# REASONING ABOUT BELIEF AND KNOWLEDGE WITH SELF-REFERENCE AND TIME

Nicholas Asher
Department of Philosophy
&
Center for Cognitive Science
The University of Texas at Austin

## ABSTRACT

In two previous papers (Asher & Kamp 1986,1987), Hans Kamp and I developed a framework for investigating the logic of attitudes whose objects involved an unlimited capacity for self-reference. The framework was the daughter of two well-known parents-- possible worlds semantics and the revisionist, semi-inductive theory of truth developed by Herzberger (1982) and Gupta (1982). Nevertheless, the offspring from our point of view was not an entirely happy one. We had argued that orthodox possible worlds semantics was an unacceptable solution to the problem of the semantics of the attitudes. Yet the connection between our use of possible worlds semantics and the sort of representational theories of the attitudes that we favor remained unclear. This paper attempts to provide a better connection between the framework developed in the previous papers and representational theories of attitudes by developing a notion of reasoning about knowledge and belief that a careful examination of the model theory suggests. This notion of reasoning has a temporal or dynamic aspect that I exploit by introducing temporal as well as attitudinal predicates.

## 1. REASONING AND A REPRESENTATIONAL THEORY OF ATTITUDES

Reasoning about propositions an agent entertains, believes, or knows involves the manipulation of structured objects by means of certain rules. A representational theory of attitudes supports such a view of reasoning, because a representational theory takes these propositions to be structured objects of the sort amenable to manipulation by rules of proof. In real life reasoning takes an agent from one mental state at one time to another at a later time; reasoning is thus essentially a dynamic process. One can abstract away the temporal element and assimilate reasoning to the paradigm of formal proof. Although it should not obscure the fact that an agent's reasoning about his beliefs or knowledge is a dynamic process, this assimilation is useful, because it leads to a precise formulation of rules for reasoning. Such a formulation amounts to a "logic for the attitudes."

That an agent's reasoning about his beliefs or knowledge employs what one might justifiably call a logic for the attitudes, does not, from the representationalist's perspective, entail that the agent's actual (or even ideal) beliefs are closed under such principles. Rather, it means that if the agent were to reason using only these principles, his reasoning would be *sound.* One task, then, of the representationally minded logician or philosopher is to determine what principles of reasoning about belief or knowledge are sound. In order to show that some set of such principles are sound, a representationalist should provide a model in which every application of such principles is sound.

The difficulty with the representationalist's conception arises when one attempts to endow the attitude predicates with the natural principles of reasoning that traditional systems of doxastic and epistemic logic suggest. The moral of the work by Kaplan, Montague (1960, 1963) and Thomason (1980) is that the combined principles of doxastic or epistemic reasoning, when applied to structured objects like representations or sentences, are not sound in contexts where, as in arithmetic, the capacity for constructing arbitrary, self-referential statements exists. One example of a self-referential statement found to cause trouble was discussed by Montague and Kaplan in connection with the Hangman Paradox; the statement says of itself that it is not known, and when minimal, standard principles of epistemic are applied to it, a contradiction quickly ensues. The motivation for Asher & Kamp (1986, 1987) was primarily to discover what principles of doxastic and epistemic reasoning could still be held in contexts where the potential for self-reference is unlimited. This led to the development of an intensional analogue to the extensional, Herzberger and Gupta models for truth, which I review now. My approach here to these intensional models differs however from the original motivation; I take them to furnish not only models of principles of correct reasoning about knowledge or belief but also but also a different perspective on what such reasoning amounts to.

## 2. MODELS, COHERENCE, AND BELIEF REVISION

To set the stage for a theory of reasoning about belief, I will need some machinery. Consider a first order language L with identity, a denumerable infinity of individual constants, and one distinguished predicate S (to be read as *is a sentence*). L(B) is the language L expanded with a 1-place predicate B (to be read as *some fixed believer K believes that*). (I'll consider belief as the only attitude to simplify matters.) A *model for* L is a quadruple M = <W, R,D, [[]]> 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; [wR] is the set of alternatives to w);
(iii) D is a non-empty set (the domain of individuals);(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, $[\![c]\!]_w \in D$; and if Q is an n-ary predicate of L, $[\![Q]\!]_w \subseteq D^n$;
(v) for each $w \in W$, $[\![S]\!]_w$ is the set of sentences of L;
(vi) each individual constant c is a *rigid designator*, i.e., for all w, $w' \in W$, $[\![c]\!]_w = [\![c]\!]_{w'}$.
(vii) for each $d \in D$ and $w \in W$, there is a constant c of L such that $[\![c]\!]_{w,M} = d$.

A *model* for L(B) is a pair <M, $[\![B]\!]$> where M is a model for L and $[\![B]\!]$ is an intension for B relative to M (i.e., a function from $W_M$ into $\wp(D_M)$) such that $\forall w \in W_M\ [\![B]\!]_w \subseteq [\![S]\!]_w$. We refer to models for L(B) simply as *models* and to models for L as *model-structures*. A model structure M is *extensional* just in case $W_M$ is a singleton and $\langle w,w\rangle \in R_M$.

An important notion for this conception of a model is the idea of *model coherence*. A model $\mathcal{M}$ is (*doxastically*) *coherent* iff the following statement is satisfied for each sentence $\psi$ and each world $w \in W_{\mathcal{M}}$:

$\psi \in [\![B]\!]_{\mathcal{M},w}$ iff $[\![\psi]\!]_{\mathcal{M},w'} = 1$ for all $w' \in [wR_{\mathcal{M}}]$.

A model structure M is *essentially incoherent* iff every model that expands M is incoherent. The notion of coherence brings together two, independent features of the models that are essential to the semantics of the attitudes, the alternativeness relation and the extension of the B predicate. The alternativeness relation in the model structure encodes plausible doxastic principles of reasoning and the basic doxastic facts that the agent may uncover through reflection; the predicate B's initial extension could represent what an agent might in fact believe. Coherent models are those models in which the agent believes (or could come to believe through reasoning) all that is doxastically possible for him to come to believe. Coherent models are those in which the agent can use all the principles of reasoning encoded in the alternativeness relation to their full effect.

Models that are incoherent may become coherent through the process of *model revision*. To define this notion, however, I need some auxillary notions. Define an *interpolation function on* a set A to be any function f from $\wp(A)^2$ into $\wp(A)$ 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 $f(A_1, A_2) \cap A2 = \emptyset$. A *revision scheme* is a function $\Re$ defined on the class of all limit ordinals such that for each $\lambda$ $\Re(\lambda)$ is an interpolation function on the set $S_L$ of sentences of L. Given a model $\mathcal{M}$ and a revision scheme $\Re$, the *revision sequence starting from* $\mathcal{M}$ *according to* $\Re$ is the sequence $\{\mathcal{M}^{\alpha,\Re}\}_{\alpha \in On}$, such that: $\mathcal{M}^{\alpha,\Re} = \langle W_{\mathcal{M}}, D_{\mathcal{M}}, R_{\mathcal{M}}, [\![]\!]^{\alpha,\Re}\rangle$, where $[\![\theta]\!]^{\alpha,\Re} = [\![\theta]\!]_{\mathcal{M}}$ for all nonlogical constants $\theta$ other than B, and $[\![B]\!]^{\alpha,\Re}$ is defined as follows:

i) $[\![B]\!]^{0,\Re}_w = [\![B]\!]_w$
ii) $[\![B]\!]^{\alpha+1,\Re}_w = \{\varphi: (\forall w' \in R_{\mathcal{M}})\ [\![\varphi]\!]_{\mathcal{M}^{\alpha}, w'} = 1)\}$
iii) $[\![B]\!]^{\lambda,\Re}_{\mathcal{M},w} = \Re(\lambda)(B^+_w, B^-_w)$, where $B^+_w = \{\varphi: (\exists\gamma<\lambda)(\forall\beta)(\gamma<\beta<\lambda \rightarrow \varphi \in [\![B]\!]^{\beta,\Re}_{\mathcal{M},w})$ and $B^-_w = \{\varphi: (\exists\gamma<\lambda)(\forall\beta)(\gamma<\beta<\lambda \rightarrow \varphi \notin [\![B]\!]^{\beta,\Re}_{\mathcal{M},w})$.

There are many different choices for revision schemes $\Re$ obeying the local stability principle. One that I will be using a great deal in this paper is the *Herzberger revision scheme* $\mathcal{H}$, in which $\mathcal{H}(\lambda)(B^+_w, B^-_w) = B^+_w$. I will call *Herzberger revision sequences* those revision sequences that employ the Herzberger revision scheme $\mathcal{H}$.

There are certain conditions under which coherence cannot be achieved no matter how many revisions are undertaken; in general, models in which paradoxical forms of self-reference are

present will not be coherent.[1] The presence of incoherent models leads to the following distinctions. $\varphi$ is *positively* (*negatively*) *stable in* a model $\mathcal{M}$ *with respect to* a revision scheme $\mathcal{R}$ *at* a world w iff $\varphi \in [\![B]\!]^{\beta, \mathcal{R}}_{\mathcal{M}, w}$ for all $\beta$ ($\varphi \notin [\![B]\!]^{\beta, \mathcal{R}}_{\mathcal{M}, w}$ for all $\beta$). $\varphi$ *stabilizes at* an ordinal $\alpha$ *in* a model $\mathcal{M}$ *with respect to* a revision scheme $\mathcal{R}$ (*at* a world w) iff $\alpha$ is the first ordinal $\beta$ such that $\varphi$ is positively or negatively stable (at w) in $\mathcal{M}^{\beta}$ with respect to $\mathcal{R}$. $\alpha$ is a *stabilization ordinal for* $\mathcal{M}$ (*at* w) *with respect to* $\mathcal{R}$ iff every $\varphi$ that stabilizes in $\mathcal{M}$ (at w) with respect to $\mathcal{R}$ stabilizes at some ordinal $\leq \alpha$ in $\mathcal{M}$ (at w) with respect to $\mathcal{R}$. If $\beta$ is any ordinal greater or equal to the first stabilization ordinal for $\mathcal{M}$ with respect to $\mathcal{R}$, the model $\mathcal{M}^{\beta, \mathcal{R}}$ is called a *metastable* model. Call $\gamma$ a *perfect stabilization ordinal for* $\mathcal{M}$ with respect to $\mathcal{R}$ just in case $\gamma$ is a stabilization ordinal for $\mathcal{M}$ with respect to $\mathcal{R}$, and $\varphi \in [\![B]\!]^{\gamma}_{\mathcal{M}, w}$ iff $\varphi$ stabilizes at some ordinal $\leq \gamma$ in $\mathcal{M}$ at w with respect to $\mathcal{R}$. The only model revision sequences that I will discuss that contain perfect stabilization ordinals will be those defined by restricting the choice of $\mathcal{R}$ to the Herzberger revision scheme $\mathcal{A}$. I shall often restrict myself to Herzberger revision sequences, as they are usually the simplest to manipulate. I will also refer to certain classes of models defined by Herzberger revision sequences; for instance, I will refer to the class of metastable models of the form $\mathcal{M}^{\alpha, \mathcal{A}}$ as the class of *Herzberger-metastable* models.

With a description of this model-theoretic machinery in hand, I want now to return to the question of reasoning about belief and the point of this machinery. If the aim were to represent correctly an agent's actual, doxastic or epistemic state, then the extension of B should not cohere with all the doxastic facts.[2] For in coherent models, agents are logically omniscient and it is precisely the lack of logical omniscience in real life agents that motivates a representationalist theory of the attitudes. On the other hand, an alternative goal is to represent what the principles of reasoning allow an agent to conclude legitimately. An agent should be able to come to believe anything that is doxastically possible for him to come to believe by legitimately reasoning about his beliefs. Then there are different ways one might represent these principles of legitimate reasoning. One way is to provide traditional axiomatizations of doxastic logic; the model-theoretical framework developed in Asher & Kamp (1986) (1987) and sketched here provides some complete and some partial axiomatizations of logics validated by various classes of models. I will say more about this approach in the next section.

A second way of using the model-theoretic framework to reflect legitimate principles of doxastic reasoning is to take seriously the idea that the process of model-revision itself captures an important aspect of correct doxastic reasoning. That aspect concerns the dynamics of doxastic reasoning, the way an agent moves from one belief state to another in reasoning about his beliefs. An original motivation for adopting the Herzberger-Gupta approach was that it appeared to capture certain aspects of the dynamic 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, 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

---

[1]One might think that any model structure which provides a general licence for self-reference is essentially incoherent. However, this is so only if the alternativeness relation satisfies certain constraints. For instance, if M satisfies the following condition,

(C1) $(\forall w \in W_M)([wR] = \emptyset \vee (\forall w' \in [wR])\,[w'R] = \emptyset)$

then M can be expanded to a coherent model. The reason for this is obvious: If w is a world such that $[wR_M] = \emptyset$ then in any expansion $\mathcal{M}$ of M all sentences will be stable at w after one revision, and if $[wR_M] \neq \emptyset$ but $(\forall w' \in [wR])\,[w'R] = \emptyset$ then every sentence becomes stable at w after at most two revisions. I refer the interested reader to Asher & Kamp (1987) for some further results on the effects of the alternativeness relation on coherence.

[2]One might require that the extension of B cohere with the "atomic" doxastic facts, though not with all their logical consequences. Let's call such models in which there is this partial coherence *normal* models. $\mathcal{M}$ is a *normal model* just in case for any atomic sentence $\varphi$ of L, $\varphi \in [\![B]\!]_{\mathcal{M}, w}$ iff $[\![\varphi]\!]_{\mathcal{M}, w'} = 1$ for all $w' \in [wR_{\mathcal{M}}]$.

Kripke's original idea. But in connection with the epistemic and doxastic paradoxes the revision aspect is especially important.

In introspective reasoning-- reasoning in which no new information is received from the outside by the agent-- the agent moves from one belief state to another by reflecting on the beliefs he already has. Sometimes, reflection simply adds to the set of beliefs; sometimes reflection leads to belief revision when the agent uncovers an inconsistency among his beliefs or perhaps some incompatibility between his explicitly held beliefs and what he is implicitly committed to. After a number of revisions one might imagine that if the agent reasons correctly, he should achieve a stable doxastic state that persists under further reflection. In some particularly bizarre circumstances like those that arise in the Knower Paradox or the Hangman paradox, however, the agent revises his beliefs but fails to achieve a stable doxastic state.

Model revisions are intended to model at least some aspects of this introspective belief revision. Model revisions also verify an agent's correct reasoning in the following sense: if an agent begins in a certain belief state $\mathbb{B}_0$, which I identify here with a set of sentences, and reasons to a state $\mathbb{B}_1$ in the absence of any new information, then that reasoning process will be sound or correct just in case for any model $\mathcal{M}$ and world w if $\mathbb{B}_0 \subseteq [\![B]\!]_{\mathcal{M}, w}$, then there is an ordinal $\alpha$ such that $\mathbb{B}_1 \subseteq [\![B]\!]^{\alpha}_{\mathcal{M}, w}$. The notion of model revision, however, does not yield a full theory of belief revision. It does not provide rules or even guidelines for which belief an agent ought to throw out when he uncovers an inconsistency.[1] The process of model revision in effect finesses the difficult epistemic problem of determining which beliefs to keep in case of conflict by legislating that the doxastic possibilities always encode the "right" beliefs. Nevertheless, model revisions point out something of interest. They provide a tractable semantics for the process belief revision. This remains a problem of belief revision, even once one has determined the epistemic problems with belief revision-- namely, which beliefs are to be kept in case of conflict. The theory of model revision also reveals some of the complexities of belief revision that those concerned with the epistemic problems of belief revision have not addressed.

## 3. CONVERGENCE AND COMPLETENESS

One way to use a model theory for the purpose of determining and motivating a logic is to look for completeness results with respect to some natural class of models. The model-theoretic framework suggests a wide variety of classes of models for consideration. I define validity with respect to a class of models in the usual way: $\varphi$ is *valid with respect to* $\mathfrak{B}$, in symbols $\mathfrak{B} \models \varphi$, iff for every $\mathcal{M} \in \mathfrak{B}$ and $w \in W_{\mathcal{M}}$ $[\![\varphi]\!]_{\mathcal{M},w} = 1$. Asher & Kamp (1987) makes shows that completeness proofs are available for classes of coherent models. There are also other notions of validity, some more plausible than others, that can be defined with respect to any class of models $\mathfrak{B}$. I will introduce one in section 4.

The process of model revision not only yields coherent models but also defines a wide class of models that "converge" under the revision process to the same coherent model, a fact first noted by Gupta (1982) for extensional models. Every such class C of convergent models arises from a single model structure; that is, there is a model structure M such that every model in C is an expansion of M. Following Gupta's terminology, I will call a model structure M *Thomasonian* just in case every model expansion of M converges to a single coherent model. Or in symbols, a model structure M is *Thomasonian* just in case there is an ordinal $\alpha$ such that

[1]For an example of the latter, see Rescher's theory in Rescher (1976).

for any models $\mathcal{M}_1$ and $\mathcal{M}_2$ expanding M and revision schemes $\mathcal{R}_1$ and $\mathcal{R}_2$ , the revisions $\mathcal{M}_1{}^{\alpha, \mathcal{R}_1}$ and $\mathcal{M}_2{}^{\alpha, \mathcal{R}_2}$ are coherent and $\mathcal{M}_1{}^{\alpha, \mathcal{R}_1} = \mathcal{M}_2{}^{\alpha, \mathcal{R}_2}$. The expansions of Thomasonian model structures are particularly interesting, because in these models correct reasoning will lead to coherence regardless of what initial extension is assigned to the belief predicate and what choice of revision scheme is used. These models are most like standard possible worlds models for the attitudes; the extension of the B predicate is eventually wholly determined by the doxastic possibilities. Once such a point is reached in the revisions of expansions of Thomasonian model structures, the model revision conception of doxastic reasoning coincides with the ordinary conception of doxastic reasoning as enshrined in the axiomatizations of standard doxastic logic. So an understanding of these models in the present framework is a good point of departure in the investigations of the model revision conception of doxastic reasoning.

To figure out what sort of model structures are Thomasonian, I need the following notion of Gupta's. Say that 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 $1 \leq i \leq n$ 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 $\langle a_1, \ldots, a_{i-1}, s', a_{i+1}, \ldots, a_n \rangle \in [\![Q]\!]_{M, w}$. Say that M is *sentence-neutral* iff every predicate of L other than B is sentence-neutral in M. Coherent models result from a sentence-neutral model-structure when the model-structure obeys the following constraint on the denotation relations it posits. For 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 containing $c_1$ as a constituent. 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. The required constraint on $<_M$ is that it be well-founded.

In fact, when these conditions are met, we not only have coherence but convergence as well.[1]

**Proposition 1.** Let M be any sentence-neutral model structure such that $<_M$ is well-founded. Then M is Thomasonian.

It is possible to extend this result by weakening the assumption of sentence neutrality. 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, 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, notes 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. There is an intensional analogue to Gupta's remark for arbitrary model structures. Moreover, one can show:

**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) $\mathcal{M}$ is an expansion of M; (vi) the set U of sentences of L which do not stabilize in $\mathcal{M}$ is included in A. Then M is Thomasonian.

Another way to extend proposition 1 is this. 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-

[1]Anil Gupta suggested the generalizations in propositions 1,2 and the one in footnote 1 on page 5 to Kamp and myself. The proofs of the generalization follow quite straightforwardly from those given for intensional model coherence in Asher & Kamp (1987). I omit the proofs here, however, because they are quite long when written out.

neutral and $<_M$ is well-founded, then it is still true that every expansion of M becomes coherent upon repeated revision and converges to the same model-- i.e., M is Thomasonian.[1]

These results establish that there is a non-trivial class of coherent models of L(B) which not only allow quantification over objects of belief but also define a corresponding class of models in ordinary possible worlds semantics that incorporate a certain amount of ability to talk about the structure of objects of belief. The completeness proof of Asher & Kamp (1987) for the class $\mathfrak{B}$ of coherent models which expand some model structure M such that (i) M is sentence-neutral and (ii) $R_M$ is transitive and reflexive on its range shows that a slightly modified version of the standard system of doxastic logic known as quantified, "weak" S4 (the propositional schemata are given in Thomason (1980)) is complete with respect to $\mathfrak{B}$ in the sense that $\varphi$ is a theorem of that logic iff $\mathfrak{B} \models \varphi$. Other completeness proofs employing the same argument are available for classes of coherent models which expand model structures with different underlying alternativeness relations.

Unfortunately, such completeness proofs do not exist for many sets of models. The reason for this becomes clear with the following proposition (also from Asher & Kamp 1987). To state it I will need names for all the sentences of L. Let $c_\varphi$ be a fixed function which maps the sentences of L one-to-one onto some coinfinite subset of $C_L$. I assume for the remainder of the paper that, for every model structure M and sentence $\varphi$, $[\![c_\varphi]\!]_M = \varphi$, and that all models are expansions of such M.

> **Proposition 3.** Let $\mathfrak{B}$ be the class of all metastable models $\mathcal{M}^{\alpha, \mathcal{R}}$ for arbitrary revision schemes $\mathcal{R}$ such that $R_{\mathcal{M}}$ is transitive and reflexive on its range.[2] Then, for any sentence $\varphi$, $\mathfrak{B} \models Bc_\psi$, where $\psi$ is the sentence $Bc_\varphi \rightarrow \varphi$, iff $\varphi$ is stable in all members of $\mathfrak{B}$.

The constraints of transitivity and reflexivity on the range of $R_{\mathcal{M}}$ ensure that in coherent models, $Bc_\psi$ is valid for any sentence $\varphi$.[3] Proposition 3 entails that whenever $\mathfrak{B}$ is such that for some decidable set S' of sentences of L the set of those members of S' which are stable throughout $\mathfrak{B}$ is not recursively enumerable, then the set of sentences of L that are valid in $\mathfrak{B}$ is not recursively axiomatizable. This situation should arise in many cases.[4]

There is, however, a weaker type of completeness result that proposition 4 and its implications do not rule out. If one is interested in only the logical validity of certain traditional doxastic "schemata", such as for example the "weak S4" axioms used in Thomason (1980), then, given some particular class of models $\mathfrak{B}$, the question precisely which schemata are validated by $\mathfrak{B}$ may admit of an answer even if the set of L-sentences that are valid throughout $\mathfrak{B}$ is not recursively axiomatizable. I will appeal to this technique in the next section.

## 4. DYNAMIC THEORIES OF REASONING

---

[1] So far the techniques developed in Asher & Kamp (1987) only yield the following, partial result:
**Proposition.** Let M be a model structure such that (i) $<_M$ is well-founded; (ii) for all $w \in W_M$ $[\![Neg]\!]_{M,w}$ is the set of all pairs $<\varphi,\psi>$ of sentences of L such that $\varphi$ is the negation of $\psi$, and $[\![Con]\!]_{M,w}$ is the set of all triples $<\varphi,\psi_1,\psi_2>$ of sentences of L such that $\varphi$ is the conjunction of $\psi_1$ and $\psi_2$; (iii) all predicates of L other than Neg, Con and B are sentence-neutral in M. Moreover let $\mathcal{M}$ be an expansion of M such that (iv) the sentence

(1) $(\exists x)(\exists y)(Neg(x,y) \ \& \ \neg B(x) \ \& \ \neg B(y))$

stabilizes in $\mathcal{M}$ at all $w \in W_M$. Then M is Thomasonian.

[2] That is, $\forall w_1 \forall w_2 \forall w_3((w_1 R w_2 \ \& \ w_2 R w_3) \rightarrow w_1 R w_3)$ and $\forall w_1 \forall w_2 (w_1 R w_2 \rightarrow w_2 R w_2)$.

[3] Their use in this framework will not yield implausibly strong axioms, and they make the models more tractable under revision. I shall appeal to them quite often in what follows.

[4] See Burgess's (1986) results where $\mathfrak{B}$ includes only standard models of arithmetic-- but if $\mathfrak{B}$ includes non-standard as well as standard models, then we have an open problem.

While the principles of reasoning suggested by ordinary doxastic logic do have some application within the semantics for the attitudes I favor, the models also suggest a quite different way of looking at reasoning about knowledge and belief in terms of belief revision. Incorporating this element of model revision explicitly into the formalization of reasoning about attitudes yields a dynamic theory of reasoning, in which reasoning is conceived as a means of moving from one mental state to another. In such a system, certain rules for reasoning will cause the agent to move to a new belief state, which may *revise*, as well as add to, the antecedent belief state.[1] These rules will be the proof theoretic analogues of the jump operation in the sequence of model revisions. This operation, recall, turns a given model into a new one by adjusting the extensions of B to the set of beliefs determined by the alternativeness relation R. The sequence of model revisions will furnish a criterion for the correctness of these rules.

A natural deduction system such as that provided by Fitch, Kalish and Montague or Lemmon (to name some familiar ones) provide a setting within which to formalize such dynamic rules of reasoning. I focus here on the system of Lemmon (1965). That system defines a proof as a sequence of lines where each line consists of (i) a natural number label, (ii) a (possibly empty) set of labels of previous lines, and (iii) a formula. The formula in (iii) is the main entry of a line. The set of labels are of course the premises of the main entry. I will alter this format slightly by entering a pair consisting of a formula and a number for the main entry of the form (n, φ).[2] The number n serves as the index of a particular belief state. The rules of Lemmon's system may all be easily restated as applying only to lines with main lines having the same ordinal index. Thus, for instance, modus ponens becomes the following principle:

If a proof contains lines

(j) $\{l_1, ..., l_n\}$ $(n, \varphi)$

(k) $m_1, ..., m_j$ $(n, \varphi \rightarrow \psi)$

then we may write down as a new line:

(l) $\{l_1, ..., l_n, m_1, ..., m_j\}$ $(n, \psi)$

Similarly, I redefine the rule of conditional proof as follows:

Suppose a proof contains lines

(j) $\{j\}$ $(n, \varphi)$

(k) $\{j, m_1, ..., m_j\}$ $(n, \psi)$

Then we may write down as a new line:

(l) $\{m_1, ..., m_j\}$ $(n, \varphi \rightarrow \psi)$

Along with the rules of proof, I also redefine derivability. We say that φ is derivable from $\psi_1$, ..., $\psi_n$, written $\psi_1, ..., \psi_n \vdash' \varphi$, just in case there is an n such that for all $m \geq n$ there is a proof of $(m, \varphi)$ from $(m, \psi_1), ..., (m, \psi_n)$. It is easy that to see that this sytem, which I'll call $\mathbb{T}_0$, provides a notion of provability that coincides with the notion of first order logical consequence.

To turn $\mathbb{T}_0$ into a rudimentary system for reasoning about the attitudes, consider the following additional rule.

Suppose a proof contains the lines,

(j) $\{l_1, ..., l_n\}$ $(n, \varphi)$

($k_1$) $\{m_{1,1}, ..., m_{1,j}\}$ $(n + 1, B(\psi_1))$[3]

. .

. .

---

[1]Some remarks Anil Gupta made in a lecture on definitions at the University of Texas, May 1987) suggested this to me.

[2]The suggestion for this particular format is due to Hans Kamp.

[3]Note that to be properly stated I should have used the constants $c_\varphi$. I will assume the more familiar notation as an abbreviation of the talk of constants.

($k_i$) $\{m_{i,1}, ..., m_{i,k}\}$ $(n + 1, B\psi_i)$

and $l_1, ..., l_n$ are the labels of lines with main entries $(n, \psi_m)$ for $1 \leq m \leq i$, then one may add as a new line:

(l) $\{m_{1,1}, ..., m_{1,j}, ..., m_{i,1}, ..., m_{i,k}\}$ $(n + 1, B\varphi)$

I call this new rule $B_0I$ (belief introduction 0), and the system that results from adding $BI_0$ to $\mathbb{T}_0$, the system $\mathbb{B}_0$. Using $B_0I$ it is easy to show that if $\vdash' \mathbb{B}_0 \varphi$ then $\vdash' \mathbb{B}_0 B(\varphi)$ and that $\vdash' \mathbb{B}_0 (B(\varphi \rightarrow \psi) \& B(\varphi)) \rightarrow B(\psi)$. But also it appears that $\mathbb{B}_0$ has the following property: if $\vdash' \mathbb{B}_0 B(\varphi)$ then $\vdash' \mathbb{B}_0 (\varphi)$, a rule which I'll call R1. I will add (R1) to $\mathbb{B}_0$, and call the resulting system $\mathbb{B}_1$.

By exploiting the notion of propositional schemata used in Asher & Kamp (1987), one can give a completeness proof for $\mathbb{B}_1$ relative to the notion of validity provided by metastable models.[1] More precisely, I 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, $\perp$ and the sentence letters $p_1, p_2, ...$, and which has besides the truth-functional connectives $\neg$, &, $\vee$ the 1-place sentence operator B. I refer to the formulae of PL as *schemata*. An *interpretation of* PL *in* L is a function I that maps each sentence letter onto a sentence of L. Every interpretation I can be extended to all formulae of PL as follows: $I(T) = (\forall x)\, x=x$; $I(\perp) = (\exists x)\, x \neq x$; $I(\neg\mu) = \neg I(\mu)$; $I(\mu \& \nu) = I(\mu) \& I(\nu)$; $I(\mu \vee \nu) = I(\mu) \vee I(\nu)$; $I(B\mu) = B(c_{I(\mu)})$. Where $\mathfrak{B}$ is a class of models for L, the schema $\mu$ *is valid in* $\mathfrak{B}$, in symbols $\mathfrak{B} \models \mu$, iff $\mathfrak{B} \models I(\mu)$ for every interpretation I. There is another more useful notion of validity for schemata, however. Say that a schema $\mu$ is *valid** with respect to $\mathfrak{B}$, which I will write as $\mathfrak{B} \models^* \mu$, iff for each instance $\underline{\mu}$ of $\mu$ there is a natural number n such that $\underline{\mu}$ is true at every world in every model in $\mathfrak{B}$ that is of the form $\mathcal{M}^{\lambda+m,\mathcal{R}}$ for some model $\mathcal{M}$, revision scheme $\mathcal{R}$, limit ordinal $\lambda$ and natural number $m \geq n$. In other words, for a schema $\mu$ to be valid* with respect to $\mathfrak{B}$, each of its instances must stabilize to truth at every world on every $\omega$-sequence of revisions that belong to $\mathfrak{B}$.

The notion of derivability provided by $\mathbb{B}_1$ coincides with the weak notion of validity provided by all metastable Herzberger expansions of some model structure M.

**Proposition 4.** Let $\mathfrak{B}$ be the class of models consisting of all metastable Herzberger expansions $\mathcal{M}^{\alpha,\mathcal{R}}$ of some model structure M and any revision scheme $\mathcal{R}$. Then for every schema $\mu$ of PL, $\mathbb{B}_1 \vdash' \mu$ iff $\mathfrak{B} \models^* \mu$.

The left to right part of the proposition is established by induction on the length of derivations in $\mathbb{B}_1$. 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 $\mu$ is a schema that is not derivable in $\mathbb{B}_1$, there will be a revision scheme $\mathcal{R}$ and an expansion $\mathcal{M}$ of M such that $I_0(\mu)$ is false in $\mathcal{M}^{\alpha,\mathcal{R}}$ for arbitrarily large $\alpha$. This is so because for any schema $\mu$ that is "contingent in" $\mathbb{B}_1$ (i.e. neither provable nor disprovable in $\mathbb{B}_1$), one can, relying on a certain normal form for $\mu$, construct an expansion $\mathcal{M}$ of M so that $\mu$ is false in $\mathcal{M}^{\lambda+m,\mathcal{R}}$ at some world w, for arbitrarily large ordinals $\lambda$ and natural numbers m. The details of this construction and of the normal form are to be found in the proof of propositions 18 & 21 in Asher & Kamp (1987).

$\mathbb{B}_1$, however, is a minimal system. It provides only the most rudimentary formalization of the sort of reasoning that takes place for instance in the Knower paradox or other semantic paradoxes or that is represented by the process of model revision. Extensional systems that are

[1] The proof follows the strategy of theorem 21 in Asher & Kamp 1987.

stronger than $\mathbb{B}_1$ are easy to construct. In such extensional systems we can prove not only $B_0I$ but also the rule (R1). Moreover, there are extensional systems that replicate the sort of reasoning that goes on in the paradoxes. Consider for example a system which contains the following introduction and elimination rules. Here first is the introduction rule ($B_{ext}I$) (for extensional B predicates).

Suppose a proof contains the line:

(j) $\{l_1, ..., l_n\}$ $(n, \varphi)$

Then we may add the new line

($k_1$) $\{l_1, ..., l_n\}$ $(n + 1, B\varphi)$

Here is the corresponding elimination rule ($B_{ext}E$).

Suppose a proof contains the line:

(j) $\{l_1, ..., l_n\}$ $(n + 1, B\varphi)$

Then we may add the new line

(k) $\{l_1, ..., l_n\}$ $(n, \varphi)$

If one interprets B as a predicate of truth or some other extensional notion, then these rules reflect the conception of reasoning about these concepts captured by the process of extensional model revision. They encode in a proof system the process of reflection that many have taken to be the intuitive motivation behind the revision conception of truth. With $B_{ext}I$ and $B_{ext}E$, it becomes possible for the reasoner to reason about information in states that intuitively occur earlier in the revision process (i.e. states that are represented by lower numbers in the main entries). With this capacity, however, comes the possibility of strengthening a rule like RAA. For suppose an agent were to reason as follows: suppose that true($\varphi$) at stage n+1 of my reflections. But that is only possible if $\varphi$ is the case at stage n. If $\varphi$'s being the case at stage n leads me to a contradiction, then I should conclude $\neg$true($\varphi$) at n+1. Such reasoning appears perfectly acceptable, but the current $\mathbb{T}_0$ formulation of RAA does not permit it. So I introduce a strengthened version of RAA, RAA*, which says the following:

(j) $\{j\}$ $(n, \varphi)$

(k) $\{j, m_1, ..., m_j\}$ $(k, \psi \,\&\, \neg\psi)$

for $k \leq n$ and where the main entries on $m_1, ..., m_j$ are all of the form $(n, \delta_1), ..., (n, \delta_j)$.

Then we may write down as a new line:

(l) $\{m_1, ..., m_j\}$ $(n, \neg\varphi)$

Call the system that results from adding to $\mathbb{T}_0$ the rules. RAA*, ($B_{ext}I$) and ($B_{ext}E$), $\mathbb{B}_{ext}$. In $\mathbb{B}_{ext}$ all of the rules of $\mathbb{B}_1$ are derivable. But also derivable is the principle (E), $B(\varphi) \leftrightarrow \neg B(\neg\varphi)$. $\mathbb{B}_{ext}$ also has a clear semantics: it is sound and complete with respect to Herzberger metastable expansions of extensional model structures.[1]

**Proposition 5.** Let $\mathfrak{B}$ be the class of models consisting of all Herzberger metastable expansions $\mathcal{M}^{\alpha, \mathcal{A}}$ of some extensional model structure M. Then for every schema $\mu$ of PL, $\mathbb{B}_{ext} \vdash' \mu$ iff $\mathfrak{B} \models^* \mu$.

The left to right version of this proof proceeds generally by induction on the length of a proof. The only difficult rules are the new ones RAA* and ($B_{ext}I$) and ($B_{ext}E$). To show that these rules are sound, we need to take account of the way these rules exploit the structure of the revision sequence. To show that RAA* is valid*, I show that the conclusion drawn by RAA* must be a consequence of its premises in any metastable model $\mathcal{M}^{\alpha+m, \mathcal{A}}$ (for sufficiently high m), for if not then some other metastable model $\mathcal{M}^{\beta, \mathcal{A}}$ will verify a contradiction. Now define the *degree* of a formula $\nu$ of PL, *deg* ($\mu$)to be the maximum of the lengths of chains of nested occurrences of B in $\mu$. To show that ($B_{ext}I$) and ($B_{ext}E$) are sound, I show by induction on the degree of $\{\psi_1, ..., \psi_n\}$ that in a particular canonical form of proof and given a derivation of $(m, \varphi)$ from $(m, \psi_1), ..., (m, \psi_n)$ in which ($B_{ext}I$) and ($B_{ext}E$) are used, $(m, \psi_1), ..., (m, \psi_n) \models^* (m,\varphi)$. Going from right to left, the presence in $\mathbb{B}_{ext}$ of the equivalence, $\vdash' B(\psi)$ iff $\vdash' \psi$, and

[1]Note that (E) is not sound with respect to limit ordinal revision stages in the Herzberger revision process.

(E) allows us to rewrite any schema $\psi$ in terms of a boolean combination of T, $\perp$, sentence letters and formulas of the form $B^n(p_i)$.[1] Rewrite each one of the distinct sentence letters $p_j$ as a formula of the form $B^0(p_j)$. Now go through and replace each one of the formulas $B^m(p_k)$ with a distinct sentence letter $p^m{}_k$. Call such a rewrite of $\&(\Gamma)$ & $\varphi$, $\varphi^*$. Since $\varphi^*$ is not a theorem of $\mathbb{B}_1$, there is an assignment of truth values to the sentence letters of $\varphi^*$ that makes $\varphi^*$ false. The trick is now to find an interpretation of $\varphi^*$, $\underline{\varphi}^*$, in our first order language L such that $[\![\underline{\varphi}^*]\!]_{\mathcal{M}^\alpha, w} = 0$, for arbitrarily large $\alpha$ and for some w. The desired translation is one that maps the distinct sentence letters $p^m{}_k$ onto L-sentences of the form $B^m(c_k)$. Now we construct a model $\mathcal{M}$ in which the constants $\{c_1,..., c_n\}$ form a self-referential set in the sense of Asher & Kamp (1987) of the following form: $[\![c_0]\!] = B(c_1)$, $[\![c_1]\!] = B(c_2)$, ..., $[\![c_n]\!] = \neg B(c_0)$. The sentences denoted by the constants $c_1,..., c_n$ then cycle through all $2^n$ possible assignments of truth values to them in metastable models. This assures us that whatever assignment of truth values to sentence letters in $\varphi^*$ is required to show that $\varphi^*$ is false, there is a metastable model $\mathcal{M}^\alpha$ for arbitrarily large $\alpha$ which replicates this pattern in its assignment of truth values at w to the sentences $B^n(c_i)$.

$\mathbb{B}_{ext}$ also replicates the sort of reasoning that someone confronted with the paradoxes might proceed through, where now each ordinal n indexes a new mental state. Suppose that $[\![b]\!] = \neg B(b)$. I assume that the agent has a theory of the denotations of the constants in his language, so that given the denotation equations like the one above the agent knows that $b = c_{\neg B(b)}$, $c_{B(b)} = c_{\neg\neg B(b)} = c_{neg(b)}$ and so on. Although this notation for constants is correct, it is clumsy. So I will just use the subscripts for the constants where no confusion results; this makes things more readable. Now suppose that the reasoner might assumes for instance that (n, B(b)). But then applying $B_{ext}E$, (n -1, $\neg B(b)$). On the other hand, an application of ($B_{ext}I$) to the premise yields (n+1, B(B(b))), or, subtituting equals for equals, (n+1, B(neg(b))). But then by (E), (n+1, $\neg B(b)$). Another application of ($B_{ext}I$) and the substitution of equals for equals yields (n+2, B(b)). The familiar pattern of reasoning cycling between two possibilities is now established, and we may see such a pattern stretched out across as many mental states as the reasoner has energy for.

Because $\mathbb{B}_{ext}$ and systems to follow reflect the dynamics of belief revision, we should perhaps no longer think of completeness proofs relative to a class of models. Rather, the relevant notion of validity should be defined relative to a class of sequences of models. More precisely, let $\Gamma$ be a collection of pairs of formulas and numbers as in the main entries of proofs of $\mathbb{B}^0$, $\mathbb{B}^1$, or $\mathbb{B}_{ext}$. Then say that a sequence of models is *appropriate* for $\Gamma \cup \{(m, \varphi)\}$ just in case there is a bijective function f from ordinals in the main entries of the lines of $\Gamma \cup \{(m, \varphi)\}$ to $\{\beta_1, ..\beta_1+j\}$ such that: if $n_1$ is the minimal number in the main entries of $\Gamma \cup \{(m, \varphi)\}$, then $f(n_1) = \beta_1$; and since for every number n in the main entries of $\Gamma \cup \{(m, \varphi)\}$ can be written as $n_1 + \mu$, then set $f(n) = \beta_1 + \mu$. Say that $\Gamma \cup \{(m, \varphi)\}$ is *verified relative to an appropriate sequence* of models $\mathcal{M}^{\beta_1}$, ..., $\mathcal{M}^{\beta_1+j}$ if for every n and $\psi$ such that if $(n, \psi) \in \Gamma$ and $\mathcal{M}^{f(n)} \models \psi$, then $\mathcal{M}^{f(m)} \models \varphi$. I will write Prov($\Gamma$, (m, $\varphi$)) just in case there is a proof of (m, $\varphi$) from $\Gamma$.[2] An induction on the length of a proof that as in proposition 5 exploits the correspondence between the structure of model revision sequences and the rules of $\mathbb{B}_{ext}$ establishes the following proposition.

**Proposition 6.** Suppose that in $\mathbb{B}_{ext}$ Prov($\Gamma$, (m, $\varphi$)). Then any appropriate sequence of metastable expansions $\mathcal{M}^\beta$, $\mathcal{M}^{\beta+1}$, ... $\mathcal{M}^{\beta+n}$ of some extensional model-structure M verifies $\Gamma \cup \{(m, \varphi)\}$.

---

[1] Again, this method of proof is to be found in Asher & Kamp (1987). I have sketched it here to give a feel for some of the character of completeness proofs there.

[2] Note that if $\varphi$ is a theorem of an indexing system, then there is a least n such that $(n,\varphi)$ is derivable from the empty set.

Note that if in $\mathbb{B}_1$ Prov($\Gamma$, (m, $\varphi$)), any sequence of metastable expansions $\mathcal{M}^{\beta}$, $\mathcal{M}^{\beta+1}$, ... $\mathcal{M}^{\beta+n}$ also verifies $\Gamma \cup \{(m, \varphi)\}$. But $\mathbb{B}_{ext}$ has a completeness property that $\mathbb{B}_1$ does not. That is,

**Proposition 7.** Let $\Gamma$ be a collection of pairs of formulas of PL and numbers and let $\varphi$ be a formula of PL. Suppose that any appropriate sequence of metastable expansions $\mathcal{M}^{\beta}$, $\mathcal{M}^{\beta+1}$, ... $\mathcal{M}^{\beta+n}$ of some extensional model-structure M verifies $\Gamma \cup \{(m, \varphi)\}$. Then in $\mathbb{B}_{ext}$, Prov($\Gamma$, (m, $\varphi$)).

Note that proposition 7 will not hold for $\mathbb{B}_1$, because $\mathbb{B}_1$ does not allow us, for instance, to predict the "flip-flop" of truth value behavior of a liar sentence in a sequence of model revisions. To prove proposition 6 for $\mathbb{B}_{ext}$, consider the normal form of those formulae in $\Gamma \cup \{\varphi\}$ described in the sketch of the proof of proposition 5, where the set of distinct sentence letters is $\{p_1, ..., p_k\}$. Since (m, $\varphi$) is not derivable from $\Gamma$, $\Gamma \cup \{(m, \neg\varphi)\}$ is derivable consistent, in the sense that using the rules of $\mathbb{B}_{ext}$ we cannot derive a contradiction from this set. Suppose that there are sentence letters $\{p_j,..., p_m\}$ occurring in $\varphi$ and in $\psi_1,..., \psi_i$ in $\Gamma$. Call the set of interpretations of $\{p_j,...,p_m\}$ needed to verify $\psi_n$ $I_n$ and the set of interpretations of $\{p_j,...,p_m\}$ needed to falsify $\varphi$, I. Since (m, $\varphi$) is not derivable from $\Gamma$, there is a function $\mathcal{F}$ from one element of each of $I_1$, ..., $I_i$ to an element of I that obeys the constraints of extensional model revision. Otherwise, $\Gamma \cup \{(m. \neg\varphi)\}$ would not be derivable consistent. The numbers associated with $\psi_1$, ..., $\psi_i$ and $\varphi$ can be ordered into a sequence with which one may correlate L-models $\mathcal{M}^{\beta}$, ... $\mathcal{M}^{\beta+k}$. Now each sentence letter $p_i$ must according to the value for $\mathcal{F}$ remain true for so long in the sequence-- a parameter which I'll call the *period* of $p_i$; because of the nature of $\mathcal{F}$, this means that if on one of the assignments $\mu$ picked out by $\mathcal{F}$ $\mu(p_i) = T$ and the period of $p_i$ is m, then $\mu(B^{m+1}(p_i)) = \perp$. It is now possible to use the method alluded to in the sketch of the proof of proposition 5 to find right set of self-referential constants to ensure that the pattern of truth value assignments to $\{p_j,..., p_m\}$ is replicated by the translations of the $p_j$ in the sequence of models $\mathcal{M}^{\beta}$, ... $\mathcal{M}^{\beta+k}$. The remaining letters not shared by $\psi_1$, ..., $\psi_i$ and $\varphi$ may be assigned the appropriate values by the procedure sketched in proposition 5.

Extensional systems like $\mathbb{B}_{ext}$, however, will not do for an analysis of the attitudes. Principle (E), when applied to belief and other attitude predicates, gives intuitively wrong predictions; just because one does not believe that there is life on other planets, one does not have to believe that there is no life on other planets-- the concept of belief allows for agnosticism. In this respect $\mathbb{B}_1$ is superior. Similarly, $\mathbb{B}_{ext}$ fails to provide the foundations for a system of reasoning about other attitudes that interact with belief like want and desire. These too must be treated intensionally. But $\mathbb{B}_{ext}$ does capture an aspect of reasoning about belief that is also reflected in reasoning about liar sentences. From the perspective of the reasoner, reasoning about beliefs is similar to reasoning about truth. For an agent always assumes any one particular belief of his is true. Also, if he concludes a proposition is true, he should rationally come to believe it. Consequently, from the perspective of the agent, there is a kind of equivalence between truth and belief.[1] So an adequate theory of reasoning about belief should somehow incorporate elements of the rules ($B_{ext}$I) and ($B_{ext}$E).

To provide such a theory, I have to be more sophisticated about reasoning about belief and other attitudes. First, I will distinguish between "belief subproofs" and main proofs by introducing yet another modification to Lemmon's notion of proof. I shall say that line n occurs within a belief-subproof just in case it is of the form

(n.) $\{k_1, ..., k_m\}$ * (n, $\varphi$),

where n is the line number, $\{k_1, ..., k_m\}$ the set of numbers indicating the premises of the line, (n, $\varphi$) the main entry, and * the indicator announcing that n. is a line in a belief subproof. The notion of a belief subproof is familiar in natural deduction systems (see for instance Fitch

[1] I am endebted to Dan Bonevac for pointing out to me this way of seeing the intuitive characteristics of $\mathbb{B}_2$.

(1974) and Bonevac (1987)). Any line entered within a belief subproof expresses a proposition that the agent at least implicitly believes; the intuitive semantics of a belief subproof is that any line within such a subproof is true at all the belief alternatives of the agent. I now add two more rules ($B_1I$) and ($B_1E$), belief introduction and elimination. ($B_1I$) says the following: Suppose a proof contains the line:

(j) $\{l_1, ..., l_n\}$ * (n, φ)

Then we may add the new line

($k_1$) $\{l_1, ..., l_n\}$ (n + 1, Bφ)

The corresponding elimination rule ($B_1E$) says:

Suppose a proof contains the line:

(j) $\{l_1, ..., l_n\}$ (n + 1, Bφ)

Then we may add the new line

(k) $\{l_1, ..., l_n\}$ * (n, φ)

These rules follow the conception of belief revision described by model revision sequences; if at stage n φ occurs within a belief subproof (which corresponds to φ's being true at all alternatives to a world w at stage n), then in reflecting upon this fact, the agent may enter B(φ) into the main proof at the next stage (which corresponds B(φ)'s being true at w at stage n+1).

But in this reflection, the believer himself should come to another belief as well; that is, he should also come to believe implicitly B(φ). This means that B(φ) should also be entered into the belief subproof. This is an important aspect of the dynamics of belief and is essential in reasoning about the paradoxes, I will make the belief predicate "quasi-extensional" by allowing the rules ($B_{ext}I$) and ($B_{ext}E$) to operate on * lines. I'll call these rules ($B_{ext}I^*$) and ($B_{ext}E^*$) respectively. Here's ($B_{ext}I^*$).

Suppose a proof contains the line:

(j) $\{l_1, ..., l_n\}$ * (n, φ)

Then we may add the new line

($k_1$) $\{l_1, ..., l_n\}$ * (n + 1, Bφ)

The corresponding elimination rule ($B_{ext}E^*$) says the following.

Suppose a proof contains the line:

(j) $\{l_1, ..., l_n\}$ * (n + 1, Bφ)

Then we may add the new line

(k) $\{l_1, ..., l_n\}$ * (n, φ)

We also need some reiteration rules for entering new formulas within * subproofs. First all theorems may be reiterated within * subproofs (REIT*). That is, suppose (n+1, φ) is a theorem, then one may enter a new line,

(m) $\{\}$ * (n, φ)

One may do this regardless of the depth of * subproofs (i.e., there can be proofs with lines that have lots of *s on them). Call the resulting system that contains $\mathbb{T}_0$, (RAA*), ($B_1I$), ($B_1E$), ($B_{ext}I^*$), ($B_{ext}E^*$) and (REIT*) the system $\mathbb{B}_2$. I now introduce yet another notion of derivability, ⊢". $\psi_1, ..., \psi_n$ ⊢" φ, just in case there is an n such that for all m ≥ n there is a proof of (m, φ) from (m, $\psi_1$), ..., (m, $\psi_n$), and the line containing (m, φ) does not also contain any *. $\mathbb{B}_2$ contains $\mathbb{B}_0$ but not $\mathbb{B}_{ext}$. (E) is not derivable within $\mathbb{B}_2$. But within $\mathbb{B}_2$ it is easy to prove ($B_0I$). Note also that if ⊢" $\mathbb{B}_2$ B(φ) then ⊢" $\mathbb{B}_2$ (φ) for largely the same reasons as before. But again we can't prove this as a derived rule, so I will add (R1) to $\mathbb{B}_2$ as well.

Unlike $\mathbb{B}_1$, $\mathbb{B}_2$ also mirrors the dynamics of reasoning about belief that is reflected in the sequence of model revisions. For example, let us consider the sort of oscillations in the extension of B in a model $\mathcal{M}$ where $[\![b]\!]_{\mathcal{M}} = \neg B(b)$. The constant b denotes a belief-theoretic analogue to the liar sentence, I'll call it the "unbeliever". Let's assume that an agent has consistent beliefs and that he believes the unbeliever at least initially. He then comes to reflect on this belief using the system $\mathbb{B}_2$. Let's see what happens in the full L-framework, which will

have, I assume, a theory of the constants denoting sentences in it, so that we can make the appropriate substitutions.

| | | | | |
|---|---|---|---|---|
| 1. | {1} | | (n, B(b)) | A. |
| 2. | {1} | * | (n-1, ¬B(b)) | 1, $B_1E$ |
| 3. | {1} | * | (n, B(¬B(b)) | 2, $B_{ext}I^*$ |
| 4. | {1} | * | (n, B(b)) | 3, definition of b |
| 5. | {1} | * | (n, ¬¬B(b)) | 4, Double negation |
| 6. | {1} | | (n+1, B(neg(b))) | 5, $B_1I$ |
| 7. | {} | | (n+1, ¬B(⊥) → (B(neg(b)) → ¬B(b))), | theorem |
| 8. | {8} | | (n+1, ¬B(⊥)) | A. |
| 9. | {1,8} | | (n+1, ¬B(b)) | 6,7,8 Modus Ponens |
| 10. | {1} | * | (n+1, B(neg(b))) | 5, $B_{ext}I^*$ |
| 11. | {1,8} | * | (n+1, ¬B(b)) | same reasoning as in 7-9 using Reit |
| 12. | {1,8} | | (n+2, B(b)) | 11, $B_1I$ |
| . . | | | . . | |
| . . | | | . . | |

The familiar pattern is now once again established. Assuming beliefs are consistent, the agent won't be able to prove B(b) and ¬B(b) at any one stage, but his doxastic states will oscillate forever between these two possibilities-- affirming one at one stage and discarding it and affirming the contradictory at the next stage.

Surprisingly as proposition 8 shows, $\mathbb{B}_2$ and $\mathbb{B}_1$ share completeness properties with repsect to the static notion of validity, although the proof of the soundness of its rules also require restrictions on the alternativeness relation that is not needed in proving the rules of the minimal system $\mathbb{B}_1$ sound.[1] Proposition 9, whose proof follows that of propositions 6 & 7, shows that $\mathbb{B}_2$ captures the dynamic aspect of belief revision in a way that $\mathbb{B}_1$ does not.

**Proposition 8.** Let $\mathfrak{B}$ be the class of models consisting of all Herzberger metastable expansions $\mathcal{M}^{\alpha, f}$ of any model structure M where $R_M$ is transitive and euclidean.[2] Then for every schema μ of PL, $\mathbb{B}_2 \vdash'' \mu$ iff $\mathfrak{B} \models^* \mu$.

The proviso about transitive and euclidean alternativeness relations guarantees that for any $w \in W_M$ and w', w'' ∈ [$wR_M$], [w'R] = [w''R]. This is needed to show that the rules ($B_{ext}I^*$) and ($B_{ext}E^*$) are verified. Otherwise the inductive proof of proposition 7 proceeds along the lines of the one in proposition 5. The right to left direction of proposition 8 is established by the same method of proof as proposition 4; the presence of the rules if $\vdash'' \mathbb{B}_2 \varphi$ then $\vdash'' \mathbb{B}_2 B(\varphi)$ and (R1) permits the same proof strategy. The restrictions on the alternativeness relation of M do not obviate the possibility of constructing countermodels for schemas that are not theorems of $\mathbb{B}_2$.

**Proposition 9.** Γ be a collection of pairs of schemata of PL and numbers and suppose that φ is a schema of PL. Then any appropriate sequence of Herzberger metastable expansions $\mathcal{M}^{\beta, f}$, $\mathcal{M}^{\beta+1, f}$, ... $\mathcal{M}^{\beta+n, f}$ of some extensional model-structure M, where $R_M$ is transitive and euclidean, verifies Γ ∪ {(φ, m)} iff in $\mathbb{B}^2$, prov(Γ, (m, φ)).

---

[1] There are other notions of validity corresponding to $\mathbb{B}^2$; one is the notion of validity$_2$ defined in Asher & Kamp (1987).

[2] That is, $\forall w_1 \forall w_2 \forall w_3((w_1Rw_2 \ \& \ w_2Rw_3) \rightarrow w_1Rw_3)$ and $\forall w_1 \forall w_2 \forall w_3((w_1Rw_2 \ \& \ w_1Rw_3) \rightarrow w_2Rw_2)$.

There are stronger proof systems than $\mathbb{B}_2$. An easy one to consider is $\mathbb{B}_3$, which is like $\mathbb{B}_2$ except for the REIT rule. $\mathbb{B}_3$ allows the reiteration of formulae of the form $B\psi$. But it does not permit the reiteration of theorems of $\mathbb{B}_3$ within * subproofs-- only theorems of $\mathbb{T}_0$.[1] This permits the derivation of $B\varphi \rightarrow BB\varphi$ as a theorem. $\mathbb{B}_3$ is sound with respect to the class of metastable expansions $\mathcal{M}^{\lambda}$ of some any structure M where $R_M$ is transitive and euclidean for $\lambda$, a limit ordinal. However, $\mathbb{B}_3$ is not complete with respect to this class of models. Moreover, proofs of $\mathbb{B}_3$ are not verified by sequences of metastable models or indeed by any natural class of sequences of models. At best, $\mathbb{B}_3$ partially captures the logic of a particular stage of revision-- limit ordinal revision stages, and many have argued that the properties of these stages are an artifact of the Herzberger construction, not an intrinsic part of the concept of belief. Thus, $\mathbb{B}_3$ appears to take a step backwards in the quest for a system that captures the dynamics of reasoning about belief.

## 5. BELIEF AND TEMPORAL REASONING

The real motivation for the S4 principle for belief comes, I think, not from a system like $\mathbb{B}_3$ , but from another direction, which requires yet again a more sophisticated treatment of belief revision. More complicated rules for belief introduction and elimination are in general non-monotonic.[2] One very general pattern such rules may take is this: the agent infers new beliefs as "defaults" on the basis of partial information; later information forces him sometimes to revise these earlier beliefs inferred on the basis of partial information. Such rules introducing defaults should also be stated as "jump" rules for moving from one mental state to another-- similar in spirit to the simple belief introduction and elimination rules of $\mathbb{B}_0$.

These considerations force us to introduce explicitly into the framework some notion of time, although I have implicitly at least characterized belief revision as a process through time. The standard way to introduce time into an account of attitudes is to add a set of times T as another sort of index in the base model structure M. Thus, the alternativeness relation is now redefined as a binary relation on $W_M \times T_M$, the set of alternatives at $\langle w,t\rangle$ in M, $[\langle w,t\rangle R_M]$, is now redefined a set of world-time pairs, and $[\![\,]\!]$ is now redefined as a function from world-time pairs to suitable extensions for its arguments. "New" information concerning "B-free" facts, i.e. those facts whose statement involves no use of the B predicate, acquired by the agent at $t' > t$ can be characterized via the alternativeness relation: $[\langle w,t\rangle R_M] \neq [\langle w,t'\rangle R_M]$.

This familiar formalism permits an integration of a variety of temporal logics with logics of the attitudes, but it misses out on one very important aspect of reasoning about belief and time. That aspect concerns the principles by means of which agents reason about their own future and past mental states and the ways these principles may actually affect the contents of future mental states. That is, so far the formalism has yielded a way of characterizing the change in beliefs over time due to the acquisition of new information of B-free facts, but we have not addressed the problem of change in an agent's beliefs over time due to the acquisition of new information that may arise in reflection or reasoning about his beliefs.

---

[1] It would be nice to allow all theorems of first order logic to be reiterated within the * subproofs. But I cannot make such a rule within the restricted propositional framework used now for the sake of the completeness results. Since completeness is beyond the reach of this system in any case, however, one might as well work out the system for first order logic.

[2] See e.g. Doyle & McDermott (1980); Moore (1982), McDermott (1986)

One might suppose that in the absence of new information about B-free facts (and this would include facts about the agent that might undermine or strengthen justifications for holding beliefs), an agent should not change his beliefs. Certainly, an agent is not justified in changing his beliefs on a whim. So if the agent's information and evidence remain constant, his beliefs are all considered and rational and such weaknesses as forgetfulness and inattention are not at issue, his beliefs should remain constant. Such reasoning leads to the plausible principle of "knowledge maintenance," which says that once something is known it is known forever.[1] The belief maintenance principles says that once something is believed, it remains so until the doxastic facts for the agent change because of new externally acquired information. Suppose that such principles were adopted within the framework of the system $\mathbb{B}_2$, so that if $\varphi$ were believed at level m, then it would also be believed at m+1. The characteristic S4 axiom would then be a consequence of $\mathbb{B}_2$ and the knowledge or belief maintenance principles construed in this way.

These principles, as one might suspect however, are not sound. They treat the process of reflection as one of pure accumulation rather than revision. Nevertheless, they are plausible default principles whose scope of legitimate application is of interest. Further, the knowledge maintenance principle is of particular interest because one version of the Hangman Paradox appears to rely on its use. To recapitulate very briefly, the prisoner, K, relies on a chain of plausible inferences beginning with the judge's decree on Sunday that the prisoner is to be hanged at 6 am on one of the next two days but that he will not know on the basis of this decree which day it will be until 30 minutes before the hanging. K supposes that the decree is true, as he has every reason to do, and then supposes, Sunday evening say, that perhaps he will be hanged on Tuesday. But that is impossible, since by Monday 6:01 am he will know on the basis of the decree when he to be hanged. So then he concludes he must be hanged on Monday, but again he concludes that is impossible according to the decree. This leads him to conclude that the decree is false. He is then quite surprised at 5:30 Monday morning when he is awakened by the hangman to be taken to the gallows!

In this bout of reasoning, K reflects about future doxastic states as well as his present one. The inference that he won't be hanged on Tuesday depends crucially on the assumption that he still believes or knows the decree to be true on Monday at 6:01 am; it is the principle of knowledge or belief maintenance that justifies this assumption. Similarly, in order to conclude that the decree is false, K must continue to suppose that he believes the decree, when he reflects on the state in which he has concluded that he cannot be hanged on Monday (call this $\mathbf{B}_1$) and so concludes that he knows that he cannot be hanged on Monday (in state $\mathbf{B}_2$). Whether one wants to postulate here a difference in time between $\mathbf{B}_1$ and $\mathbf{B}_2$ is perhaps controversial, but we have already seen reason enough to distinguish between these states, and something like the principle of knowledge maintenance preserves the belief in the decree. If this analysis is correct, it shows that reflective reasoning about future mental states may, even in the absence of new, B-free or K-free information, change the beliefs the agent presently holds. This happens in the presence of paradoxical propositions like the judge's decree. What is needed now is an account of how this is possible.

To make explicit the connection between reflective belief revision and temporality, certain stages of the revision process have to be correlated with times in some way. If times have already been introduced as an independent parameter, then the model revision process must be redefined so as to assign the model revisions times, so that in the interpretation of various

---

[1] I should perhaps add to the principle of knowledge maintenance the proviso, 'as long as the B-free facts do not change'. For one could imagine that a change in the B-free facts could change some of the beliefs that the agent uses to justify his knowledge claim and so undermine it. For yet another puzzle in which the principle of knowledge maintenance suffers for perhaps these reasons, see Fred Dretske's discussion of zebras and cleverly disguised mules in Dretske (1970).

temporal operators or predicates, the appropriate model revisions that designate future or past belief states of the agent may be used. But if situations in which the B-free facts do change are not the primary interest (as is the case here), then a simpler solution, though perhaps an ultimately unsatisfactory one, is available. That solution uses the indices already present in the model theory of revision (or perhaps only certain designated ones)[1] to stand for the belief states of the agent at various times. In order to mimic the notion of later information coming to an agent's beliefs from the outside, one might appeal to a notion like *model perturbation.*. A *model perturbation of* a model $\mathcal{M}$ *at a world* w is a function on $R_{\mathcal{M}}$ and w that reassigns R alternatives to w. A model perturbation of a model $\mathcal{M}$ may introduce "momentary incoherence" between model revisions-- incoherence, because $[\![B]\!]_{\mathcal{M}, w}$ may no longer reflect the doxastic facts encoded by $[wR_{\mathcal{M}}]$ after the perturbation of $\mathcal{M}$, but only momentary incoherence, because after one revision coherence is restored. One useful type of model perturbation to consider is a *monotone decreasing* perturbation $\mathcal{F}$ of the alternativeness relation in $\mathcal{M}$ at w such that $\mathcal{F}(R_{\mathcal{M}}, w) \subseteq [wR_{\mathcal{M}}]$.[2] Using the notion of a model perturbation, we might rephrase the belief maintenance principle as follows: once something is believed it remains believed in the absence of non-monotonically-decreasing model perturbations.

I will follow out this simplistic approach to introducing time explicitly into the reasoning characterized by the model revision process. I will restrict myself here to just Herzberger revision sequences, since they are the simplest to understand. I will drop all explicit reference to revision schemes for the remainder of the paper. I will introduce certain predicates or operators into the language L(B) concerning these temporal moments as I have interpreted them. To be specific I introduce $\boxdot$ and $\triangledown$ as 1-place predicates to L(B); their intuitive interpretation is 'forever' and 'next' respectively. With these temporal predicates, the knowledge maintenance principle is expressible as follows:

(KM) $\quad K(\varphi) \rightarrow \boxdot(K(\varphi))$

These predicates alone, however, will not be sufficient to reproduce the complete Hangman argument. Although we could use temporal predicates to express the Hangman argument, it is most easily expressed in a tensed predicate calculus. Indeed, a goal of the theory of reasoning about belief and time is the interpretation of the tensed predicate logic within this framework and along the lines I have proposed for the temporal predicates. But this promises to be a complex task,[3] and a simpler temporal theory is useful to sort out basic issues; so I won't embark on an interpretation of full predicate logic here.

Given that these predicates talk about the stages of model revision, they must be interpreted with care. In particular, we cannot, without falling into contradiction, adopt the following interpretation, which parallels our earlier interpretation of the predicate B:

$$\varphi \in [\![\boxdot]\!]^{\alpha}_{\mathcal{M},w} = 1 \text{ iff } \forall \beta \geq \alpha\ [\![\varphi]\!]^{\beta}_{\mathcal{M},w} = 1$$

Rather what we must do is to complicate the notion of model revision considerably. $\boxdot$ and $\triangledown$ must be evaluated with respect to an entire *sequence of model revisions.* How long should this sequence be? One natural sequence that I will use here is the one defined by the original model revision process concerning the predicate B up to the second perfect stabilization ordinal. Because of Herzberger's (1982) grand cycle theorem, this includes the entire range of distinct metastable models. Perhaps the most straightforward implementation of this idea is to suppose that the sequence of model revisions is indexed to two ordinals to produce the

---

[1]It is natural to consider only successor stages in a revision sequences as naturally correlated with times. It is indeed possible to do so and preserve all the machinery developed here, but in the interests of simplicity if not intuition, temporal moments will be simply correlated with ordinals.

[2]There are model perturbations that bear out our predictions and verify our default rules of belief revision-- these are certain kinds of monotone decreasing perturbations. But there are others (even monotone decreasing sequences) that do not and which then force the revision of those predictions.

[3]Briefly, we would want to map time variables onto functions of ordinals in such a way that the restrictions on the time variables were satisfied relative to the ordinal indexes of the model revisions.

following sort of sequence: $\mathcal{M}^{0,0}$, $\mathcal{M}^{1,0}$, ..., $\mathcal{M}^{\alpha_0,0}$, $\mathcal{M}^{0,1}$, $\mathcal{M}^{1,1}$..., $\mathcal{M}^{\alpha_1,1}$,...., $\mathcal{M}^{0,\gamma}$, ..., $\mathcal{M}^{\alpha_\gamma,\gamma}$, where $\alpha_0$, ...., $\alpha_\gamma$ are second perfect stabilization ordinals. I define the extension of B largely as before:

i) $[\![B]\!]^{0,\beta}_w = [\![B]\!]_w$

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

iii) For limit ordinal $\lambda$, $[\![B]\!]^{\lambda,\beta}_w = \{\varphi: (\exists \delta<\lambda)(\forall \gamma)(\delta \leq \gamma < \lambda \rightarrow \varphi \in [\![B]\!]^{\gamma,\beta}_w)\}$.[1]

The extension of $\boxdot$ is now defined as follows:

i) $[\![\boxdot]\!]^{\delta,0}_w = [\![\boxdot]\!]_w$

ii) $[\![\boxdot]\!]^{\delta,\beta+1}_w = \{\varphi: (\forall \gamma) (\alpha_\beta \geq \gamma \geq \delta \rightarrow [\![\varphi]\!]_{\mathcal{M}^{\gamma,\beta}, w} = 1)\}$

iii) For limit ordinal $\lambda$, $[\![\boxdot]\!]^{\delta,\lambda}_w = \{\varphi: (\exists \delta<\lambda)(\forall \gamma)(\delta \leq \gamma < \lambda \rightarrow \varphi \in [\![\boxdot]\!]^{\delta,\gamma}_w)\}$.

The definition for $\triangledown$ is analogous. These definitions assure us that there are $\alpha_{1,0}$, ...., $\alpha_{1,\gamma}$ second perfect stabilization ordinals and that the sequence of models is well-defined. The model revision process with respect to the first model index is just as before.

This interpretation of the temporal predicates is imperfect to be sure. For one thing it validates the following, very strong commutativity principles: $\boxdot(B(\varphi)) \leftrightarrow B(\boxdot(\varphi))$ and $\triangledown(B(\varphi)) \leftrightarrow B(\triangledown(\varphi))$, in the absence of any model perturbations. So in the absence of new information, this interpretation entails that agents be omniscient with respect to the futures (and if one wishes also the pasts) of their doxastic possibilities. Recall, however, that I am not trying to describe agents' actual capacities for reasoning about belief and time. I am trying to describe the correct principles for such reasoning. Viewed from this perspective, the commutativity principles are no more egregious than logical omniscience: they merely entail that in the absence of new information, it is legitimate and consistent to follow out the consequences of ones temporal beliefs! But of course shifts in the underlying doxastic possibilities due to the acceptance of new information by an agent are such a prevalent part of our mental life that it is difficult to think of how our habits of prediction, planning and reevaluation should fare under the absence of it.[2]

The temporal predicate $\boxdot$ in effect is a predicate of stable truth within this model-theoretic framework. It allows us to construct a temporal analogue to the extended liar: consider the denotation equation $[\![b]\!] = \neg\boxdot(b)$. The sentence $\neg\boxdot(b)$ induces an analogous pattern of instability in the extension of the predicate $\boxdot$, as the unbeliever does for the predicate B. That is, suppose, for example, that $b \notin [\![\boxdot]\!]^{0,0}_w$ in some particular model. By the semantics for $\boxdot$, $b \notin [\![\boxdot]\!]^{\beta,0}_w$, for all $\beta \leq \alpha_0$. But then $b \in [\![\boxdot]\!]^{\beta,1}_w$, for all $\beta \leq \alpha_1$. Again by the semantic definitions, $b \notin [\![\boxdot]\!]^{\beta,0}_w$, for all $\beta \leq \alpha_2$, and so on. Thus, $\boxdot$ is a predicate that behaves in the predictable patterns already explored in the literature on the type-free semantics of truth and the attitudes.

The interpretation of $\boxdot$ also forces a redefinition of certain basic concepts. For the temporal predicates, as well as attitude predicates, now introduce instabilities into the process of model revision. I will distinguish between *local* stability and *global* stability; the local/global distinction applies to the definitions of stabilization, perfect stability and metastability as well. All these definitions assume a model revision sequence with no model perturbations. $\varphi$ is

---

[1] This simplification of course follows from the definition of the Herzberger revision scheme.

[2] The absence of shifts in the set of alternatives can make quite a difference. Without them we seem to get into certain puzzles about the "next moment." Suppose I believe that in the next moment I will fall asleep or that in the next moment I will die, examples suggested by Dan Bonevac. According to the principles advanced here, in the next moment I must believe that I am asleep or believe that I am dead! Something is amiss here-- in fact two things. First, for real life agents, there are not always "next" cognitive states; the sequence of revisions abstracts away from this limitation and correspondingly suffers when we countenance beliefs that involve the lack of a next cognitive state. The other element that gives these beliefs a sort of paradoxical air is the lack of any changing of set of alternatives. If I don't die in the next moment, my set of alternatives will surely change, which will account for the fact why I don't believe in that next moment that I am dead.

*positively (negatively) locally stable in* a model $\mathcal{M}$ *at* a world w iff for some $\beta$ $\varphi \in [\![B]\!]^{\gamma,\beta}_{\mathcal{M},w}$ for all $\gamma$ ($\varphi \notin [\![B]\!]^{\gamma,\beta}_{\mathcal{M},w}$). I will also use the terminology $\varphi$ is *locally stable in* a model $\mathcal{M}$ *at* a world w *with respect to* an ordinal $\beta$ iff for all $\gamma$ $\varphi \in [\![B]\!]^{\gamma,\beta}_{\mathcal{M},w}$ or $\varphi \notin [\![B]\!]^{\gamma,\beta}_{\mathcal{M},w}$. $\varphi$ *locally stabilizes at* an ordinal $\alpha$ *in* a model $\mathcal{M}$ (*at* a world w) *with respect to* an ordinal $\beta$ iff $\alpha$ is the first ordinal $\gamma$ such that $\varphi$ is positively or negatively stable (at w) in $\mathcal{M}^{\gamma,\beta}$. The other definitions of local stabilization ordinal, local perfect stabilization ordinal and local metastable model follow straightforwardly.[1]

The generalization of local stability is *global stability*. $\varphi$ is *positively (negatively) globally stable in* a model $\mathcal{M}$ *at* a world w iff for all $\gamma$ and for all $\beta$ $\varphi \in [\![B]\!]^{\gamma,\beta}_{\mathcal{M},w}$ ($\varphi \notin [\![B]\!]^{\gamma,\beta}_{\mathcal{M},w}$). $\varphi$ *globally stabilizes at* an ordinal $\alpha$ *in* a model $\mathcal{M}$ (*at* a world w) iff $\alpha$ is the first ordinal $\beta$ such that $\forall\gamma$ $\varphi$ is positively or negatively stable (at w) in $\mathcal{M}^{\gamma,\beta}$. $\alpha$ is a *global stabilization ordinal for* $\mathcal{M}$ (*at* w) iff every $\varphi$ that stabilizes in $\mathcal{M}$ (at w) stabilizes at some ordinal $\leq \alpha$ in $\mathcal{M}$ (at w). Call $\gamma$ a *global perfect stabilization ordinal for* $\mathcal{M}$ just in case $\gamma$ is a stabilization ordinal for $\mathcal{M}$, and $\varphi \in [\![B]\!]^{\delta,\gamma}_{\mathcal{M},w}$ iff $\varphi$ stabilizes at some ordinal $\leq \gamma$ in $\mathcal{M}$ at w. If $\beta$ is any ordinal greater or equal to the first perfect stabilization ordinal for $\mathcal{M}$, the model $\mathcal{M}^{\beta}$ is called a *globally metastable* model.

Given these definitions, it is obvious that if $\varphi$ is globally stable throughout a class of models $\mathfrak{B}$, then it is also locally stable throughout $\mathfrak{B}$ for any ordinal $\delta$. The converse of course isn't true. Even further, local stability at $\alpha$ for all $\alpha$ does not entail global stability. Global and local stability may affect each other too in subtle ways. A sentence may be locally stable at w in $\mathcal{M}$ with respect to $\delta$ but not with respect to some other ordinal. Consider for instance, the following set of denotation equations: $[\![c]\!]_M = \neg\Box(b) \vee \neg B(d)$, $[\![b]\!]_M = \neg\Box(b)$ and $[\![d]\!]_M = \neg B(d)$. It may indeed be the case that b is true at w in $\mathcal{M}^{0,\delta}$ but false at w in $\mathcal{M}^{0,\delta+1}$. Then c is stably true at w in $\mathcal{M}^{0,\delta}$ and so locally stable at w in $\mathcal{M}$ with respect to $\delta$ but not stable at w in $\mathcal{M}^{0,\delta+1}$.

But how do these instabilities now affect reasoning about belief or other attitudes? Having already ascertained that the locally metastable models whose alternativeness relation is euclidean and transitive correspond to a certain weak form of reasoning about belief simpliciter, one natural, but quite restricted class of models to consider as defining the logic of reasoning about belief and time is to consider the class, MP, of locally and globally metastable models of the form $\mathcal{M}^{\gamma,\alpha}$, where $\alpha$ is a global, perfect stabilization ordinal. Another class, ML, of models to consider consists of locally and globally metastable models of the form $\mathcal{M}^{\gamma,\alpha}$, where $\alpha$ is a limit ordinal. Both the model classes MP and ML validate the following axiom schemata and rules for $\Box$:

(F1) $\Box\varphi \rightarrow \varphi$
(F2) $(\Box(\varphi \rightarrow \psi) \,\&\, \Box\varphi) \rightarrow \Box\psi$
(F3) $\Box\varphi \rightarrow \Box\Box\varphi$
(F4) $\Box\varphi \rightarrow \neg\Box\neg\varphi$
(F5) If $\varphi$ is a theorem of first order logic, then $\Box\varphi$ is a theorem

Together with the commutativity rules, these provide a relatively acceptable logic for the temporal predicate $\Box$. These axioms, given the intended interpretation of $\Box$ in this restricted set up, seem acceptable. At locally and globally metastable models that are not in MP or ML, the

---

[1] $\alpha$ is a *local stabilization ordinal for* $\mathcal{M}$ (*at* w) iff every $\varphi$ that stabilizes in $\mathcal{M}$ (at w) stabilizes at some ordinal $\leq \alpha$ in $\mathcal{M}$ (at w). Call $\gamma$ a *local perfect stabilization ordinal for* $\mathcal{M}$ just in case $\gamma$ is a stabilization ordinal for $\mathcal{M}$, and $\varphi \in [\![B]\!]^{\gamma,\beta}_{\mathcal{M},w}$ iff $\varphi$ stabilizes at some ordinal $\leq \gamma$ in $\mathcal{M}$ at w and for some $\beta$. If $\beta$ is any ordinal greater or equal to the first perfect stabilization ordinal for $\mathcal{M}$, the model $\mathcal{M}^{\beta}$ is called a *local metastable* model.

logic differs; in particular (F1) and (F3) are no longer valid in all such models.[1] The semantics of $\nabla$ is such that all locally and globally metastable models validate the following principles:

(N1) $\nabla\neg\varphi \leftrightarrow \neg\nabla\varphi$
(N2) $(\nabla(\varphi \rightarrow \psi) \,\&\, \nabla\varphi) \rightarrow \nabla\psi$
(N3) if $\vdash''\ \varphi$ then $\vdash''\ \nabla\varphi$.

Within the classes ML or MP, however, the rule if $\vdash \varphi$ then $\vdash \Box(\varphi)$ will not be valid, and neither will it be the case that if $\vdash \Box(\varphi)$ then $\vdash \varphi$. This complicates considerably efforts to come up with completeness results for axiom schemata for $\Box$ and $\nabla$. In particular the method used in propositions 4 and 8 is not available.

The nature of the counterexamples to the knowledge maintenance principle and belief maintenance principle in the absence of non-monotonically-decreasing model perturbations can now be made more precise.

**Proposition 10.** Let $\mathfrak{B}$ be a class of globally metastable expansions of some model structure M such that $R_M$ is transitive and euclidean. Then: (i) if $\varphi$ is globally stable throughout $\mathfrak{B}$, then $\mathfrak{B} \models B(\varphi) \rightarrow \Box(B(\varphi))$; (ii) if $\mathfrak{B} \models B(\varphi) \rightarrow \Box(B(\varphi))$, then $\varphi$ is locally stable throughout $\mathfrak{B}$. Moreover, if $\varphi$ is not locally stable in $\mathfrak{B}$, then there are counterinstances to $B(\varphi) \rightarrow \Box(B(\varphi))$ at arbitrarily large successor, limit and perfect stabilization ordinals.

Proof sketch: Obviously if $\varphi$ is globally stable throughout $\mathfrak{B}$, then $\mathfrak{B} \models B(\varphi) \rightarrow \Box(B(\varphi))$. To show (ii), suppose that $\mathfrak{B} \models B(\varphi) \rightarrow \Box(B(\varphi))$. But suppose that $\varphi$ is not locally stable throughout $\mathfrak{B}$. So for some w and $\mathcal{M}$ in $\mathfrak{B}$, $[\![\varphi]\!]_{\mathcal{M}\beta,\gamma,w} = 1$ and $[\![\varphi]\!]_{\mathcal{M}\beta+1,\gamma,w} = 0$ for arbitrarily large $\beta$ and for all $\gamma$. Since $\varphi$ is locally unstable that means that w must belong to the range of R; since otherwise, all sentences would eventually locally stablize everywhere. Suppose $w \in [w_1R]$. Because $R_{\mathcal{M}}$ is transitive and euclidean, $\varphi$ will have for sufficiently large $\alpha$ the same truth value at all $w' \in [w_1R]$. So if we pick a sufficiently large $\beta$, $[\![\varphi]\!]_{\mathcal{M}\beta,\gamma+1,w'} = 1$ for all $w' \in [w_1R]$. So $[\![B\varphi]\!]_{\mathcal{M}\beta+1,\gamma+1,w'} = 1$. Note, however, that it won't be the case that for all $\delta \geq \beta+1$, $[\![B(\varphi)]\!]_{\mathcal{M}^{\delta,\gamma},w_1} = 1$, since $\varphi$ is locally unstable at w. So $\Box B\varphi$ will be false in $\mathcal{M}^{\alpha+1,\gamma+1}$ at w. We have shown if $\varphi$ is locally unstable at w for arbitrarily large successor ordinals $\alpha$, then $B(\varphi) \rightarrow \Box(B(\varphi))$ is false at w in $\mathcal{M}^{\alpha,\gamma}$. Let $\beta$ be a globally perfect stabilization ordinals, then since $B(\varphi) \rightarrow \Box(B(\varphi))$ is false at w at $\mathcal{M}$ at $\mathcal{M}^{\alpha,\gamma}$ for $\gamma > \alpha$, then $B(\varphi) \rightarrow \Box(B(\varphi))$ is false at w in $\mathcal{M}^{\alpha,\beta}$. Finally, suppose that $\beta$ is a limit ordinal. Because $\varphi$ is locally unstable at w, $\varphi$ never stabilizes on any end segment of $\beta$ and so then $B(\varphi) \rightarrow \Box(B(\varphi))$ is false at w in $\mathcal{M}^{\alpha,\beta}$.

Proposition 10 shows the knowledge and belief maintenance principles to have, in the absence of model perturbations, much the same status as the axiom for belief $B(B(\varphi \rightarrow \varphi)$ in proposition 3. As we have seen, such principles are not valid, though for reasoning about ordinary propositions they may well be efficient. They are sound when applied to knowledge of "normal" statements or propositions-- i.e., propositions that eventually stabilize under revision (of a knowledge predicate). The temporal and non-temporal versions of the Hangman Paradox also fall under a single theme: the agent employs plausible default principles of reasoning about knowledge and belief on a proposition that fails to stabilize; and in reasoning about his knowledge of this proposition, the agent undermines his very knowledge of it. But to locate the source of the problem with reasoning about paradoxical statements in this way has required the development of a dynamic theory of reasoning about belief and knowledge. I have argued that the process of model revision captures at least some of the significant aspects of such a dynamic theory of reasoning. The proof systems developed in section 4 constitute a first attempt to employ the insights of the process of model revision within a formal calculus for reasoning about belief and knowledge.

---

[1] There are other options for the logic of $\Box$ of course. I leave to another time the question of sorting out which of these is worth pursuing. Also I won't pursue the question of whether to use these axioms within the systems of natural deduction or whether to adopt natural deduction analogues of them.

first attempt to employ the insights of the process of model revision within a formal calculus for reasoning about belief and knowledge.

## Acknowledgments

Many of the results in this paper have developed from joint work with Hans Kamp on the semantic paradoxes for belief. In particular theorems 1, 2, 3 and the theorem in footnote 1 on page 5 are the result of our joint work. I am also endebted to him for many specific suggestions. Thanks also go to Anil Gupta and Dan Bonevac who also contributed several suggestions. The responsibility for any errors, of course, is completely mine.

## References

N. Asher & H. Kamp: 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.

N. Asher & H. Kamp: 1987, 'Self-Reference, Attitudes and Paradox,' forthcoming in the Proceedings of the 1986 Conference on Property Theory at The University of Massachusetts at Amherst.

D. Bonevac: 1987, *Deduction*, Palo Alto: Mayfield Press.

J. Burgess: 1986, 'The Truth is Never Simple', *Journal of Symbolic Logic* 51, pp. 663-681.

J. Doyle & D. McDermott: 1980, 'Non-Monotonic Logic I', *Artificial Intelligence*, 13, pp. 41-72.

F. Dretske: 1970, 'Epistemic Operators', *Journal of Philosophy*, vol. 67, pp. 1007-1023.

F. Fitch: 1974, *Elements of Combinatory Logic*, New Haven: Yale University Press.

A. Gupta: 1982, 'Truth and Paradox,' *Journal of Philosophical Logic* 12, pp. 1-60.

H. Herzberger: 1982, 'Notes on Naive Semantics,' *Journal of Philosophical Logic* 12, pp. 61-102.

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

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

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

E. J. Lemmon: 1968, *Beginning Logic*, Indianapolis: Hackett.

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

R. Moore: 1983, 'Semantical Considerations on Nonmonotonic Logic,' SRI Technical Note.

N. Rescher: 1976, *Plausible Reasoning: An Introduction to the Theory and Practice of Plausibilistic Inference*, Assen: Van Gorcum.

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