $(B1)$, $(B2')$, $(B3)$, $(B4)$ as a single universal sentence. Therefore, if it is to contain the information that the schemata are valid, it must do so by having each one of the instances of those schemata as a separate axiom. That, however, is possible only if T has names for all the sentences of L. Thus in order to state the theorem, we must assume the existence of such names. So let c_φ be a fixed function which maps the sentences of L one-to-one onto some coinfinite subset of C_L. As the restriction to models that conform to this naming function involves no significant loss of generality, we assume for the remainder of the paper that, for every model structure M and sentence φ, $[\![c_\varphi]\!]_M = \varphi$, and that all models are expansions of such M.

Let $\mathscr{B}$ be the class of all coherent models $\mathscr{M}$ which expand some model structure M such that (i) M is sentence-neutral and (ii) R_M is transitive and reflexive on its range. Let T be the first order theory whose set of axioms $\mathscr{A}$ is defined as follows. Let $\mathscr{A}_0$ be the set consisting of (a) $(\forall x)(B(x) \rightarrow S(x))$; (b) $(\forall \vec{z})(\forall x, y)((Sx \;\&\; Sy) \rightarrow (Q(z_1, \ldots, z_{i-1}, x, z_{i+1}, \ldots, z_n) \leftrightarrow Q(z_1, \ldots, z_{i-1}, y, z_{i+1}, z_n))$ for each n-place predicate Q of L other than B and for each $i = 1, \ldots, n$; (c) $S(c_\varphi)$ for each sentence φ of L; (d) $B(c_\varphi) \rightarrow B(c_\psi)$, where φ is any sentence of L and ψ is the sentence $B(c_\varphi)$; (e) $B(c_\varphi)$ for each theorem φ of first order logic; (f) $B(c_{\varphi \rightarrow \psi}) \rightarrow (B(c_\varphi) \rightarrow B(c_\psi))$ for all sentences φ and ψ; (g) $B(c_\psi)$, where ψ is the sentence $B(c_\varphi) \rightarrow \varphi$ for some sentence φ of L. Let $\mathscr{A}$ be the closure of $\mathscr{A}_0$ under the principle $(B2')$:

$$\text{if } \varphi \in \mathscr{A}, \text{ then } B(c_\varphi) \in \mathscr{A}.$$

We can now state:

PROPOSITION 13. *Let T and $\mathscr{B}$ be as defined above. Then, for any sentence φ of L, $T \vdash \varphi$ iff for every $\mathscr{M} \in \mathscr{B}$ and $w \in W_{\mathscr{M}}$ $[\![\varphi]\!]_{\mathscr{M}, w} = 1$.*

Our proof of Proposition 13, which we omit, follows lines familiar from the literature on modal logic: suppose it is not the case that $T \vdash \varphi$. Then construct a semantic tableau for $\neg\varphi$. This tableau will not close and thus supply a model in which $\neg\varphi$ is true at some world.[39]

Proposition 13 not only brings out the similarities, as well as some of the differences, between our approach and that of traditional modal logic; it also provides a paradigm for analogous theorems that apply to classes which include incoherent models. Unfortunately there exists a serious obstacle to such results. The difficulty we run into is the following. When $\mathscr{M}$ is incoherent, then the familiar correspondences between properties of $R_{\mathscr{M}}$ and the validity of propositional schemata tends to break down. For example, suppose that $R_{\mathscr{M}}$ is transitive, and reflexive on its range, that $[\![b]\!]_{\mathscr{M}} = \neg B(b)$ and that w is a member of $W_{\mathscr{M}}$ such that $[wR] \neq \emptyset$ and $(\forall w' \in [wR])[w'R] \neq \emptyset$. Then the instance of $(B4)$ which we obtain by substituting b for φ, i.e. $B(\boldsymbol{B}(b) \rightarrow \neg\boldsymbol{B}(b))$, will fail in $\mathscr{M}$ at w either at all the odd or else at all the even finite stages $\geqslant 1$. Consequently the sentence will also be false in $\mathscr{M}$ at w at stage ω; indeed, this will be the case at all limit stages. So the schema fails even though the "corresponding" condition on R, reflexivity on its own range, is satisfied. Thus in incoherent models schemata can fail for two quite distinct reasons — either because of the structure of R or because of the nefarious effects of self-reference. A completeness proof pertaining to model classes which contain incoherent models will have to cope with both these factors and the ways in which they may interact.

The fact about $(B4)$ which we have illustrated here with the help of $\boldsymbol{b}$ is a quite general one, as attested by the following proposition. To simplify notation, let us abbreviate the sentence $B(c_\psi)$, where ψ is the sentence $B(c_\varphi) \rightarrow \varphi$, as $B4(\varphi)$.

PROPOSITION 14. *Suppose that $\mathscr{M}$ is a model such that $R_{\mathscr{M}}$ is transitive and Euclidean, and that φ is a sentence that does not stabilize in $\mathscr{M}$ at any ordinal α. Then there is a $w \in W_{\mathscr{M}}$ such that for arbitrarily large ordinals β $[\![B4(\varphi)]\!]_{\mathscr{M}^{\beta}, w} = 0$. Moreover, among these ordinals there are successor ordinals, limit ordinals and perfect stabilization ordinals.*

Proof. Observe that if φ never stabilizes in $\mathscr{M}$, then there must be a pair consisting of a sentence φ and a world w such that φ changes

truth value in $\mathcal{M}^\alpha$ at w for arbitrarily large α. It must be the case that $w \in \mathcal{R}an(R_\mathcal{M})$, for if all sentences were eventually to stabilize in $\mathcal{M}$ at all members of $\mathcal{R}an(R_\mathcal{M})$, they would stabilize also on the remaining worlds. Observe that the switches of truth value must always occur at successor ordinals. Because of the structure of $R_\mathcal{M}$, φ will have for sufficiently large α the same truth value at all $w' \in [wR]$ in $\mathcal{M}^\alpha$. Thus, there will be arbitrarily large ordinals β such that $[\![\varphi]\!]_{\mathcal{M}^\beta, w'} = 1$ for all $w' \in [wR]$ and $[\![\varphi]\!]_{\mathcal{M}^{\beta+1}, w} = 0$. So $[\![\varphi]\!]_{\mathcal{M}^\alpha, w} = 0$, while $[\![B(c_\varphi)]\!]_{\mathcal{M}^\alpha, w} = 1$. Since $B(c_\varphi) \rightarrow \varphi$ fails at w in $\mathcal{M}^\alpha$, $B4(\varphi)$ fails at w in $\mathcal{M}^{\alpha+1}$. To establish the last part of the proposition, note that if γ is the limit of any unbounded sequence of ordinals α such that $B4(\varphi)$ fails at w in $\mathcal{M}^\alpha$ then $B4(\varphi)$ fails at w in $\mathcal{M}^\gamma$. And if α is a perfect stabilization ordinal for $\mathcal{M}$ then $[\![B]\!]_{\mathcal{M}^\alpha, w} \subseteq [\![B]\!]_{\mathcal{M}^\beta, w}$ for all $\beta \geqslant \alpha$. So, since $B4(\varphi)$ fails at w in $\mathcal{M}^\beta$ for some $\beta \geqslant \alpha$, it also fails at α. □

Proposition 14 makes explicit that incoherent models may provide counterexamples to schemata that are validated by the Kripke frames which these models contain. However, such counterinstances always involve unstable sentences. When the sentences that replace the sentence letters in a given schema μ are all stable in $\mathcal{M}$ then the resulting sentence will be true in $\mathcal{M}$ (at any $w \in W_\mathcal{M}$) if μ is valid on the corresponding Kripke frame.

We can combine this observation with Proposition 14 into a single statement, to the effect that $B4(\varphi)$ fails at arbitrarily large α iff φ does not stabilize. To cast this statement in the form that best suits our purpose we need to introduce two further notions. For any class $\mathcal{B}$ of models and sentence φ we say that φ is *stable throughout* $\mathcal{B}$ iff φ is stable in all the members of $\mathcal{B}$. We say that φ is *valid in* $\mathcal{B}$, in symbols $\mathcal{B} \models \varphi$, iff for all $\mathcal{M}$ in $\mathcal{B}$ and w in $W_\mathcal{M}$ $[\![\varphi]\!]_{\mathcal{M}, w} = 1$.

PROPOSITION 15. *Let $\mathcal{B}$ be the class of all metastable models $\mathcal{M}$ such that $R_\mathcal{M}$ is transitive and Euclidean. Then, for any sentence φ, $\mathcal{B} \models B4(\varphi)$ iff φ is stable throughout $\mathcal{B}$.*

Proposition 15 entails that whenever $\mathcal{B}$ is such that for some decidable set S' of sentences of L the set of those members of S' which are stable throughout $\mathcal{B}$ is not recursively enumerable, then the set of sentences of L that are valid in $\mathcal{B}$ is not recursively axiomatizable. One strategy for establishing the consequent of this last statement would be

to take for S' some decidable set of arithmetical sentences. For instance assume, as at the end of II.1.4, that 0, ′, +, · are predicate constants of L, and let N be a 1-place predicate of L distinct from B and S. Let Q^N be the conjunction of the axioms of Robinson's system Q, relativized to N, and let AR be $(Q^N \,\&\, B(c_{Q^N}))$. Let S' be the set of sentences of the form $AR \rightarrow \varphi^N$, where φ is a sentence containing no non-logical constants other than 0, ′, +, · and B. Evidently S' is decidable. Suppose $\mathscr{M}$ is a member of $\mathscr{B}$. Since AR is stable in all metastable models, any sentence $AR \rightarrow \varphi^N$ will be stably true in $\mathscr{M}$ at any $w \in W_{\mathscr{M}}$ at which AR is (stably) false. When AR is true in $\mathscr{M}$ at w, then in w and in all the worlds in $[wR_{\mathscr{M}}]$ the extension of N in $\mathscr{M}$ is a model of Q. Thus, $AR \rightarrow \varphi^N$ is stably true throughout $\mathscr{B}$ iff φ^N is stably true in all members of $\mathscr{B}$ whose restrictions to N are models of Q. If $\mathscr{B}$ were such that all such restrictions were standard models of arithmetic, the complexity of the class of stable φ^N would be Σ^1_2-complete, as follows from Theorem 12.3 of Burgess (1986). However, this is not a plausible assumption about $\mathscr{B}$, and we know at present of no way to establish the non-axiomatizability of $\mathscr{B}$-validity in the absence of this assumption.

This argument appears to rule out the possibility of giving, for a substantial number of prima facie plausible definitions of validity, completeness theorems modeled on that given in Proposition 14. There is, however, a weaker type of completeness result that our most recent observations do not rule out. We have in our informal discussions in this section focused repeatedly on the logical validity of what we have been referring to as "schemata", among them in particular $(B1)$—$(B4)$. Given some particular class of models $\mathscr{B}$, the question precisely which schemata are validated by $\mathscr{B}$ may admit of a simple answer even if the set of L-sentences that are valid throughout $\mathscr{B}$ does not admit of recursive axiomatization. To make this precise it is convenient to introduce a language of propositional doxastic logic in which belief is represented as a sentential operator. So let PL be the language whose atomic sentences are T, $\bot$ and the sentence letters $p1, p2, \ldots$, and which has besides the truth-functional connectives $\neg$, &, $\vee$ the 1-place sentence operator B. We will refer to the formulae of PL as *schemata.* By an *interpretation of PL in L* understand a function I which maps each sentence letter onto a sentence of L. Every interpretation I can be extended to all formulae of PL as follows: $I(\mathrm{T}) = (\forall x)x = x$; $I(\bot) = (\exists x)x \neq x$; $I(\neg\mu) = \neg I(\mu)$; $I(\mu \,\&\, \nu) = I(\mu) \,\&\, I(\nu)$;

$I(\mu \vee \nu) = I(\mu) \vee I(\nu)$; $I(B\mu) = B(c_{I(\mu)})$. When μ is a schema containing only one sentence letter p_i we will write $\mu(\varphi)$ instead of $I(\mu)$ when I is any interpretation such that $I(p_i) = \varphi$. Where $\mathscr{B}$ is a class of models for L, we say that the schema μ *is valid in* $\mathscr{B}$, in symbols $\mathscr{B} \models \mu$, iff $\mathscr{B} \models I(\mu)$ for every interpretation I.

In the next section we will prove a result to the effect that a schema is valid in a certain class $\mathscr{B}$ iff it is derivable from some specified theory T of propositional modal logic. The force of that result will be negative in so far as it establishes that for the class $\mathscr{B}$ to which it applies there are no non-trivial propositional schemata whatever. The model classes that are under discussion in this section, however, do not necessarily generate propositional doxastic logics that are quite that weak. Some of them do validate non-trivial schemata, as is made clear by the next two propositions. We expect that the sets of schemata validated by the particular classes discussed below are in fact axiomatizable within *PL*. However, we have no such results to offer at the present time.

In the proof of Proposition 14 we saw that $(B4)$ fails even in metastable models, and at limit as well as at successor stages. For the other schemata in our list $(B1)$—$(B4)$ the situation is somewhat different. $(B1)$, for instance, can fail, in models with transitive alternativeness relation, at successor stages but not at limit stages. This fact is contained in Proposition 16.

PROPOSITION 16. (i) *Suppose* $\mathscr{M}$ *is a model such that* $R_{\mathscr{M}}$ *is transitive.* (a) *If* α *is any ordinal* $\geqslant 1$, *then for all* $w \in W_{\mathscr{M}}$ *every instance of* $(B2)$ *and* $(B3)$ *is true in* $\mathscr{M}^{\alpha}$ *at* w. (b) *If* α *is a limit ordinal then, moreover, every instance of* $(B1)$ *is true in* $\mathscr{M}^{\alpha}$ *at* w.

(ii) *Suppose that* $R_{\mathscr{M}}$ *is transitive and Euclidean and that* $\mathscr{M}^{\alpha}$ *is incoherent for all* α. *Then there are a sentence* φ *and a world* $w \in W_{\mathscr{M}}$ *such that for arbitrarily large successor ordinals* β, $B1(\varphi)$ *fails in* $\mathscr{M}^{\beta}$ *at* w.

Proof. (i.a) is trivial. To prove (i.b) let α be any limit ordinal λ. Suppose $[B(c_\varphi)]^{\lambda}_{\mathscr{M}, w} = 1$. By our revision rule for limit ordinals, $(\exists\gamma)(\forall\beta)(\gamma \leqslant \beta < \lambda \rightarrow [B(c_\varphi)]^{\beta}_{\mathscr{M}, w} = 1)$. So by our revision rule for successor ordinals, for each $\beta > \gamma$ and each $w' \in [wR]$, $[\varphi]^{\beta}_{\mathscr{M}, w'} = 1$. Suppose $w' \in [wR]$. Then by the transitivity of $R_{\mathscr{M}}$, for each $w'' \in [w'R]$ $[\varphi]^{\beta}_{\mathscr{M}, w''} = 1$. So $\varphi \in [B]^{\beta}_{\mathscr{M}, w'}$ and so $[B(c_\varphi)]^{\beta}_{\mathscr{M}, w'} = 1$. So, putting $\psi = B(c_\varphi)$, we have that $(\forall w' \in [wR])[B(c_\psi)]^{\beta}_{\mathscr{M}, w'} = 1$ for

each successor ordinal β such that $\gamma < \beta < \lambda$. Let δ be such that $\lambda = \gamma + \delta$. It follows by straightforward induction on δ that $(\forall\beta)(\gamma \leqslant \beta \leqslant \lambda \rightarrow [\![B(c_\psi)]\!]^{\beta}_{\mathscr{M}, w} = 1)$. So in particular $[\![B(c_\psi)]\!]^{\lambda}_{\mathscr{M}, w} = 1$.

(ii) is proved in the same way as Proposition 14. □

A fact closely related to those recorded in Proposition 14 and Proposition 16.ii is that these are natural classes C of ordinals such that the set of sentences true in all models at all worlds at all members of C is not closed under Principle $(B2')$. The next Proposition makes this explicit.

PROPOSITION 17. *Suppose that $\mathscr{M}$ is a model such that $R_{\mathscr{M}}$ is transitive and Euclidean, and that for all α $\mathscr{M}^{\alpha}$ is incoherent. Then there is an instance ψ of $(B1)$ and a $w \in W_{\mathscr{M}}$ such that for arbitrarily large ordinals β $[\![B(c_\psi)]\!]_{\mathscr{M}^{\beta}, w} = 0$; moreover, if β is a perfect stabilization ordinal then $[\![B(c_\psi)]\!]_{\mathscr{M}^{\beta}, w} = 0$ while $[\![\psi]\!]_{\mathscr{M}^{\beta}, w} = 1$.*

Proof. As in the proof of Proposition 14 we can assume that there exist a sentence φ and a world w such that $w \in \mathscr{R}an(R_{\mathscr{M}})$ and that at arbitrarily large ordinals α $B(c_\varphi)$ is false at w in $\mathscr{M}^{\alpha}$ while φ is true in $\mathscr{M}^{\alpha}$ at all $w' \in [wR]$. Then $B(c_\varphi)$ is true at w in $\mathscr{M}^{\alpha+1}$. Observe that the assumption that $w \in \mathscr{R}an(R_{\mathscr{M}})$ entails that $w \in [wR]$. So, if $\theta = B(c_\varphi)$, $B(c_\theta)$ is false at w in $\mathscr{M}^{\alpha+1}$. So $B(c_\varphi) \rightarrow B(c_\theta)$ is false at w in $\mathscr{M}^{\alpha+1}$. So, taking ψ to be the sentence $B(c_\varphi) \rightarrow B(c_\theta)$, $B(c_\psi)$ is false at w in $\mathscr{M}^{\alpha+2}$.

As before, since this holds for arbitrarily large ordinals and the sentence is of the form $B(\delta)$ for some δ, it will fail in $\mathscr{M}^{\alpha}$ at w in particular for all perfect stabilization ordinals α. We already saw in Proposition 16 that for such α ψ itself is true in $\mathscr{M}^{\alpha}$ at w. □

Besides the specific information these last three propositions give about the behavior of the schemata $(B1)$—$(B4)$ and a couple of close variants, they also prove the general point that when validity is defined as truth throughout the class of models $\mathscr{B}$, the set of schemata that are thereby singled out as valid will depend not just on the constraints that are imposed on R but also on the types of revision stages that are admitted in $\mathscr{B}$. Indeed, summarizing what we have found above, and adding a few further facts of the same kind:

Suppose $\mathscr{B}$ is the class of all model structures M such that R_M is transitive and Euclidean, $\mathscr{B}_0$ is the class of all expansions of model

structures in $\mathscr{B}$, $\mathscr{B}_1$ is the class of all models of the form $\mathscr{M}^\alpha$ where $\mathscr{M} \in \mathscr{B}_0$ and $\alpha \geqslant 1$, $\mathscr{B}_2$ is the class of all models of the form $\mathscr{M}^\alpha$ where $\mathscr{M} \in \mathscr{B}_0$ and α is a limit ordinal, $\mathscr{B}_3$ is the class of all models of the form $\mathscr{M}^\alpha$ where $\mathscr{M} \in \mathscr{B}_0$ and α is a perfect stabilization ordinal for $\mathscr{M}$, and $\mathscr{B}_4$ is the class of all models of the form $\mathscr{M}^\alpha$ where $\mathscr{M} \in \mathscr{B}_0$ and $\mathscr{M}^\alpha$ is meta-stable. Let for $i = 0, \ldots, 4$ Val_i be the set of all sentences of L which are true in all members of $\mathscr{B}_i$ at each of their worlds. We say that a schema from the set $\{(B1)$—$(B4)\}$ is *validated by* $\mathscr{B}_i$ if all its instances belong to Val_i. Then out of these four schemata, $\mathscr{B}_0$ validates none; $\mathscr{B}_1$ validates $\{B2, B3\}$; $\mathscr{B}_2$ validates $\{B1, B2, B3\}$; $\mathscr{B}_3$ validates $\{B1, B2, B3\}$; and $\mathscr{B}_4$ validates $\{B2, B3\}$.

The $\mathscr{B}_i$ form only a small selection from a much larger family of classes all of which have some prima facie plausibility as bases for a definition of validity. The extensive experience with possible worlds semantics for modal logics has demonstrated that there is in general no hope of finding intuitive criteria that narrow such families down to a single class which yields "the correct" definition of validity. In particular, it is usually unfeasible to come up with non-circular justifications for the conditions on the alternativeness relation R, in terms of which these classes have often been defined. In relation to the present semantics this problem is amplified, for the prima facie plausible classes may vary not only with respect to the conditions which they impose on R, but also with respect to the types of model revision which they admit.

With respect to this second dimension of variation there appear to be some natural guidelines for what should be admitted and what not. In a model where some sentences that would stabilize upon revision are nevertheless unstable, $[\![B]\!]$ is an unnecessarily flawed record of what is believed (according to the relation R and the definition of truth); it seems reasonable to see this removable defect as disqualifying the model, and to ignore such models in definitions of validity.

This limits the family of relevant classes to those which contain only metastable models. But that still leaves room for further restrictions — e.g. to the semistable models, or to models which represent revision stages of certain ordinal types, such as, say, successor stages or limit stages. We have just seen that it can make a difference to validity whether or not such further restrictions are imposed. But we do not know of any compelling reasons that speak either for or against them.

To those who see the central task of doxastic logic as that of determining the one "true" logic of belief, this multiplicity of possibilities

poses a dilemma. It is a dilemma for which we have no solution. In fact, we suspect that no solution exists. If one wants to settle for some particular system of doxastic logic, then this will have to be, to some extent, a matter of decision. It cannot be a matter of discovery alone.

But it ought to be a matter of *informed* decision. In particular, the decision should be informed by a proper perspective of the options from which the choice is made, and such understanding can be gained only by patiently exploring the entire field of possibilities. However, a thorough exploration of that field would be a task of immense proportions. The options we have so far considered cover no more than a small corner of it. In the next and final section we consider a few others. But that will still leave an unknown territory of which we cannot even guess the true dimensions.

II.2.3.

Gupta (1982) and Belnap (1982) note a curious consequence of Herzberger's rule for limit stages, which we incorporated into the model theory developed in II.1.2. The phenomenon to which they draw attention manifests itself within our framework as follows. Let $\mathscr{M}$ be a model such that $R_{\mathscr{M}}$ is transitive and there is at least one world w in $W_{\mathscr{M}}$ such that $[wR] \neq \emptyset$ and $(\forall w' \in [wR])[w'R] \neq \emptyset$. Suppose moreover that $[\![b]\!]_{\mathscr{M}} = \neg B(b)$ and $[\![c]\!]_{\mathscr{M}} = \neg B(c)$. Then we find a curious asymmetry in the behavior of certain Boolean compounds of $\boldsymbol{b}$ and $\boldsymbol{c}$. For instance the sentences $B(b) \vee \neg B(c)$ and $\neg B(b) \vee B(c)$ stabilize in $\mathscr{M}$ at w, whereas $B(b) \vee B(c)$ and $\neg B(b) \vee \neg B(c)$ do not. This does seem odd indeed. For it might well be that from an intuitive point of view $\boldsymbol{b}$ and $\boldsymbol{c}$ have nothing to do with each other,[40] in which case it is very hard to see what could be responsible for the stability of, say, $B(b) \vee B(c)$ that would not equally apply to $B(b) \vee \neg B(c)$. It is easy to verify these facts, and in doing so one realizes that the discrepancy arises because of Herzberger's revision rule for limit stages. We are not sure that this is in itself sufficient reason to reject Herzberger's rule. But the phenomenon seems incongruous enough to justify the search for an alternative.

There exists, we believe, a consensus that every revision rule for limit stages must obey the following constraint. Suppose that $\mathscr{M}$ is a model and that $\mathscr{M}^\beta$ has been defined for all $\beta < \lambda$. Then we can, for any $w \in W_{\mathscr{M}}$, divide the sentences of L into three categories:

(LS.i) sentences φ for which there is a $\gamma < \lambda$ such that $(\forall\beta)(\gamma < \beta < \lambda \rightarrow \varphi \in [\![B]\!]^{\beta}_{\mathcal{M}, w})$;

(LS.ii) sentences φ for which there is a $\gamma < \lambda$ such that $(\forall\beta)(\gamma < \beta < \lambda \rightarrow \varphi \notin [\![B]\!]^{\beta}_{\mathcal{M}, w})$;

(LS.iii) those sentences φ to which neither (i) nor (ii) applies.

From the perspective of the limit of the unbounded sequence of revisions $\{\mathcal{M}^{\beta}\}_{\beta < \lambda}$, the sentences of the first type appear as positively stable at w, those of the second type as negatively stable at w. Accordingly any revision rule applying at stage λ should assign to B an extension at w that contains the former sentences and excludes the latter.

We will refer to this criterion as the *local stability principle*. Besides this principle there do not appear to be any other clear constraints that an alternative to Herzberger's rule should satisfy. Indeed, Belnap's (1982) approach to the treatment of limit stages is premised on the conviction that there are none. Herzberger's rule itself can be described as the one which makes $[\![B]\!]^{\lambda}_{\mathcal{M}, w}$ as small as the local stability principle permits. Gupta (1982) proposes a different rule, according to which the sentences of type (LS.iii) are put into $[\![B]\!]^{\lambda}_{\mathcal{M}, w}$ depending on whether or not they belong to $[\![B]\!]^{0}_{\mathcal{M}, w}$. This rule, which might be said to be equipped with the memory of an elephant, has consequences that strike us as much more unpalatable than any of the apparent oddities associated with Herzberger's. For instance, if we define validity as truth at all worlds in all metastable expansions of some natural class of model structures (e.g. the class of all model structures in which R is transitive and Euclidean) then no schema that isn't a theorem of truth-functional logic will come out as valid. The reason for this is that if a schema μ does not have the form of a tautology we can always interpret its sentence letters as sentences of L that are unstable at w in some model $\mathcal{M}$ and choose $[\![B]\!]^{0}_{\mathcal{M}, w}$ in such a way that μ is refuted in $\mathcal{M}$. It is easily seen that in such a situation there will always be arbitrarily large limit ordinals λ from which each of the relevant sentences is locally unstable at w. Gupta's rule guarantees that at each such λ $[\![B]\!]^{\lambda}_{\mathcal{M}, w}$ agrees on all the sentences that interpret sentence letters occurring in μ. Consequently μ is refuted at w in $\mathcal{M}^{\lambda}$. (The details of this argument will become explicit in the proof of Proposition 18 below.)[41]

This last observation reveals that there is yet another dimension to the question which models should be considered relevant to the definition of validity. This is a dimension that we ignored in our discussion of

the question in II.2.1; we never raised, when contemplating which models $\mathcal{M}^{\alpha}$ should be included in the class $\mathcal{B}$, the question whether any restrictions should be placed on the "initial" intension of B, i.e. on $[\![B]\!]_{\mathcal{M}^0}$. We avoided the issue because at that point it would have been difficult to explain its relevance. Now, however, we can see that under certain conditions the initial intension of B will have a lasting effect that no amount of revision can undo, and that consequently restrictions on the initial intension may make a difference to the set of valid schemata even if the class $\mathcal{B}$ only contains metastable models.

Someone who would wish to define validity on the basis of a semantics in which Herzberger's rule has been replaced by Gupta's, would, on pain of ending up with an essentially vacuous doxastic logic, have to impose restrictions on initial intensions. But what could those restrictions be? Evidently if the restrictions are to save the emerging logic from triviality, the initial extensions must validate for any schema that is to qualify as valid at least all interpretations that involve unstable sentences. We do not know of many ways in which such restrictions can be expressed in non-question begging terms. There is, as far as we can see, only one constraint that can be stated without circularity and which guarantees that some non-tautological schemata stand a chance of qualifying as valid. This is the condition which demands that the initial extensions be all empty. Herzberger calls extensional models which satisfy this condition *primary*. We adopt this term also, and call a model $\mathcal{M}$ *primary* iff $(\forall w \in W_{\mathcal{M}})[\![B]\!]_{\mathcal{M},w} = \emptyset$. Metastable revisions of primary models validate certain non-tautologous schemata even when Gupta's rule is used instead of Herzberger's. However, this is hardly exciting news, for in relation to primary models the two limit rules produce exactly the same revision sequences. This entails in particular that for models of this kind Gupta's rule will produce the same oddities that were noted in connection with Herzberger's rule, so that even from that perspective it does not constitute an improvement.[42]

It may be that the local stability principle is the only intuitively justifiable constraint on limit rules. If this is so the definition of logical validity ought to be insensitive to which of those rules is being used. This is the position of Belnap (1982), who proposes a revision policy according to which we may at each limit stage choose any model that can be obtained from the sequence of preceding stages in a way compatible with the local stability principle. There are several ways in which this idea can be made precise. Here we present one.

The central notion in Belnap's proposal is what he calls a "bootstrap-

ping policy". We will use the term "revision scheme" instead. By an *interpolation function on* a set A we understand any function f from $\mathcal{P}(A)^2$ into $\mathcal{P}(A)$ such that whenever $A_1, A_2 \subseteq A$ and $A_1 \cap A_2 = \emptyset$ then $f(A_1, A_2) \supseteq A_1$ and $f(A_1, A_2) \cap A_2 = \emptyset$. By a *revision scheme* understand a function $\mathcal{R}$ defined on the class of all limit ordinals such that for each λ $\mathcal{R}(\lambda)$ is an interpolation function on the set S_L of sentences of L. Given a model $\mathcal{M}$ and a revision scheme $\mathcal{R}$, the *revision sequence starting from* $\mathcal{M}$ *according to* $\mathcal{R}$ is the sequence $\{\mathcal{M}^{\alpha,\mathcal{R}}\}_{\alpha \in On}$, defined by: $[\![B]\!]^{0,\mathcal{R}}_{\mathcal{M},w} = [\![B]\!]_{\mathcal{M},w}$; $[\![B]\!]^{\alpha+1,\mathcal{R}}_{\mathcal{M},w}$ is defined as in Section II.1.2; and $[\![B]\!]^{\lambda,\mathcal{R}}_{\mathcal{M},w} = \mathcal{R}(\lambda)(B^+_w, B^-_w)$, where $B^+_w = \{\varphi : (\exists\gamma < \lambda)(\forall\beta)(\gamma < \beta < \lambda \rightarrow \varphi \in [\![B]\!]^{\beta,\mathcal{R}}_{\mathcal{M},w})\}$ and $B^-_w = \{\varphi : (\exists\gamma < \lambda)(\forall\beta)(\gamma < \beta < \lambda \rightarrow \varphi \notin [\![B]\!]^{\beta,\mathcal{R}}_{\mathcal{M},w})\}$. With respect to any sequence $\{\mathcal{M}^{\alpha,\mathcal{R}}\}_{\alpha \in On}$, we can distinguish between those sentences that stabilize at w and those that do not, and similarly for the various other notions relating to stability that were introduced earlier. We can in particular distinguish between those $\mathcal{M}^{\alpha,\mathcal{R}}$ in which every sentence that stabilizes at any world is stable at that world, and those for which this is not so. We refer to the former again as *metastable* models. Among the metastable $\mathcal{M}^{\alpha,\mathcal{R}}$ there will be some such that $[\![B]\!]^{\alpha}_{\mathcal{M}} = [\![B]\!]^{\beta}_{\mathcal{M}}$ for arbitrarily large β. A model of this last kind will be called *recurrent*. When λ is a limit ordinal, $w \in W_{\mathcal{M}}$ and $(\exists\gamma < \lambda)(\forall\beta)(\gamma \leqslant \beta < \lambda \rightarrow [\![\varphi]\!]_{\mathcal{M}^{\beta},w} = 1)$ we say that φ is *locally stably true in* $\mathcal{M}$ *at* w *f(rom the) p(erspective) o(f)* λ *according to* $\mathcal{R}$. Similarly, if $(\exists\gamma < \lambda)(\forall\beta)(\gamma \leqslant \beta < \lambda \rightarrow [\![\varphi]\!]_{\mathcal{M}^{\beta},w} = 0)$ φ is *locally stably false in* $\mathcal{M}$ *at* w *fpo* λ *according to* $\mathcal{R}$; and if

$$(\forall\gamma < \lambda)((\exists\beta)(\gamma \leqslant \beta < \lambda \,\&\, [\![\varphi]\!]_{\mathcal{M}^{\beta},w} = 1) \,\&\, (\exists\beta)(\gamma \leqslant \beta < \lambda \,\&\, [\![\varphi]\!]_{\mathcal{M}^{\beta},w} = 0))$$

we say that φ is *locally unstable in* $\mathcal{M}$ *at* w *fpo* λ *according to* $\mathcal{R}$.

Once again, this revision concept yields a considerable variety of classes of models in terms of which validity might be defined. Although no clear conceptual criteria seem available for choosing from among these classes, here too those consisting only of metastable models would appear to be among the most natural candidates. In relation to the present revision concept, however, the choice turns out to be of little material import. For nearly every intuitively reasonable choice results in a logic that contains virtually no non-tautologous schemata at all. Proposition 18, though it does not exhaust the entire spectrum of relevant possibilities, establishes this fact for a representative family of such classes.

Let T be the system of modal propositional logic, formulated in the language PL, whose axioms are the instances of all truthfunctional tautologies and which is closed under the following three inference rules:

M.P.	R1	R2
$\vdash \varphi, \vdash \varphi \rightarrow \psi$	$\vdash \varphi$	$\vdash \neg\varphi$
$\vdash \psi$	$\vdash B\varphi$	$\vdash \neg B\varphi$

PROPOSITION 18. *Let $\{c_i\}_{i \in \omega}$ be a denumerable coinfinite subset of $C_L - C_\varphi$, where C_φ is the range of the function c_φ. Let M be an extensional model structure such that for each $i \in \omega$ $[\![c_i]\!]_M = \neg B(c_i)$, and let $\mathscr{B}$ be the class of models consisting of all metastable expansions $\mathscr{M}^{\alpha, \mathscr{R}}$ of M, for arbitrary revision schemes $\mathscr{R}$. Then for every schema μ of PL, $T \vdash \mu$ iff $\mathscr{B} \models \mu$.*

Proof. The proof from left to right proceeds by a straightforward induction on the length of proofs in T. The proof from right to left rests on the following idea. Let I_0 be the interpretation which assigns to each sentence letter p_i the sentence $\neg B(c_i)$. Then, whenever μ is a schema that is not derivable in T, there will be a revision scheme $\mathscr{R}$ and an expansion $\mathscr{M}$ of M such that $I_0(\mu)$ is false in $\mathscr{M}^{\alpha, \mathscr{R}}$ for arbitrarily large α. This is so because for any schema μ that is "contingent in" T (i.e. neither provable nor disprovable in T) we can construct an expansion $\mathscr{M}^{\lambda, \mathscr{R}}$ of M for arbitrarily large limit ordinals λ so that μ is locally unstable in $\mathscr{M}$ fpo λ according to $\mathscr{R}$.

To prove the right-to-left direction in detail we proceed as follows. By the *degree* of a formula μ of PL, $deg(\mu)$, we understand the maximum of the lengths of chains of nested occurrences of B in μ. For ease of notation we will write $\boldsymbol{\mu}$ instead of $I_0(\mu)$; where A is a set of formulae of PL, $\boldsymbol{A}$ is the set of all $\boldsymbol{\mu}$ such that $\mu \in A$.

With every formula μ of PL we associate what we call a *decoration of* μ, d_μ. d_μ is a function from a certain subset of the set of subformulae $B\nu$ of μ to the set $\{\mathrm{T}, \bot\}$. This function is determined as follows. We first consider the subformulae $B\nu$ of μ such that ν does not contain any occurrences of B. For each of these subformulae we put $d_\mu(B\nu) = \mathrm{T}$ if ν is a tautology (of ordinary propositional logic), and put $d_\mu(B\nu) = \bot$ if ν is the negation of a tautology. For the remaining formulae $B\nu$ with B not occurring in ν d_μ is undefined. We now look at

the subformulae $B\nu$ where ν contains only unnested occurrences of B. Let ν' be the formula obtained from ν by replacing each subformula $B\beta$ of ν for which d_μ is defined by $d_\mu(B\beta)$ and replacing the remaining subformulae $B\beta_1, \ldots, B\beta_n$ of ν by distinct sentence letters $q_1, \ldots, q_n$ which do not occur in ν. If ν' is a tautology then we put $d_\mu(B\nu) = T$, if ν' is the negation of a tautology, we put $d_\mu(B\nu) = \bot$, and otherwise $d_\mu(B\nu)$ remains undefined. We then look at the subformulae $B\nu$ such that ν contains only subformulae $B\beta$ of degree 2, repeat the same procedure, etc.

It is easy to establish the following facts concerning d_μ:

(1) (i) if $d_\mu(B\nu) = \mathrm{T}$ then $T \vdash B\nu$.

(ii) if $d_\mu(B\nu) = \bot$ then $T \vdash \neg B\nu$.

(2) Let I be any interpretation and let $\mathcal{M}$ be a member of $\mathcal{B}$. Then, if $d_\mu(B\nu) = \mathrm{T}$, $[\![I(B\nu)]\!]_{\mathcal{M}} = 1$,[43] and if $d_\mu(B\nu) = \bot$, $[\![I(B\nu)]\!]_{\mathcal{M}} = 0$.

(3) Let I and $\mathcal{M}$ be as under (2). Suppose μ' is obtained from μ in the same way as the formulae ν' were obtained from the subformulae ν of μ. Suppose that V is an assignment of truth values to the sentence letters of μ' and that for each Boolean constituent γ of μ which is not in the domain of d_μ $[\![I(\gamma)]\!]_{\mathcal{M}} = 1$ iff V assigns T to the sentence letter q of μ' corresponding to γ. Then $[\![I(\mu)]\!]_{\mathcal{M}} = V(\mu')$.

(1) and (2) are both proved by induction on the complexity of ν. (3) is an immediate consequence of (2).

We call a formula of *PL prime* if it is either T, $\bot$, a sentence letter or a formula of the form $B\nu$. Note that if a formula ν of *PL* is prime and different from T and $\bot$ then ν is always either of the form $B(c)$ or of the form $\neg B(c)$ for some constant c.

Let μ be a formula of *PL* such that $\neg T \nvdash \mu$. Assume that $\{p_1, \ldots, p_k\}$ includes all the sentence letters that occur in μ. For $n \leqslant \deg(\mu)$ let A_n be the set of all prime formulae of *PL* of degree $\leqslant n$ that are subformulae of μ and are not in the domain of d_μ. We will show, by induction on n, that the following is true for $n \leqslant \text{degree}(\mu)$:

(*) There exist

(i) a function r_n defined on A_n which maps each $\varphi \in A_n$ to a partial function from the class Lim of all limit ordinals to $\{0, 1\}$,

(ii) a closed unbounded subsequence $\{\alpha^n_\beta\}_{\beta \in On}$ of Lim, and
(iii) for each subset S of A_n an unbounded subsequence Γ_S of $\{\alpha^n_\beta\}_{\beta \in On}$, consisting exclusively of ordinals that are successor ordinals in $\{\alpha^n_\beta\}_{\beta \in On}$,

such that if $\mathscr{M}$ is any expansion of M and $\mathscr{R}$ any revision scheme that is *compatible with* r_n,[44] then

(a) for every $\varphi \in A_n$ $\mathrm{Dom}(r_n(\varphi)) \cap \{\alpha^n_\beta\}_{\beta \in On} = \Gamma_{S1} \cup \ldots \cup \Gamma_{Sv}$, where $S1, \ldots, Sv$ are all the subsets of A_n.
(b) if $S \neq S'$ then $\Gamma_S \cap \Gamma_{S'} = \emptyset$
(c) for every ordinal β and every $\varphi \in A_n$ φ is locally unstable in $\mathscr{M}$ fpo α^n_β according to $\mathscr{R}$, and
(d) For every $\gamma \in \Gamma_S$, $[\![B(c_\varphi)]\!]_{\mathscr{M}^\gamma, \mathscr{R}} = 1$ iff $\varphi \in S$.[45]

The role of the functions r_n is to impose increasingly strong constraints on the revision schemes $\mathscr{R}$ that are to be considered. The significance of r_n for $\mathscr{R}$ is that it tells us for those combinations of φ and α for which it is defined that if φ is locally unstable in $\mathscr{M}$ fpo α according to $\mathscr{R}$ then $[\![B(c_\varphi)]\!]_{\mathscr{M}^\alpha, \mathscr{R}} = r_n(\varphi)(\alpha)$. The following definition guarantees that this is indeed the effect which compatibility with r_n has: let f be any partial function from the set of sentences of L to partial functions from Lim to $\{0, 1\}$, and let $\mathscr{R}$ be a revision scheme. Then $\mathscr{R}$ is *compatible with* f iff for each φ and α such that $f(\varphi)(\alpha)$ is defined, and each pair B^+, B^- of disjoint subsets of S_L such that $\varphi \notin B^+ \cup B^-$, $\varphi \in \mathscr{R}(\alpha)(B^+, B^-)$ iff $f(\varphi)(\alpha) = 1$.

It is easy to see that (*) gives us the desired result. For let $B\beta_1, \ldots, B\beta_n$ be the Boolean constituents of μ that begin with B. Since not $T \vdash \mu$, it follows from (1) that μ' (the result of replacing the $B\beta_i$ for which d_μ is defined by $d_\mu(B\beta_i)$, and the remaining $B\beta_i$ by new sentence letters q_i) is not a tautology. So there is an assignment V of truth values to the sentence letters of μ' such that $V(\mu') = 0$. Let S be a subset of $A_{\deg(\mu)}$ such that for each Boolean constituent $B\beta$ of μ such that $d_\mu(B\beta)$ is undefined, $B\beta \in S$ iff $V(B\beta) = 1$. Then it follows from (*.iii.d) and (3) that if, for $n = \deg(\mu)$, $\mathscr{R}$, $\mathscr{M}$ and Γ_S satisfy (*) and $\gamma \in \Gamma_S$ then $[\![B(\mu)]\!]_{\mathscr{M}^\gamma, \mathscr{R}} = 0$.

To prove (*) let $\mathscr{M}$ be any expansion of M. First suppose that $n = 0$. A_0 consists just of the sentences $\neg B(c_i)$ such that p_i occurs in μ. So every member of A_0 will be locally unstable in $\mathscr{M}$ from any limit ordinal according to any revision scheme whatever. Let $\{\alpha^0_\beta\}_{\beta \in On}$ be

the sequence of all limit ordinals. Assign to the subsets $S1, \ldots, Sm$ of A_0 disjoint unbounded sequences $\Gamma_{S1}, \ldots, \Gamma_{Sm}$ of limit ordinals, and let r_0 be the function defined on A_0 such that for each $\varphi \in A_0$ $r_0(\varphi)$ is the function: $\Gamma_{S1} \cup \ldots \cup \Gamma_{Sm} \rightarrow \{0, 1\}$ such that for $\alpha \in \Gamma_{S1} \cup \ldots \cup \Gamma_{Sm}$ $r_0(\neg B(c_i))(\alpha) = 0$ iff $\alpha \in \Gamma_{Si}$. It is not hard to see that if $\mathscr{R}$ is compatible with r_0, then for every $\gamma \in \Gamma_{Si}$ and $\varphi \in A_0$ $[\![B(c_\varphi)]\!]_{\mathscr{M}^\gamma, \mathscr{R}} = 1$ iff $\varphi \in Si$. Note in this connection that because of the local instability of the sentences c_i neither of the sets $\{\varphi : (\exists \gamma < \lambda)(\forall \beta)(\gamma < \beta < \lambda \rightarrow \varphi \in [\![B]\!]^\beta_{\mathscr{M}\mathscr{R}})\}$ and $\{\varphi : (\exists \gamma < \lambda)(\forall \beta)(\gamma < \beta < \lambda \rightarrow \varphi \notin [\![B]\!]^\beta_{\mathscr{M}\mathscr{R}})\}$ will contain c_i for any limit ordinal λ. So the compatibility of $\mathscr{R}$ with r_0 guarantees that the sentences $B(c_\varphi)$ have the required truth values.

Now suppose we have associated with A_{n-1} a function r_{n-1}, a closed unbounded sequence $\{\alpha^{n-1}_\beta\}_{\beta \in On}$ of limit ordinals and unbounded subsequences Γ_S of $\{\alpha^{n-1}_\beta\}_{\beta \in On}$ for all subsets S of A_{n-1} such that (*) is satisfied for $n - 1$. Let $\boldsymbol{B\nu}_1, \ldots, \boldsymbol{B\nu}_p$ be all the members of $A_n - A_{n-1}$. For each $j \leqslant p$ there is a B-free formula ν''_j of PL such that (i) $\boldsymbol{\nu}_j$ is the result of substituting sentences $\boldsymbol{B\beta}_1, \ldots, \boldsymbol{B\beta}_t$ for the sentence letters q_i in ν''_j, and (ii) ν'_j is obtained from ν''_j by substituting $d_\mu(B\beta_i)$ for q_i whenever $d_\mu(B\beta_i)$ is defined. For ease of notation let us assume that d_μ is defined for $\boldsymbol{B\beta}_1, \ldots, \boldsymbol{B\beta}_s$ and not for the remaining formulae $\boldsymbol{B\beta}_{s+1}, \ldots, \boldsymbol{B\beta}_t$. Since $B\nu_j$ is not in the domain of d_μ, ν'_j is contingent, and so there are valuations V_{j1} and V_{j2} on $\{q_1, \ldots, q_s\}$ which make ν'_j true and false, respectively. Let S_{j1} be a subset of A_{n-1} such that for $i = 1, \ldots, s$ $\psi_i \in S_{j1}$ iff $V_{j1}(q_i) = 1$ and let S_{j2} be a subset of A_{n-1} that is similarly related to V_{j2}. By assumption there are unbounded sequences Γ_{j1} and Γ_{j2}, each consisting of successor ordinals in the sequence $\{\alpha^{n-1}_\beta\}_{\beta \in On}$, so that for all $\gamma \in \Gamma_{j1}$ and $\psi \in A_{n-1}$ $r_{n-1}(\psi)(\gamma) = 1$ iff $\psi \in S_{j1}$, and likewise for Γ_{j2} and S_{j2}. In this way we can associate with each of the sentences $\varphi_j (j = 1, \ldots, p)$ a pair of ordinal sequences Γ_{j1}, Γ_{j2}. Using these we can construct for any ordinal δ an ω-sequence $\{\gamma_i\}_{i \in \omega}$ of ordinals $> \delta$ all of which are members of $\mathrm{Dom}(r_{n-1}(\varphi)) \cap \{\alpha^{n-1}_\beta\}_{\beta \in On}$ such that $\{\gamma_i\}_{i \in \omega}$ has an infinite intersection with each of the sets Γ_{ju}, for $j = 1, \ldots, p$; $u = 1, 2$. Let γ be the limit of $\{\gamma_i\}_{i \in \omega}$. It is easily verified from the induction hypothesis that if $\mathscr{R}$ is a revision scheme compatible with r_{n-1} then for $j = 1, \ldots, p$ $\boldsymbol{\nu}_j$ will be locally unstable in $\mathscr{M}$ according to $\mathscr{R}$ fpo the limit of $\{\gamma_i\}_{i \in \omega}$. Since for every ordinal δ we can construct such a

sequence $\{\gamma_i\}_{i\in\omega}$ consisting entirely of ordinals $> \delta$, we can select an unbounded subsequence $\{\alpha'^n_\beta\}_{\beta\in On}$ of $\{\alpha^{n-1}_\beta\}_{\beta\in On}$ so that each α'^n_β is the limit of such a sequence $\{\gamma_i\}_{i\in\omega}$. It is easy to see that

(4) if $\mathscr{R}$ is compatible with r_{n-1}, then $\nu_1, \ldots, \nu_p$ are all locally unstable in $\mathscr{M}$ according to $\mathscr{R}$ fpo every α'^n_β.

Whenever a sentence φ is unstable in $\mathscr{M}$ according to $\mathscr{R}$ fpo all members of a sequence of ordinals that is cofinal with a limit ordinal λ then φ is also unstable in $\mathscr{M}$ according to $\mathscr{R}$ fpo λ. So the closure of $\{\alpha'^n_\beta\}_{\beta\in On}$, to which we will refer as $\{\alpha^n_\beta\}_{\beta\in On}$, also satisfies (4). Moreover, since $\{\alpha^{n-1}_\beta\}_{\beta\in On}$ is closed and $\{\alpha'^n_\beta\}_{\beta\in On} \subseteq \{\alpha^{n-1}_\beta\}_{\beta\in On}$, $\{\alpha^n_\beta\}_{\beta\in On} \subseteq \{\alpha^{n-1}_\beta\}_{\beta\in On}$. We now choose for all the subsets $S1, \ldots, Sv$ of A_n disjoint unbounded subsequences $\Gamma_{S1}, \ldots, \Gamma_{Sv}$ of ordinals that are successors in $\{\alpha^n_\beta\}_{\beta\in On}$ and define a function r' on A_n such that for each $\varphi \in A_n$ $r'(\varphi)$: $\Gamma_{S1} \cup \ldots \cup \Gamma_{Sv} \to \{0, 1\}$ is that function such that for all $i = 1, \ldots, v$, $\alpha \in \Gamma_{Si}$, $\varphi \in A_n$, $r'(\varphi)(\alpha) = 1$ iff $\varphi \in Si$. Let $r_n = r' \cup r_{n-1}$. Evidently if $\mathscr{R}$ is compatible with r_n then it is compatible with r_{n-1}. So it follows from the induction hypothesis that all members of A_n are locally unstable in $\mathscr{M}$ according to $\mathscr{R}$ fpo each ordinal α^n_β. This establishes (*iii.c). (*iii.d) then follows in view of the definition of r' and (3). (*iii.a) and (*iii.b) are direct consequences of the choice of the Γ_{Si} and the definition of r_n. □

An immediate corollary of the last proposition is

PROPOSITION 19. *Suppose that $\mathscr{B}'$ is a class of models which includes the class $\mathscr{B}$ of Proposition* 18. *Then the schemata that are valid in $\mathscr{B}'$ are included among the theorems of T.*

Proposition 19 pinpoints the predicament into which our search for a well-motivated and yet non-trivial doxastic logic seems to have landed us. The problematic status of particular revision rules for limit stages led us to try and eliminate the element of arbitrariness that they appear to introduce into the analysis of validity we have offered. The remedy we tried, first suggested by Belnap in relation to truth, was to include in the classes of models in terms of which validity is defined the results of using any one of the intuitively possible revision schemes (i.e. any of those schemes that obey the local stability principle.) But as a result we

find ourselves left with virtually no doxastic logic to speak of. The natural reaction to this is to wonder if there are perhaps not after all other admissibility criteria besides the local stability principle which narrow the class of revision schemes down, so that more schemata come out valid than are derivable within T.

As a matter of fact, Proposition 18 implies — and a careful look at its proof will confirm this — that some revision schemes are rather odd. Suppose that $\mu \leftrightarrow \mu'$ is a tautology of truthfunctional logic and that $I(\mu)$ and $I(\mu')$ are both locally unstable in $\mathcal{M}$ at all $w' \in [wR_{\mathcal{M}}]$ from the limit ordinal λ (according to, say, any $\mathcal{R}$ whatever). Then there will be revision schemes $\mathcal{R}$ which put $I(\mu)$ into $[\![B]\!]^{\lambda,\mathcal{R}}_{\mathcal{M},w}$ while leaving $I(\mu')$ out. One might well want to object to this, on the grounds that the equivalence between μ and μ' is so fundamental that any revision policy ought to treat them alike. Thus, a good revision scheme $\mathcal{R}$ should obey a second admissibility criterion, viz. that whenever $\varphi \leftrightarrow \varphi'$ instantiates a theorem of classical sentential logic, then for all λ, B^+, B^- $\mathcal{R}(\lambda)(B^+, B^-)$ contains either both of φ, φ' or neither. Let us call this the (*sentential*) *equivalence criterion.* If this criterion is adopted and conditions are otherwise as stated in Proposition 18, then the set of valid schemata is somewhat larger; it coincides with the set of theorems of the *PL*-theory T', which we obtain from T by adding for every tautology $\mu \leftrightarrow \mu'$ the formula $B\mu \leftrightarrow B\mu'$ as an axiom.

It is tempting to strengthen the sentential equivalence criterion in various ways. For instance, we could modify it so that it covers not only equivalences of sentential logic, but also those of quantification theory. This will further strengthen the set of valid sentences of L, although no difference arises at the level of propositional schemata. Another modification would be to demand for every schema μ that all its instances $I(\mu)$ and $I(\mu')$ are to be treated in the same way if $\mu \leftrightarrow \mu'$ is valid. It is not so obvious, however, how we can state this requirement without getting entangled in the circularity that is involved in the formulation we have just given. (Recall that the notion of validity itself depends on the constraints that we impose on $\mathcal{R}$!) One way, to be sure, in which we can avoid this particular circularity is by stipulating that the revision schemes must all be such that any model obtained by means of them must satisfy all instances of theorems of the theory T'', where T'' is the theory we obtain by adding to T the rule $\vdash \varphi \leftrightarrow \varphi' \Rightarrow \vdash B\varphi \leftrightarrow B\varphi'$. But if our goal is to arrive at a logic that is licensed by

independent semantic motivations then this will of course not do. An axiomatic characterization of the valid schemata should be *derivable* from the definition of validity (including the specification of the admissibility criteria for revision schemes); it should not be stipulated. By formulating the criterion in this way we merely replace one kind of circularity by another.

Another direction in which we might seek escape from the predicament Propositions 18 and 19 seem to present is by reasoning along the following lines. We argued that the motivation for the revision approach does not tell us much about what should be done at limit stages; it was this that suggested the Belnap treatment of those stages in lieu of the Herzberger rule. But this, one might say, is tantamount to admitting that limit stages should not really be regarded as "proper" stages. In particular, they should be excluded from a semantic characterization of validity. Suppose that, in the spirit of this reflection, we define validity in terms of classes of metastable models of the form $\mathscr{M}^{\beta+1}$. More formally, for any class $\mathscr{B}$ of models say that a schema μ is $valid_1(\mathscr{B})$ iff every instance of μ is true at every world in every model in $\mathscr{B}$ that is of the form $\mathscr{M}^{\beta+1,\mathscr{R}}$ for some $\mathscr{M}$, $\mathscr{R}$ and β. For validity$_1$ matters look a little better. In particular, schema $B3$ will now be valid, and if, as in Proposition 18, we restrict attention to extensional models the much stronger schema

$$(E) \qquad B\varphi \leftrightarrow \neg B \neg \varphi$$

is valid as well. We note that, on the other hand, (E) is not valid in the class of models that are of the form $\mathscr{M}^{\lambda,H}$, where λ is a limit ordinal and H is the Herzberger revision rule.

The thought behind this modification of our earlier definitions of validity (which included limit stages among the relevant models) may be pushed even farther. From the present perspective limit stages should be ignored since they permit exceedingly perverse choices for the extensions of B. This latitude does not prevent them from contributing to the purpose for which they were originally included, that of permitting the stabilization of sentences that need more than ω revisions to achieve stability. But it does entail that limit stages can, when the revision scheme is crazy enough, be distressingly untidy. Unfortunately, however, that untidiness is not limited to the limit stages themselves.

For instance, if $\mathcal{M}$ is extensional and φ and ψ are tautologically equivalent sentences such that $\varphi \in [B]_{\mathcal{M}^\lambda}$ and $\psi \notin [B]_{\mathcal{M}^\lambda}$, then for all natural numbers n $B^n\varphi \in [B]_{\mathcal{M}^{\lambda+n}}$ and $B^n\psi \notin [B]_{\mathcal{M}^{\lambda+n}}$.[46] Thus it looks as if the refuse of the limit stages spills arbitrarily far along the ω-sequences of successor stages which they initiate. This suggests an even stronger constraint on the models we want to consider, but one that is dependent on the particular sentence under consideration: say that μ is $valid_2(\mathcal{B})$ iff for each instance $\boldsymbol{\mu}$ of μ there is a natural number n such that $\boldsymbol{\mu}$ is true at every world in every model in $\mathcal{B}$ that is of the form $\mathcal{M}^{\lambda+m,\mathcal{R}}$ for some model $\mathcal{M}$, revision scheme $\mathcal{R}$, limit ordinal λ and natural number $m \geqslant n$. In other words, for μ to be $valid_2$ (with respect to $\mathcal{B}$) all that is required is that each of its instances stabilize to truth on every ω-sequence of revisions (that belong to $\mathcal{B}$).

Validity$_1$ and validity$_2$ do indeed differ from each other in the way we might have expected. For instance, in a model of the form $\mathcal{M}^{\lambda+1}$, where λ is a limit ordinal, (E) and $(B3)$ will both be valid; but the schemata which we obtain when we apply rule R_1 to them,

$$B(B\varphi \leftrightarrow \neg B\neg\varphi) \quad \text{and}$$
$$B((B\varphi \wedge B(\varphi \rightarrow \psi) \rightarrow B\psi),$$

are not. This means in particular that R_1 is not valid$_1$. On the other hand it is easily verified that R_1 is valid$_2$.

The ability which validity$_2$ has for neutralizing the chaotic effects that the Belnap approach has on the extensions of B at limit stages is reflected by the fact that the set of valid$_2$ schemata does not change when we restrict the class $\mathcal{B}$ of models, with respect to which validity is being defined, to those metastable stages that are obtained with the Herzberger rule. The next Proposition 21 makes this explicit, and its proof indicates how the difference between the Herzberger and Belnap approaches loses its impact on validity if limit stages are excluded in the manner of validity$_2$.

PROPOSITION 20. *Let T_2 be the theory obtained through adding to T the schemata $(B3)$ and (E) as axioms. Let $\mathcal{B}$ be the class defined in the statement of Proposition* 18. *Then the set of schemata that are $valid_2(\mathcal{B})$ coincides with the set of theorems of T_2. Moreover, if $\mathcal{B}'$ is*

the class of all extensional models that are metastable according to the definition of II.1 (*i.e. using the Herzberger rule for limit stages*) *then the set of schemata that are valid*$_2(\mathscr{B}')$ *also coincides with the set of* T_2 *theorems.*

Proof. It is straightforward to show that all theorems of T_2 are valid in the two senses Proposition 20 refers to.

To show that only the theorems of T_2 are valid in the relevant senses, we argue as follows. The presence of schema (E) in T_2 enables us to rewrite any schema μ as a Boolean combination of sentence letters and formulae of the form $B^n p_i$.[47] For further ease of notation we will write '$B^0 p_i$' for 'p_i'. Suppose that μ is a schema written in this form and that $T_2 \nvdash \mu$. We will find an instance $\boldsymbol{\mu}$ of μ such that

(1) for some model $\mathscr{M}$ in $\mathscr{B}'$ $[\![\boldsymbol{\mu}]\!]_{\mathscr{M}^{\omega+n}} = 0$ for arbitrarily large n.

Since the set of schemata that are valid$_2(\mathscr{B}')$ includes each of the other three classes mentioned in Proposition 20, this will establish the proposition.

Assume that the sentence letters occurring in μ are all among $p_1, \ldots, p_k$, and that $\deg(\mu) = m$. Then the Boolean constituents of μ are all among $B^0 p_1, \ldots, B^m p_1, B^0 p_2, \ldots, B^m p_2, \ldots, B^0 p_k, \ldots, B^m p_k$. Replacing these constituents by distinct sentence letters $p_1^0, \ldots, p_1^m, p_2^0, \ldots, p_2^m, \ldots, p_k^0, \ldots, p_k^m$, we obtain a B-free formula μ'. Since μ is not a theorem of T_2, there is an assignment A of truth values to the sentence letters of μ' such that $A(\mu') = 0$. Our task thus reduces to finding sentences $\varphi_1, \ldots, \varphi_k$ of L such that interpreting the sentence letters $p_1, \ldots, p_k$ by means of the $\varphi_1, \ldots, \varphi_k$ gives an instance $\boldsymbol{\mu}$ of μ for which (1) holds. To this end we make use of an observation in Herzberger (1982). Let $\mathscr{M}$ be an extensional model such that, for m given constants $c_0, \ldots, c_m$, $[\![c_i]\!]_{\mathscr{M}} = B(c_{i+1})$ for $i = 0, \ldots, m-1$, and $[\![c_m]\!]_{\mathscr{M}} = \neg B(c_0)$. Moreover, we assume that these constants are not involved in any other self-referential links in $\mathscr{M}$; i.e. if we restrict $<_{\mathscr{M}}$ to the set $\{c : (\exists i < m) c <_{\mathscr{M}} c_i\}$ and then subtract the pairs $\langle c_i, c_{i+1}\rangle$, for $i = 0, \ldots, m-1$, and $\langle c_m, c_0\rangle$ from this set, then the resulting relation is well-founded. Under revision the truth values of the c_i in $\mathscr{M}$ go through the following cyclic pattern of periodicity $2 \cdot (m+1)$, of which we display the segment beginning with stage ω:

Stage	c_0	c_1	...	c_{m-2}	c_{m-1}	c_m
ω	F	F		F	F	T
$\omega+1$	F	F		F	T	T
,	,	,		,	,	,
,	,	,		,	,	,
$\omega+m$	T	T		T	T	T
$\omega+m+1$	T	T		T	T	F
$\omega+m+2$	T	T		T	F	F
,	,	,		,	,	,
,	,	,		,	,	,
$\omega+2m+1$	F	F		F	F	F
$\omega+2m+2$	F	F		F	F	T

For $r = 0, \ldots, m$ let φ^r be the conjunction $\alpha_1^r \,\&\, \ldots \,\&\, \alpha_m^r$, where for $j = 0, \ldots, m$ $\alpha_j^r = \boldsymbol{c}_j^r$ if the intersection of the j-th column and the r-th row of the above array contains a T, and $\alpha_j^r = \neg \boldsymbol{c}_j^r$ if the intersection contains an F. Let for $i = 1, \ldots, k$, $\varphi_i = \mathrm{V}_{r \in R} \varphi^r$, where $R = \{r : 0 \leqslant r \leqslant m \,\&\, A(p_i^{(m-r)}) = 1\}$. Since all the $\boldsymbol{c}_i$ are locally unstable fpo ω, for all $i \leqslant k$ $\boldsymbol{c}_i \notin [\![B]\!]_{\mathscr{M}\omega}$. It is easy to verify that for $i = 1, \ldots, k$, $r = 0, \ldots, m$: $[\![B^r \varphi_i]\!]_{\mathscr{M}\omega+m,\mathscr{A}} = [\![\varphi_i]\!]_{\mathscr{M}\omega+m-r,\mathscr{A}} = A(p_i^r)$. So if $\boldsymbol{\mu}$ is the result of substituting the φ_i for the p_i in μ, then $[\![\boldsymbol{\mu}]\!]_{\mathscr{M}\omega+m,\mathscr{A}} = 0$. Because of the cyclical character of the $\boldsymbol{c}_i$ we can infer that $[\![\boldsymbol{\mu}]\!]_{\mathscr{M}\omega+2h\cdot(m+1),\mathscr{A}} = 0$ for arbitrarily large h. □

When we also admit non-extensional models within the class we use to define validity then schema (E) will clearly no longer count as valid. However, if we restrict attention to successor stages, validity remains for the schema $(B3)$. In fact, we can prove the following:

PROPOSITION 21. *Let $\mathscr{B}''$ be the class of all metastable models resulting from the expansion of any model structure M using the Belnap approach for revision at limit stages. Let T_1 be the theory obtained by removing from the theory T of Proposition* 18 *the rule R_2, and adding the schemata $(B3)$ and $B(\varphi \,\&\, \neg\varphi) \rightarrow B\psi$ as axioms. Then the set of all schemata that are valid$_2(\mathscr{B}'')$ coincides with the set of theorems of the theory T_1.*

Like Proposition 20, Proposition 21 is not affected when we restrict attention to the Herzberger revision scheme. Moreover, imposing restrictions on the alternativeness relation $\mathscr{R}$ has comparatively little effect. If we restrict attention to models in which $\mathscr{R}$ is serial then the axiom $B(\varphi \wedge \neg\varphi) \rightarrow B\psi$ must be strengthened to $\neg B(\varphi \wedge \neg\varphi)$. But otherwise such restrictions seem to add no new validities. For instance, even if we restrict attention to models $\mathscr{M}$ in which $\mathscr{R}_{\mathscr{M}}$ is the universal relation, the change from $B(\varphi \wedge \neg\varphi) \rightarrow B\psi$ to $\neg B(\varphi \wedge \neg\varphi)$ is the only one needed in T_1.

We already noted that validity$_1$ invalidates (R_1). It also invalidates (R_2), as is evident from the fact that, given (E), (R_1) and (R_2) are interderivable. This, it seems to us, shows that validity$_1$ is not a particularly natural notion, not at any rate in the context of extensional models. For we regard it as counterintuitive that a certain sentence form, such as that identified by (E), should qualify as valid, and yet that sentences asserting that sentences of that form are true should not count as valid. This unfortunate property validity$_1$ shares with a hierarchy of similar notions, of which it forms the bottom rung. Define a schema to be *valid*^{n} iff all its instances are true in all models of the form $\mathscr{M}^{\lambda+k}$ with λ a limit ordinal and $k \geqslant n$. (So validity$_1$ coincides with validity^1!) While the set of valid^n schemata strictly increases with increasing n, all the validity concepts in this progression suffer from the same drawbacks as validity$_1$; moreover, no motivated choice between them seems at all possible.

The illustrations we gave of the difference between validity$_1$ and validity$_2$ also show a difference between validity$_1$ and what looks a priori like an intermediate notion between validity$_1$ and validity$_2$, that according to which a schema μ is valid respect to $\mathscr{B}$ iff there is a number n such that for every instance $\boldsymbol{\mu}$ of μ and every model $\mathscr{M}^{\lambda+k}$, with λ a limit ordinal and k a natural number $> n$, $\boldsymbol{\mu}$ is true in $\mathscr{M}^{\lambda+k}$. In fact, the proof of Proposition 20 shows that for the extensional case this new concept gives the same set of schemata as validity$_2$. This appears to remain true when we consider validity in relation to non-extensional model structures.

When reflecting on the importance of Propositions 18, 19, 20 and 21 we should not forget that theorems of the type they exemplify characterize validity in what may be considered a rather crude way. There are substantial families of sentences of L which verify many more schemata than are valid universally. To be precise, if $\mathscr{B}$ is a class of models and $\mathscr{F}$ any set of sentences, say that the schema μ is *verified*

by $\mathcal{F}$ in $\mathcal{B}$ iff for every interpretation $I(\mu)$ of μ such that $\mathrm{Ran}(I) \subseteq \mathcal{F}$, model $\mathcal{M} \in \mathcal{B}$ and every $w \in W_{\mathcal{M}}$, $[\![I(\mu)]\!]_{\mathcal{M},w} = 1$. Then, in particular, if $\mathrm{St}(\mathcal{B})$ is the set of all sentences stable throughout $\mathcal{B}$, the set of schemata of our propositional language *PL* verified by $\mathrm{St}(\mathcal{B})$ in $\mathcal{B}$ will include the set of all those that are valid (in the sense of modal logic) on the set $\mathcal{A}$ of frames represented in $\mathcal{B}$ — i.e. the set of all possible worlds structures $\langle W_{\mathcal{M}}, R_{\mathcal{M}} \rangle$ with $\mathcal{M} \in \mathcal{B}$. This suggests that in those cases where it is impossible to axiomatize the set of valid sentences of *L*, it might be possible to obtain results of a type intermediate between those exemplified by Proposition 13 on the one hand and Proposition 18 on the other, results which specify for one or more families of sentences the sets of schemata verified by those families. Such results would give a much better impression of how much doxastic logic survives in the definitions considered in this paper than all-or-nothing results in the style of Propositions 18 and 19. Whether such a more refined approach is really going to bring us substantially more, is a matter that still needs investigation. It may well be that the only two sentence sets that are relevant in this context are (i) the set of all sentences, and (ii) the set of sentences that stabilize throughout the class of models used to define validity. So far we have not found an interesting example of a set intermediate between these two and which validates a set of schemata which lies properly between the schemata sets validated by the sets (i) and (ii).

The alternatives we have explored in this section remain fairly close to the semantics we introduced in II.1.2. But there are a number of other possibilities some of which depart much more radically from the framework we have used. One of these, which still remains quite close to the original, is a somewhat different elaboration of Belnap's idea. This version circumvents the notion of a revision scheme. Instead revisions at limit stages produce sets of models. In particular at stage ω we associate with any starting model $\mathcal{M}$ the set of all models that could be obtained from the sequence $\{\mathcal{M}^n\}_{n \in \omega}$ by any choice of $[\![B]\!]$ that agrees with the local stability principle. From this point onwards the revision process has to deal with sets of models rather than single models at both successor and limit stages. The precise definition of this process requires some care, but we omit the details.

We mention this option not so much for its own interest, but because it naturally leads to others that constitute more significant departures from our original semantics. For instance, we could, instead of adopting

the revision process just indicated, use the set $\mathcal{M}^{\omega}$ of models obtained at stage ω as the basis of a supervaluation on the original model $\mathcal{M}$. This is tantamount to adopting at that stage a Kripke model, in which the extension and anti-extension of B in w contain only those φ which belong to all or none of the models, respectively, in the set $\mathcal{M}^{\omega}$. In this way we are led to a theory which is a hybrid between that of Kripke (1975) and those of Herzberger (1982) and Gupta (1982). The proper setting for a study of the various options that arise if we proceed in this direction would, we expect, be some intensional version of the "Unified Theory" sketched in Herzberger (1982),

While there are substantial differences between the various possibilities we have so far mentioned, they are nonetheless all variations on the general theme of a theory which combines the inductive and semi-inductive procedures of Kripke and others with the possible worlds account of intensionality. A much wider range of options opens up when we abandon the theme itself, for instance by embracing a different analysis of the intensional. In Part I of this paper, we criticized possible worlds semantics for its inability to provide a satisfactory explication of many attitudinal notions. Some of our criticisms had nothing to do with self-reference and apply no less to the semantics presented in this paper than they do to possible worlds semantics in its more familiar forms. Indeed, we expressed our doubts about the framework employed in this study, and made a commitment to search for one that meets those criticisms. Whether such a theory will shed new light on attitudinal logic is at this point a matter of speculation. But whether it does or not, it will be needed in any case.

NOTES

[1] See Montague (1963).

[2] Usually S_4 and S_5 are given as systems that include an inference rule (the rule of necessitation, which allows passage from $\vdash \varphi$ to $\vdash \Box\varphi$). However, it is possible to reformulate these systems as sets of axioms, e.g. by adopting as axioms all sentences obtained by prefixing $\Box$ to an instance of one of the axiom-schemata that are included in the familiar formulations of these systems.

[3] The system has the property that belief in certain intuitively self-evident statements, such as e.g. the single axiom of Robinson's Q, entails belief in any statement whatsoever, including transparent contradictions.

It has been put to us that the principle $(B4)$ lacks the intuitive plausibility that attaches to the principles $(B1)$–$(B3)$. So it may be appropriate to say a few words in its defense. One argument in favor of $(B4)$ runs like this: As G. E. Moore was the

first to observe, one cannot coherently believe that one believes that φ and yet that φ is false. Thus, the conjunction $B(\varphi)$ & $\neg\varphi$ is doxastically impossible. In other words we have $\neg(\neg B\neg)(B(\varphi)$ & $\neg\varphi)$, in view of the fact that $(\neg B\neg)$ expresses doxastic possibility. This last formula, however, is equivalent to $B(B(\varphi) \rightarrow \varphi)$, given $(B1)$—$(B3)$.

A second defense rests on a different principle, according to which a believer has complete access to what he believes. This means that for each φ the believer knows, and thus believes, that $B(\varphi)$, or else knows, and thus believes, that $\neg B(\varphi)$. Moreover, he will believe $B(\varphi)$ (if and) only if he believes that φ. So we have either $B(B(\varphi))$ & $B(\varphi)$ or else $B(\neg B(\varphi))$. Given $(B1)$—$(B3)$ each of these two disjuncts entails $B(B(\varphi) \rightarrow \varphi)$.

When reflecting on the plausibility of $(B4)$ it is important to remain attentive to the difference between the claim that all instances of $(B4)$ are valid — which is the claim at issue here — and the principle $B(\forall x)(B(\varphi) \rightarrow \varphi)$, which given $(B1)$—$(B3)$ properly entails $(B4)$. This last principle does not seem to be plausible to us. In fact, to believe $(\forall\varphi)(B(\varphi) \rightarrow \varphi)$ is a form of hubris that some of us may succumb to; but it is surely not part of doxastic logic.

It should also be noted in this connection that $(B1)$—$(B4)$ is not the only incompatible set of doxastic principles that yields incompatibility results. Other incompatible sets (taken from Rob Koons', 1987) are for instance: $(J1)$ $J(\neg J(\varphi)) \rightarrow \neg J(\varphi)$; $(J2)$ $J(\varphi)$ where φ is an axiom of first order logic; $(J3)$ $J(\varphi \rightarrow \psi) \rightarrow (J(\varphi) \rightarrow J(\psi))$; $(J4)$ $J(\alpha)$, where α is the conjunction of the axioms of Robinson's arithmetic; and $(J5)$ $J(\beta)$, where β is an instance of one of $(J1)$—$(J4)$; and the closure under the rule $\vdash\varphi \Rightarrow \vdash I(\varphi)$ of the principles $(I1)$ $I(\varphi \rightarrow \psi) \rightarrow (I(\varphi) \rightarrow I(\psi))$; $(I2)$ $I(\varphi) \rightarrow I(I(\varphi))$; and $(I3)$ $\neg I(\bot)$.

[4] The approach is based on Discourse Representation Theory. See Asher (1986); Kamp (1985).

[5] See footnote 2.

[6] Here we assume that particular predicates of $L(T)$ have been designated to represent the arithmetical notions of successor, $+$ and $\cdot$.

[7] For a more detailed description of Kripke models, see Section II.1 below. We say that $\langle W, R, [\ ]\rangle$ is a model of the theory T iff there exists $w \in W$ such that each sentence of T is true in $\langle W, R, [\ ]\rangle$ at w.

[8] See Montague (1968), Montague (1970). We will sometimes refer to Montague's Intensional Logic as 'IL'.

[9] Perhaps attitudes to the effect that a certain sentence expresses a certain content are not quite properly called beliefs. It appears for instance that we say 'a realized/did not realize that such and such is expressed by sentence S', rather than 'a believed that such and such is expressed by sentence S'. But in any case the attitudes in question are similar enough to belief for this not to affect the argument.

[10] Note that the route by which this person may find himself entangled in doxastic or epistemic paradox is not much different from that which will lead him into trouble with the liar paradox. Note also in this connection that from the internal perspective the logic of belief is arguably much stronger than that defined by $(B1)$—$(B4)$ and in fact not all that different from the logic of truth, as given by Schema T. We will return to this point in Sections II.2.2 and II.2.3.

[11] Kaplan and Montague (1962) and Montague (1963). Of course, in the present case the language would be a "language of thought" and potentially quite different from the languages in which propositional attitudes are expressed publicly. The motivation for the work of Kaplan and Montague relates rather to public language.

[12] When we noted the inexpressibility of the expression relation in the predecessor of this paper, Asher and Kamp (1986), we failed to realize that precisely the same point had been made more than a decade ago by Parsons (see Parsons (1974)). We apologize for the oversight.

The impossibility of representing in IL the expression relation between propositions and the Gödel numbers of the sentences expressing them has, we have just seen, nothing to do with the presence or absence of attitudinal predicates but arises independently, by virtue of the fact that IL contains the sentence forming operator $\check{}$. But in slightly different systems, which lack $\check{}$, but instead contain, say, a belief predicate B satisfying $(B1)$—$(B4)$, addition of E, with the axioms we just gave for it, would also be impossible. In such systems E could be used to define an attitudinal predicate of numbers B', such that $(B'(n) = (\exists p)(E(n, p) \,\&\, B(p))$. Under suitable conditions it is possible to show that the principles governing B also hold for B'. The contradiction then follows as in Montague (1963) or Thomason (1980).

[13] Strictly speaking it is the combination of attitudinal (e.g. epistemic or doxastic) logic and the underlying general logic that leads to paradox. So it would seem in principle conceivable that consistency could be secured by limiting the underlying logic rather than the logics of the particular attitudinal notions involved; or, alternatively, by giving up a little of both. We do not know if these options have any practical value. We will ignore them for most of the paper, but will return to them briefly in Section II.2.3.

[14] See Gupta (1982) and Herzberger (1982).

[15] See Quine (1953).

[16] "Reject" admittedly is rather vague, and might be said to beg precisely the question at issue, viz. whether (1) is false or just incapable of being known. But it can be made more precise in the following way. Certain declarative speech acts, among them judges' decrees, carry the implication that they constitute knowledge for those to whom they are addressed; their truth is warranted by the authority of the person or institution from which they issue. It is possible to construe this component of the meaning of such utterances as involving some sort of self-reference. The statement appears to be saying, among other things, that it itself is known to its addressees. On the face of it this is "self-reference" of a pragmatic, rather than a syntactic or semantic, sort. But we can, without producing too much distortion, make it formally explicit by construing the decree as overtly self-referential, viz. by representing it as:

$$(2) \qquad (M \,\&\, \neg(\exists t < t_m)K_t M) \vee (T \,\&\, \neg(\exists t < t_t)K_t T) \,\&\, K_{t_0}(2),$$

where t_0 denotes the time at which the decree is issued. (We have ignored the distinction between knowledge and belief here, but this does not affect the point. Note that although (2) is explicitly self-referential, it is quite different from the self-referential versions of the decree that can be found in two earlier treatments of the paradox, which emphasize its self-referential character, viz. Kaplan and Montague (1962) and Shaw (1958)).

If the decree is interpreted as (2), then K's inference that it must be false is above suspicion.

[17] An actual person in K's predicament would presumably stop chasing his own tail pretty soon, and wait for things to come in a state of bemused apprehension.

[18] See e.g. Doyle and McDermott (1980); Moore (1983), McDermott (1986).

[19] In K's case it is not a matter of the sentence being falsified by the state of affairs, but rather of its being possible in the light of it; but that is an irrelevant discrepancy for the point we are concerned to make.

[20] See in particular Asher (1986), Asher (1987).

[21] Some such logics can be found in Asher (1986) and Asher (1987). We intend to study these and others in forthcoming work.

[22] In our framework it is not possible to eliminate global self-reference entirely, for the quantifiers range over sets that include the sentences of the language. Thus any sentence that contains a quantifier will contain quantification over a domain to which it itself belongs. This kind of impredicativity can be rendered harmless in several ways, for instance by ensuring that the atomic formulae in which the bound variable occurs are true of some quantifier-free sentence. For details see Section II.1.4.

[23] The assumption that every individual in the domain is named by a constant is made strictly for reasons of convenience. It obviates the need to fuss with assignment functions in stating the truth conditions of quantified sentences. Strictly speaking, our procedure is not entirely sound; for instance, it excludes models whose domain has a cardinality exceeding that of the set of constants of L. Such problems are easily overcome by allowing the set of constants to be expanded when necessary. The issue is familiar from standard model theory and needs, we trust, no further elaboration.

[24] This perspective on the adjustment procedure defined above for $[B]_{\mathcal{K},w}$ is motivated by the observations made in Section 1.5 about the role revision plays in the hangman paradox. It is not evident, however, that the revision procedure we adopt in our formal analysis correctly captures the kind of belief revision we identified in I.5. We argued there that, as an effect of the subject's reflections on a paradoxical sentence, his beliefs actually change. If the totality of his beliefs is changing and that totality is reflected in the set of doxastic alternatives, then that set ought to change also. In the present formalization of belief revision, however, the relation R remains constant. The only changes that can happen to the set of alternatives for w concern the extensions of B in the members of this fixed set of alternatives. But if these extensions do no more than record what is determined already in some other way, then a change of extension cannot, it would seem, be regarded as a change from one possible world into another. For this reason the relevance of our formal treatment to the phenomenon of belief change remains in the last analysis problematic. A satisfactory resolution of this problem is, we believe, possible only within a more explicitly representational account.

[25] We will consider some alternatives to Herzberger's rule in II.2.3.

[26] Note that since L has identity, designative self-reference can always be mimicked by quantificational self-reference. For instance suppose that $[b]_M$ is the sentence $\neg B(b)$. Then, since $\neg B(b)$ is logically equivalent to $(\forall x)(x = b \rightarrow \neg B(x))$, this last sentence says that some sentence logically equivalent to itself is not believed. Since in the semantics presented here belief is preserved by logical equivalence, the sentence will be true if and only if it is not believed. Similarly, if $[b]_M = \neg B(c)$ and $[c]_M = B(b)$, then the sentences

(1) $(\forall x)(x = c \rightarrow \neg B(x))$

and

(2) $(\forall x)(x = b \rightarrow B(x))$

can be interpreted as making assertions about each other in M, (1) that (2) is not believed and (2) that (1) is believed. And so on.

[27] We have not pursued the question whether the results of this section can also be established for model structures M in which $<_M$ is loop-free but not well-founded. From our present perspective this question seems of little importance.

[28] In any such model $\mathcal{M}$, $\varphi \in [B]_w^{a+1}$ iff $[\varphi]_w^a = 1$ and if $\mathcal{M}$ is coherent then $\varphi \in [B]_w$ iff $[\varphi]_w = 1$. Thus, B behaves like a truth predicate.

[29] As Gupta is concerned with truth, his proof applies directly only to extensional model structures. However, it can easily be modified to apply to non-extensional structures as well.

[30] In response to a request of the editors of the present volume to reduce the length of the original manuscript, we have cut the present section to the bare minimum that remains here. We chose to make the cuts in this part, as most of its propositions are fairly straightforward generalizations of results found in Gupta (1982) although the method we have used to prove our propositions differs substantially from his. The interested reader may consult the unabridged version in a technical report published by the Center for Cognitive Science at the University of Texas.

[31] It would be possible to relax the constraints on M somewhat and to require only, say, that at each w the interpretations of O, $'$, $+$, $\cdot$ provide a model of Robinson's Q. But little would be gained from such a generalization for our present purpose.

[32] We do not know whether there exists a single natural condition on R which is both necessary and sufficient for the essential incoherence of any model structure M verifying the denotation relation ($C3$). Arguably this is problem of little conceptual importance as the condition would have to be tailored to the particular case of self-reference at hand. Moreover, the following consideration indicates that the solution could not be very simple. A necessary condition for essential incoherence that is relevant to model structures in which R is not transitive as well as to structures in which it is, is that the inverse of R be not well-founded. But this condition is not sufficient, as the following model structure M, in which R is indeed not-well-founded, shows. Let M be sentence-neutral, satisfy ($C3$) and be such that $W_M = \{w_0, w_1\}$ and R_M is the relation $\{\langle w_0, w_1\rangle, \langle w_1, w_0\rangle\}$. Let $\mathcal{M}$ be the expansion of M obtained by putting $[B]_{\mathcal{M}, w_0} = \{b\}$ and $[B]_{\mathcal{M}, w_1} = \emptyset$. Then $\mathcal{M}$ will be coherent and thus M not essentially incoherent. When one generalizes from this example, one realizes that a condition on R which is both necessary and sufficient for essential incoherence would have to be fairly complicated. To appreciate this, consider a similar model structure with worlds and an alternativeness relation R_M which consists of a set of disjoint loops: if R_M contains a loop with an odd number of elements then M is essentially incoherent; on the other hand, if R_M consists only of loops with even number of elements then M is not essentially incoherent.

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

[34] This set appears in several papers on the liar paradox. Herzberger (1982), for instance, discusses some of its formal properties. Of course, the discussions in these papers are restricted to the case where B behaves as a truth predicate.

[35] For instance in any model structure of the kind considered in Proposition 5 we can define, by means of some formula $\beta(x, y)$ of L, the relation which holds between n and m iff both n and m are natural numbers and m is the Gödel number of the sentence $B^n(\exists x)x = x \;\&\; \neg B^{n+1}(\exists x)x = x$, where for any sentence φ $B^1\varphi = B(\varphi)$ and $B^{i+1}\varphi = B(B^i\varphi)$. Suppose that $[wR] \neq \emptyset$ for all $w \in W_M$ and that $\mathcal{M}$ is an expansion of M such that $[B]_{\mathcal{M}, w} = \emptyset$ for all $w \in W_M$. Then for any w $n + 1$ will be the only natural number k such that $(B^n(\exists x)x = x \;\&\; \neg B^{n+1}(\exists x)x = x) \in [B]^k_{\mathcal{M}, w}$. Now let $\psi(x)$ be any arithmetical formula of L whose only free variable is x and let $\theta(y)$ be the formula $(\forall x)(\psi(x) \;\&\; \beta(x, y)$. Then the sentence $(\exists y)(\theta(y) \;\&\; B(y))$ belongs to $[B]^k_{\mathcal{M}, w}$ iff $k - 2$ is a member of the set of natural numbers defined by ψ in M. So since for every arithmetically definable set E the set $\{k + 2: k \in E\}$ is also arithmetically definable, every arithmetically definable set is the φ-profile at some world in some expansion of M for some sentence φ. Note, moreover, that the assumption we have made about R is quite weak. In particular, it is satisfied by all extensional model structures.

[36] Some of the extant solutions to the liar paradox differ in the components to which they propose revisions. Thus Kripke's semantics leads naturally (though not necessarily!) to a weakening of the underlying logic as well as of the principles of the "logic of truth" (i.e. the instances of schema T). The Tarskian solution, involving a hierarchy of metalanguages, is tantamount to eliminating the devices for self-reference.

[37] To the question: which are the relevant worlds of $\mathcal{M}$, there seem to us to be only two plausible answers: either all worlds in $W_{\mathcal{M}}$ count as relevant, or else only the "real" worlds of $\mathcal{M}$ do. Let us explain what we mean by the second of these answers. The possible worlds semantics for belief tries to capture the content of the subject's beliefs through the set of all worlds compatible with his beliefs. In the model theory we have adopted here these belief worlds are treated on a par with the world in which the subject is situated and has his beliefs. But intuitively the two kinds of worlds have a different status. If a valid sentence that speaks of the subject's beliefs is to be one that is true no matter what the beliefs are that the subject either does or might hold, then the definition of validity ought to be restricted to those worlds which represent such possible states of affairs. The (actual and/or possible) belief worlds need not be among those. The distinction between "real" worlds and belief worlds may have some formal repercussions, but the matter seems to us of too marginal an interest to the present discussion to merit closer attention.

[38] See Van Benthem (1983).

[39] When written out the proof is quite long while revealing next to nothing that is new. The need to ensure that the constructed countermodel has a constant domain and that at each world the extension of S consists of precisely the sentences of L complicates matters somewhat, and seems to require a judicious application of the Omitting Types Theorem (see Chang and Keisler (1973)). One can avoid these complications if one is prepared to alter the notion of a model as defined in II.1.2 by (i) dropping the requirement that D is constant and (ii) changing the condition on $[S]$ into the weaker requirement that $[S]_w$ include the sentences of L.

[40] For example, suppose that on a given evening one of the two authors of this paper

concludes the draft of a section with the dispirited words: 'Anyway, my coauthor won't believe a single one of all the statements I am making in this draft.' Suppose also that the second author is finishing that same evening a draft of a different section in a similarly despondent mood, and concludes it with the words: 'I don't believe a single one of the things I have said in this draft.' Suppose, moreover, that the second author does in fact neither believe any one of the statements of the first author's draft, with the possible exception of its last statement, nor any of the statements in his own draft, again with the possible exception of its last statement. Then once the second author has read the first author's draft, the last statements of each of the two drafts are paradoxical. But it is hard to see, at least without additional information about the case, any reason why the disjunction of the two statements should be incapable of achieving the stability that is within the reach of the disjunction of one and the negation of the other.

[41] In an addendum to his paper as it appears in Martin (1984), Gupta himself expresses dissatisfaction with this revision rule for limit ordinals. We do not know whether his reasons relate to the complaint voiced here.

[42] Another rule to deal with limit stages is the "maximizing rule", which makes $[B]_{\mathcal{M},w}$ as large as is compatible with the local stability principle. This rule seems quite odd from a conceptual viewpoint; it apparently embodies the principle that one should adopt as beliefs all that hasn't been permanently disqualified as such in the course of the unbounded sequence of revisions that the new intension is to sum up. One should be wary of a logic built on such a foundation. In fact, the logics that can be defined with the help of this rule appear to be quite curious. In particular, if $\mathscr{B}$ is a class that includes $\mathscr{M}^\lambda$, where $\mathscr{M}$ is metastable and λ is a limit ordinal, then $(B3)$ will typically not be valid in $\mathscr{B}$. On the other hand, if $\mathscr{B}$ consists exclusively of metastable $\mathscr{M}^\lambda$ with λ a limit ordinal, where moreover $R_{\mathscr{M}}$ is transitive and reflexive on its range, then $(B1)$, $(B2)$ and $(B4)$ will be among the valid schemata in $\mathscr{B}$. We have not investigated precisely what logics can be obtained with the help of this rule.

[43] Since W_M consists of the single world w_0, we will leave out all references to worlds for the remainder of the proof. The implicit reference will always be to w_0.

[44] We define compatibility below.

[45] Inspection of the proof reveals that there is no need to consider all the ordinals. In fact for given μ we can carry out what is essentially the construction described here on the ordinal $w^{\deg(\mu)}$.

[46] For natural numbers n $B^n\varphi$ is defined as follows: $B^0\varphi$ is φ; $B^{n+1}\varphi$ is $B(c_{B^n\varphi})$.

[47] See footnote 46.

REFERENCES

Asher, N.: 1986, 'Belief in Discourse Representation Theory', *Journal of Philosophical Logic*, pp. 137—189.

Asher, N.: 1987, 'A Typology of Attitude Verbs and their Anaphoric Properties', *Linguistics and Philosophy* **10**, pp. 125—197.

Asher, N. and Kamp, H.: 1986, 'The Knower's Paradox and Representational Theories of Attitudes', in *Theoretical Aspects of Reasoning about Knowledge*, ed. J. Halpern. Los Angeles: Morgan Kaufmann, pp. 131—148.

Belnap, N.: 1982, 'Gupta's Rule of Revision Theory of Truth', *Journal of Philosophical Logic* **12**, pp. 103–126.
Benthem, van J.: 1983, *Modal Logic and Classical Logic*, Naples: Bibliopolis.
Burgess, J.: 1986, 'The Truth is Never Simple', *Journal of Symbolic Logic* **51**, pp. 663–681.
Chang, C. and Keisler, J.: 1973, *Model Theory*, Amsterdam: North Holland.
Doyle, J. and McDermott, D.: 1980, 'Non-Monotonic Logic I', *Artificial Intelligence* **13**, pp. 41–72.
Gupta, A.: 1982, 'Truth and Paradox', *Journal of Philosophical Logic* **12**, pp. 1–60.
Herzberger, H.: 1982, 'Notes on Naive Semantics', *Journal of Philosophical Logic* **12**, pp. 61–102.
Herzberger, H.: 1982, 'Naive Semantics and the Liar Paradox', *Journal of Philosophy* **79**, pp. 479–497.
Kamp, H.: 1985, 'Context, Thought and Communication', *Proceedings of the Aristotelian Society*, 1984–85.
Kaplan, D. and Montague, R.: 1960, 'A Paradox Regained', *Notre Dame Journal of Formal Logic* **1**, pp. 79–90.
Koons, R.: 1987, Analogues of the Liar Paradox in Epistemic Logic, Ph.D. Thesis, UCLA.
Kripke, S.: 1975, 'Outline of a New Theory of Truth', *Journal of Philosophy* **72**, pp. 690–715.
Martin, R.: 1984, *New Essays on Truth and the Liar Paradox*, OUP, p. 198.
Montague, R.: 1963, 'Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability', *Acta Philosophica Fennica* **16**, pp. 153–167.
Montague, R.: 1968, 'Pragmatics', in *Contemporary Philosophy: A Survey*, ed. R. Klibansky, Florence: La Nuova Italia Editrice, pp. 102–122.
Montague, R.: 1970, 'Pragmatics and Intensional Logic', *Synthese* **22**, pp. 68–94.
Moore, R.: 1983, 'Semantical Considerations on Nonmonotonic Logic', SRI Technical Note.
Quine, W. V. O.: 1953, 'On a So-Called Paradox', *Mind* **62**, 65–67.
Shaw, R.: 1958, 'The Paradox of the Unexpected Examination', *Mind* **67**, pp. 382–384.
Thomason, R.: 1980, 'A Note on Syntactical Treatments of Modality', *Synthese* **44**, pp. 391–395.

The University of Texas at Austin,
Austin, TX, U.S.A.