10

DOXASTIC LOGIC

Michael Caie

There are at least three natural ways of interpreting the object of study of doxastic logic. On one construal, doxastic logic studies certain general features of the doxastic states of actual agents. On another construal, it studies certain general features of *idealized* doxastic states. While on yet another construal, doxastic logic provides a normative account of what features an agent's doxastic state ought to have on pain of irrationality.

The field of doxastic logic was initiated by Hintikka (1962), where techniques from modal logic were employed to model doxastic and epistemic states and to characterize certain valid principles governing such states. The theory presented by Hintikka provides an account of certain synchronic principles governing doxastic states. That theory, however, is silent on the question of how the doxastic states of an agent over time are or should be related. Later work, initiated by Alchourrón, Gärdenfors, and Makinson (1985), sought to provide an account of how an agent will or should revise her doxastic state in the light of new evidence. According to the accounts developed out of Alchourrón et al., a characterization of an agent's doxastic state should include, in addition to the set of beliefs the agent has, a characterization of the agent's belief revision policy.

One of the characteristic features of the models developed by Hintikka is that they provide a natural way of modeling the higher-order beliefs of an agent, i.e., the agent's beliefs about her own beliefs. The models developed out of Alchourrón et al., however, don't provide a natural way of characterizing an agent's higher-order beliefs about her belief revision policy. More recent work in dynamic doxastic logic has attempted to remedy this defect by providing a semantics for an object language that includes not only a unary belief operator but also a binary belief revision operator.

In Section 1, I'll outline the theory developed by Hintikka and briefly discuss how this theory looks given each of the above construals. In Section 2, I'll consider the theory of belief revision that developed out of Alchourrón et al. In Section 3, I'll discuss more recent work in dynamic doxastic logic. And, finally, in Section 4, I'll consider some paradoxes of doxastic logic and the bearing that these have on some of the accounts considered in Section 1–3.

1 STATIC DOXASTIC LOGIC

In Section 1.1, I'll first provide a quick overview of the basic theory developed by Hintikka (1962). In Section 1.2, I'll then consider how these doxastic models may be extended to characterize the doxastic states of multiple agents and various collective doxastic properties. The presentation of this material will work under the assumption that the theory serves to characterize certain features of an *idealized* doxastic state. Having outlined the basic theory, however, in Section 1.3, I'll consider how the theory looks under alternate interpretations.

1.1 *Basic Doxastic Logic*

Let $\mathcal{L}$ be a propositional language. We assume that $\mathcal{L}$ includes the Boolean connectives $\neg$ and $\vee$, and in addition a unary operator B_α. The intuitive gloss of B_α will be "Alpha believes that…". Other connectives may be defined in the standard manner. Being a sentence of $\mathcal{L}$ is characterized as follows.

- If ϕ is an atomic propositional sentence letter, then ϕ is a sentence.
- If ϕ and ψ are sentences, then so is $\phi \vee \psi$.
- If ϕ is a sentence, then so are $\neg\phi$ and $B_\alpha\phi$.
- Nothing else is a sentence.

A *Kripke model* for our language $\mathcal{L}$ is a tuple $M = \langle W, R_\alpha, [\![\cdot]\!]\rangle$. W is a set of points that we'll call *possible worlds*. R_α is a binary relation on W, i.e., $R_\alpha \subseteq W \times W$, that we'll call the *accessibility relation*. And $[\![\cdot]\!]$ is the *interpretation function* mapping propositional letters to sets of possible worlds.

We can think of the accessibility relation R_α as serving to represent the set of worlds that are doxastic possibilities for an agent α relative to some world w. In particular, if w' is such that $\langle w, w'\rangle \in R_\alpha$, then we can think of w' as being a possible world that is left open given all that the agent believes at w.

The truth of a sentence ϕ at a world w in a Kripke model M (for short: $[\![\phi]\!]^w_m = 1$) may be defined as follows.

- If ϕ is a propositional letter, then $[\![\phi]\!]^w_m = 1$ just in case $w \in [\![\phi]\!]$.
- $[\![\neg\phi]\!]^w_m = 1$ just in case $[\![\phi]\!]^w_m \neq 1$.
- $[\![\phi \vee \psi]\!]^w_m = 1$ just in case $[\![\phi]\!]^w_m = 1$ or $[\![\psi]\!]^w_m = 1$.

- $[\![B_\alpha \phi]\!]^w_m = 1$ just in case $[\![\phi]\!]^{w'}_m = 1$, for every w' such that $wR_\alpha w'$.

We let $\vdash_x \phi$ mean that that there is a sequence of formulas $\phi_1, \ldots, \phi$ such that each item in the sequence is either an axiom of the logical system X or follows from items earlier in the sequence by one of the inference rules of X. Then $\vdash_k$ may be characterized as follows.

AXIOMS OF K

(P) Axioms of propositional logic

(K) $B_\alpha(\phi \to \psi) \to (B_\alpha \phi \to B_\alpha \psi)$.

INFERENCE RULES OF K

(MP) $(\vdash_k \phi \ \wedge \ \vdash_k \phi \to \psi) \Rightarrow \vdash_k \psi$.

(N) $\vdash_k \phi \Rightarrow \vdash_k B_\alpha \phi$.

Let $\mathcal{K}$ be the class of Kripke models, and let $\models_{\mathcal{K}} \phi$ mean that, for every Kripke model M, and every $w \in W$, $[\![\phi]\!]^w_m = 1$. Then two basic results in modal logic are:[1]

THEOREM 1. $\vdash_k \phi \Rightarrow \models_{\mathcal{K}} \phi$.

THEOREM 2. $\models_{\mathcal{K}} \phi \Rightarrow \vdash_k \phi$.

THEOREM 1 tells us that $\vdash_k$ is sound with respect to the class of models $\mathcal{K}$, and THEOREM 2 tells us that $\vdash_k$ is complete with respect to this class of models.

The first assumption that we'll make is that the beliefs of an *idealized* doxastic state may be represented by a Kripke model. Given the soundness of $\vdash_k$, it follows from this assumption that:

(B_N) if ϕ is a logical validity, then ϕ is believed by an idealized doxastic agent;

(B_K) an idealized doxastic agent believes all of the logical consequences of her beliefs.

If, in addition, we assume that, for any model $M \in \mathcal{K}$, there is some idealized doxastic state that is represented by M, then the soundness and completeness of $\vdash_k$ entail that (B_N) and (B_K) provide a complete characterization of those properties that are shared by every idealized doxastic state.

We can characterize certain subsets of $\mathcal{K}$ by properties of R_α.

1 For proofs of these results see, e.g., Hughes and Cresswell (1996), Chellas (1980), or Blackburn, de Rijke, and Venema (2001).

DEF. We say that R_α is *serial* just in case, for every $w \in W$, there is some $w' \in W$ such that $wR_\alpha w'$

DEF. We say that R_α is *transitive* just in case, for every $w, w', w'' \in W$, if $wR_\alpha w'$ and $w'R_\alpha w''$, then $wR_\alpha w''$.

DEF. We say that R_α is *Euclidean* just in case, for every $w, w', w'' \in W$, if $wR_\alpha w'$ and $wR_\alpha w''$, then $w'R_\alpha w''$.

We'll let $\mathcal{K}_D$ be the subset of Kripke models whose accessibility relation is serial, $\mathcal{K}_4$ the subset of Kripke models whose accessibility relation is transitive, and $\mathcal{K}_5$ the subset of Kripke models whose accessibility relation is Euclidean.

Assume that $\mathcal{L}$ has only propositional letters p and q. Then we can represent a Kripke model for $\mathcal{L}$ by a diagram like Figure 1. In this model

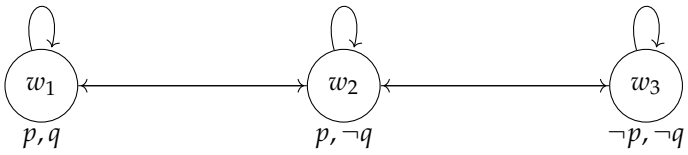

Figure 1: Diagram of a Kripke model

R_α is reflexive and serial, but neither transitive nor Euclidean.

We can further consider the logical systems that arise when one adds certain axioms to K. For example, as additional possible axioms, we have:

(D) $B_\alpha\phi \rightarrow \neg B_\alpha \neg \phi$.

(4) $B_\alpha\phi \rightarrow B_\alpha B_\alpha \phi$.

(5) $\neg B_\alpha \phi \rightarrow B_\alpha \neg B_\alpha \phi$

We'll denote the result of adding some axiom X to K, KX. We have, then, the following soundness and completeness results.

THEOREM 3. $\vdash_{kd} \phi \Leftrightarrow \models_{\mathcal{K}_D} \phi$.

THEOREM 4. $\vdash_{kd4} \phi \Leftrightarrow \models_{\mathcal{K}_D \cap \mathcal{K}_4} \phi$.

THEOREM 5. $\vdash_{kd45} \phi \Leftrightarrow \models_{\mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5} \phi$.

These results tell us the following. First, if we assume that each idealized doxastic state can be modeled by some $M \in \mathcal{K}_D$, then we have:

(B_D) an idealized doxastic agent's beliefs will be consistent.

And, if we further assume that, for each $M \in \mathcal{K}_D$, there is some idealized doxastic state that is represented by M, then it follows that (B_N), (B_K)

and (B_D) provide a complete characterization of those properties that are shared by every idealized doxastic state.

Second, if we assume that, in addition, each idealized doxastic state can be modeled by some $M \in \mathcal{K}_D \cap \mathcal{K}_4$, then we also have:

> (B_4) if an idealized doxastic agent believes ϕ, then she will believe that she believes ϕ.

We'll call this property *positive transparency*.

If we further assume that, for each $M \in \mathcal{K}_D \cap \mathcal{K}_4$, there is some idealized doxastic state that is represented by M, then it follows that (B_N), (B_K), (B_D) and (B_4) provide a complete characterization of those properties that are shared by every idealized doxastic state.

Finally, if we assume that, in addition, each idealized doxastic state can be modeled by some $M \in \mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5$, then we also have:

> (B_5) if an idealized doxastic agent fails to believe ϕ, then she will believe that she fails to believe ϕ.

We'll call this property *negative transparency*.

If we further assume that, for each $M \in \mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5$, there is some idealized doxastic state that is represented by M, then it follows that (B_N), (B_K), (B_D), (B_4), and (B_5) provide a complete characterization of those properties that are shared by every idealized doxastic state.

In the doxastic logic literature, it is typically assumed that every idealized doxastic state can be represented by some element of $\mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5$, and that each element of this set accurately represents some idealized doxastic state. Given this assumption, then, the logic governing the operator B_α is the modal logic KD45.

Note that the following principle is not assumed to hold:

> (T) $\quad B_\alpha\phi \rightarrow \phi$.

That is, we do not assume that our idealized doxastic states are error-free. An ideal belief state, on this picture, need not be one that only includes true beliefs.

> DEF. We say that R_α is *reflexive* just in case, for every $w \in W$, $wR_\alpha w$.

The axiom (T) is guaranteed to hold in any Kripke model whose accessibility relation is reflexive. Importantly, then, we do not assume that a model M representing an idealized doxastic state has a reflexive accessibility relation.

1.2 *Group Beliefs*

So far, our doxastic models have treated the doxastic state of only a single agent. This restriction, however, can be easily relaxed. Instead of a single operator B_α, let $\mathcal{L}$ now contain a series of operators $B_{\alpha_1}, B_{\alpha_2}, \ldots, B_{\alpha_r}$. A Kripke model for $\mathcal{L}$ will be a tuple $\langle W, R_{\alpha_1}, R_{\alpha_2} \ldots, R_{\alpha_r}, [\![\cdot]\!] \rangle$, and truth-at-a-point in such a model will be defined in the obvious way. As with the case of our individual models, we can impose various restrictions for each R_{α_i} such as seriality, transitivity etc. We'll let $\mathcal{K}^\alpha$ be the set of Kripke models for $\mathcal{L}$. We'll let D^α, $\mathcal{K}^\alpha_4$, $\mathcal{K}^\alpha_5$ be, respectively, the set of Kripke models for $\mathcal{L}$ such that each accessibility relation is serial, transitive, and Euclidean.

This type of model allows us to simultaneously represent the doxastic states of multiple agents. Furthermore, it allows us to represent certain collective properties of the doxastic states of groups that cannot be represented by a set of individual Kripke models for each agent in the group.[2] Let's consider how such features may be represented in this sort of model. In particular, we will consider how in such models we can represent the group doxastic properties of *common belief* and *distributed belief*.

1.2.1 *Common Belief*

What is it for ϕ be a matter of common belief amongst a group of agents? The intuitive idea is that common belief is a matter of each agent in the group believing ϕ, and each agent believing that each agent believes ϕ, and each agent believing that each agent believes that each agent believes $\phi \ldots$, and so on, ad infinitum.[3] This sort of group doxastic property can be represented in our models as follows.

> DEF. We will say that w and w' are *n-connected* just in case there is some series of worlds $w_1, \ldots, w_{n+1}$ such that $w = w_1$, $w_{n+1} = w'$ and for each pair $\langle w_i, w_{i+1} \rangle$ there is some R_{α_j} such that $w_i R_{\alpha_j} w_{i+1}$. We'll write $w R_n w'$ to indicate that w and w' are n-connected.

So, for example, the set of 1-connected worlds will just be those pairs of worlds such that there is some j such that $w R_{\alpha_j} w'$, while the set of 2-connected worlds will just be those pairs of worlds that are connected via the belief accessibility relation of at most two agents, etc.

2 See, e.g., Fagin, Halpern, and Vardi (1991), Halpern and Moses (1992), Halpern and Moses (1984), and Halpern, Moses, and Vardi (1995) for important work on the doxastic and epistemic properties of groups.

3 The concepts of common belief and common knowledge and the role that these play in reasoning were introduced in Lewis (1969). See also Aumann (1976) for another influential early treatment of these ideas. See Barwise (1988) for alternative analyses of the notions of common belief and common knowledge.

Let the schematic abbreviation $B_{\alpha^n}\phi$ be inductively characterized as follows:

DEF.

1. $B_{\alpha^1}\phi =_{\text{df}} B_{\alpha_1}\phi \wedge B_{\alpha_2}\phi \wedge \ldots \wedge B_{\alpha_r}\phi$,
2. $B_{\alpha^n}\phi =_{\text{df}} B_{\alpha_1}B_{\alpha^{n-1}}\phi \wedge B_{\alpha_2}B_{\alpha^{n-1}}\phi \wedge \ldots \wedge B_{\alpha_r}B_{\alpha^{n-1}}\phi$.

So B_{α^1} abbreviates the claim that each agent in the group believes ϕ. While B_{α^n} abbreviates the claim that each agent in the group believes that everyone in the group believes ϕ to level $n-1$. As a further definitional abbreviation, we let:

DEF. $M^n_\alpha\phi =_{\text{df}} B_{\alpha^1}\phi \wedge \ldots \wedge B_{\alpha^n}\phi$.

We will read $M^n_\alpha\phi$ as saying that there is *mutual belief of degree n* that ϕ amongst $\alpha_1, \ldots, \alpha_r$.

Given these definitions, it follows that the truth of $B_{\alpha^n}\phi$ and $M_{\alpha^n}\phi$, relative to a point w, in a model M, may be characterized as follows.

$[\![B_{\alpha^n}\phi]\!]^w = 1$ just in case $[\![\phi]\!]^{w'} = 1$, for every w' such that wR_nw'.

$[\![M_{\alpha^n}\phi]\!]^w = 1$ just in case $[\![\phi]\!]^{w'} = 1$, for every w' such that there is some $1 \leq i \leq n$ such that wR_iw'.

So far we haven't added any expressive power to our language. Each operator B_{α^n} and M_{α^n} is merely an abbreviation for some formula already in $\mathcal{L}$. Suppose, however, that we wanted to say that ϕ is a matter of common belief amongst $\alpha_1, \ldots, \alpha_r$. The natural way to express this is to say that, for each n, $M_{\alpha^n}\phi$ holds. Expressing common belief in this way, though, would require quantificational devices or devices of infinite conjunction that our language lacks. This doesn't, however, mean that we can't express the property of common belief in a propositional modal language. To do so, however, we need to add a new operator to our language.

Let $\mathcal{L}$, then, be the language that includes, in addition to each of the operators B_{α_i}, an operator C_α. A Kripke model for our new language $\mathcal{L}$ will still be a tuple $M = \langle W, R_{\alpha_1}, R_{\alpha_2} \ldots, R_{\alpha_r}, [\![\cdot]\!]\rangle$. Given our relations R_{α_i}, we can define the following relation on points in W:

DEF. Let $R_1^+ = \bigcup_{n\geq 1} R_n$.

R_1^+ is the so-called *transitive closure* of R_1.[4] Given this definition, we have that wR_1^+w' just in case there is some n such that wR_nw'. Thus wR_1^+w'

4 This is the smallest transitive relation containing R_1. To see why this is a transitive relation, assume that we have $w_1R_1^+w_2$ and $w_2R_1^+w_3$. Then we have that there is some n such that $w_1R_nw_2$, i.e., w_1 and w_2 are n-connected. And we also have that there is some m such that $w_2R_mw_3$, i.e., w_2 and w_3 are m-connected. But, given this, it follows that we have $w_1R_{n+m}w_3$, i.e., w_1 and w_3 are $(n+m)$-connected.

holds just in case there is some finite length path connecting w and w' via the accessibility relations R_{α_i}.

Note that since R_1^+ is definable in terms of the $R_{\alpha_i} \in M$, we do not need to include this relation in M in order to appeal to it in characterizing the truth of certain sentences in such a model.

We can now characterize the truth of a sentence $C_\alpha\phi$ relative to a world of evaluation in M as follows:

$$[\![C_\alpha\phi]\!]^w = 1 \text{ just in case } [\![\phi]\!]^{w'} = 1, \text{ for every } w' \text{ such that } wR_1^+w'.$$

What are the logical properties governing the common belief operator C_α? We can characterize the logic of this operator as follows. Let K_α be the multi-modal logic characterized by each instance (N), and the relevant instance of (K), for each operator B_{α_i}. (Similarly for KD_α, $KD4_\alpha$ and $KD45_\alpha$.) And let K_α^c (or KD_α^c, $KD4_\alpha^c$ and $KD45_\alpha^c$) be the axiomatic system we get by adding to K_α (or KD_α, $KD4_\alpha$ and $KD45_\alpha$) the following axiom and rule of inference:

(C_1) $\quad C_\alpha\phi \rightarrow M_\alpha^1(\phi \wedge C_\alpha\phi)$.

(R_1) $\quad \vdash_{K_\alpha^c} \phi \rightarrow M_\alpha^1(\psi \wedge \phi) \Rightarrow \vdash_{K_\alpha^c} \phi \rightarrow C_\alpha\psi$.

We, then, have the following soundness and completeness result:[5]

THEOREM 6. $\quad \vdash_{K_\alpha^c} \phi \Leftrightarrow \models_{\mathcal{K}^\alpha} \phi$.

Similar results show that KD_α^c is sound and complete with respect to D^α, that $KD4_\alpha^c$ is sound and complete with respect to $D^\alpha \cap \mathcal{K}_4^\alpha$, and that $KD45_\alpha^c$ is sound and complete with respect to $D^\alpha \cap \mathcal{K}_4^\alpha \cap \mathcal{K}_5^\alpha$. The logic governing C_α, then, is characterized by adding (C_1) and (R_1) to the axioms governing the operators B_{α_i}.

Now it's clear that the structural properties of the accessibility relation R_1^+ will supervene on the structural properties of the accessibility relations R_{α_i}. Importantly, though, the structural properties of R_1^+ may be distinct from those of R_{α_i}. In certain cases, R_1^+ may have additional structural properties to those of R_{α_i}, while in other cases, R_1^+ may lack certain structural properties had by each of the R_{α_i}. Given such discrepancies, then, the logic governing common belief may be distinct from the logic governing individual belief. Let's consider, briefly, some ways in which common belief may inherit some of the logical properties governing individual belief and some ways in which the logic governing common belief may come apart from the logical properties governing individual belief.

First, note that the transitive closure of a serial relation is also serial. If, then, each R_{α_i} is serial, R_1^+ will also be serial. And so, if the logic

5 See Halpern and Moses (1992) for a proof of this. Halpern and Moses (1992), in fact, includes an additional axiom stating $M_\alpha^1\phi \leftrightarrow (B_{\alpha_1}\phi \wedge \ldots \wedge B_{\alpha_1}\phi)$. This, however, is a *definitional* truth and so is not strictly speaking required as an axiom.

governing individual beliefs includes the principle (D), then so will the logic governing common belief. Thus, if the individual agents that we are representing are such that their beliefs are guaranteed to be consistent, then so too will the common beliefs of this group.

We have, then:

(D) $\models_{\mathcal{D}^\alpha} C_\alpha\phi \rightarrow \neg C_\alpha\neg\phi$.

Next, note that, given its definition, R_1^+ is guaranteed to be transitive. In particular, it will satisfy this property whether or not all R_{α_i} do. In this manner, then, R_1^+ may have a structural feature that some R_{α_i} lack. Given that R_1^+ is transitive, we have then:

(4) $\models_{\mathcal{K}^\alpha} C_\alpha\phi \rightarrow C_\alpha C_\alpha\phi$.

Common belief, then, is guaranteed to be positively transparent, even if individual beliefs are not.

Finally, note that while each R_{α_i} being serial entails that R_1^+ is serial, it does *not* follow that if each R_{α_i} is Euclidean then R_1^+ will also be Euclidean.[6] Thus, even if the logic of individual belief entails that individual belief is negatively transparent, it does not follow that common belief must also be negatively transparent.

1.2.2 *Distributed Belief*

What is it for ϕ to be a matter of distributed belief? The intuitive idea is that ϕ is a distributed belief amongst some group of agents just in case ϕ is a consequence of what all of the agents believe.

To express this notion, we'll introduce the operator D_α to our language $\mathcal{L}$. A model for $\mathcal{L}$ will still be a tuple $M = \langle W, R_{\alpha_1}, R_{\alpha_2} \ldots, R_{\alpha_r}, [\![\cdot]\!]\rangle$. We define the following relation amongst the members of W, given our relations R_{α_i}.

DEF. Let wR_dw' just in case for every R_{α_i}, $wR_{\alpha_i}w'$.

Truth-at-a-world for a formula $D_\alpha\phi$, in a model M, can, then, be characterized as follows:

$[\![D_\alpha\phi]\!]^w = 1$ just in case $[\![\phi]\!]^{w'} = 1$, for every w' such that wR_dw'.

We can characterize the logic of this operator as follows. Again let K_α be the multi-modal logic characterized by each instance (N), and the relevant instance of (K), for each operator B_{α_i}. And let K_α^d be the axiomatic system we get by adding to K_α the relevant instance of (K) for the operator D_α, as well an axiom of the following form, for each α_i:

(D_1) $D_\alpha\phi \rightarrow B_{\alpha_i}\phi$.

We, then, have the following soundness and completeness result:[7]

6 See Lismont and Mongin (1994), Colombetti (1993), and Bonanno and Nehring (2000) for proofs and discussion of this result.

7 See Halpern and Moses (1992).

THEOREM 7. $\vdash_{K^d_\alpha} \phi \Leftrightarrow \models_{\mathcal{K}^\alpha} \phi$.

Similarly, if we let $K4_\alpha$ ($K45_\alpha$) be the the multi-modal logic characterized by each instance (N), and the relevant instances of (K) and (4) (and (5)), for each operator B_{α_i}, and let $K4^d_\alpha$ ($K45^d_\alpha$) be the axiomatic system we get by adding to $K4_\alpha$ ($K45_\alpha$) the relevant instances of (K) and (4) (and (5)) for the operator D_α, then we can also show that $K4^d_\alpha$ ($K45_\alpha$) is sound and complete with respect to $\mathcal{K}^\alpha_4$ ($\mathcal{K}^\alpha_4 \cap \mathcal{K}^\alpha_5$).

It's clear that the structural properties of R_d will supervene on the structural properties of the accessibility relations R_{α_i}. While, though, R_d may inherit certain structural properties from R_{α_i}, other structural properties of R_d may be distinct from those of R_{α_i}.

First, note that if each R_{α_i} is transitive, then R_d is transitive. Similarly, if each R_{α_i} is Euclidean, then R_d is Euclidean. It's for this reason that if the logic governing the B_{α_i} includes the principles (4) or (5), so too will the logic governing D_α.

Importantly, however, it doesn't follow from the fact that each R_{α_i} is serial, that R_d is also serial. For it doesn't follow from the fact that, for each w, and each R_{α_i}, there is w' such that $wR_{\alpha_i}w'$, that for each w there is some w' such that wR_dw'. For while, for some w, it may be the case that, for each R_{α_i}, there is some w' such that $wR_{\alpha_i}w'$, this w' need not be the same in each case. Even, then, if the logic governing each B_{α_i} includes the principle (D), it doesn't follow this will be a principle governing D_α. Even, then, if each individual's beliefs are consistent, the distributed beliefs of the group need not be.

1.3 *Some Remarks on the Interpretation of the Formalism*

So far, we've been assuming that our doxastic models represent the doxastic states of certain *idealized* agents. And, as we've noted, there is a standard assumption in the literature that the logic governing such states is KD45. It would, however, be a mistake, I think, to take there to be a substantive question of whether or not the logic governing idealized doxastic states *really* is KD45 or some other logic. Instead, I think it is much more natural to think of these principles as simply *codifying* a certain idealization. On this view, then, there are various types of idealized doxastic states that we might investigate. For example, we might consider those doxastic states that can be represented by some model in $\mathcal{K}$. Doxastic states of this type would be logically omniscient and closed under logical consequence, but perhaps not consistent or perhaps not positively or negatively transparent. Or we might consider those doxastic states that can be represented by some model in $\mathcal{K} \cap \mathcal{D} \cap \mathcal{K}_4$. Doxastic states of this type would be logically omniscient, closed under logical consequence, consistent and

positively transparent, but not negatively transparent. And so on. Now different types of idealized doxastic states may be useful or illuminating for different purposes, but it seems implausible to me that any one of these idealizations stands out as being of significantly greater theoretical importance than all of the others.

One might, however, endorse the bolder hypothesis that our doxastic models are, in fact, intended to represent necessary features of any possible doxastic state. Given this view, the question of whether such models are appropriate and, if so, which constraints should be endorsed, becomes a substantive question. Some have, indeed, argued that the nature of doxastic states makes it the case that they may be represented by models in $\mathcal{K}$.[8] Others have argued that the nature of doxastic states makes it the case that principles such as (B_4) and (B_5) will be satisfied and so models in $\mathcal{K}_4$ or $\mathcal{K}_5$ may serve to represent such states.[9] These claims, however, are quite controversial, and it would take us too far afield now to assess their plausibility.[10] Suffice it to say, there are certain accounts of the nature of doxastic states according to which such states may in fact be accurately represented by the sorts of models we've been considering, while, according to other accounts, certain doxastic states—indeed the types of doxastic states that actual agents tend to have—cannot be accurately represented by the sorts of models that we've been considering.

A third way of interpreting our doxastic models, which should be distinguished from the first interpretation, has it that the role of this class of models is to codify certain general principles that a rational agent's doxastic state *ought* to satisfy. Here we might profitably consider, as an analogous view, the Bayesian account of credal rationality. According to a subjective Bayesian, the class of probability functions defined over some algebra $\mathcal{A}$ represents the class of rationally permissible credal states defined over $\mathcal{A}$. Those features, then, that are common to all such functions represent rational requirements on any credal state defined over such an algebra. Similarly, one might hold that the class of models $\mathcal{K}$ (or the class $\mathcal{K} \cap \mathcal{D} \cap \mathcal{K}_4$ etc.) represent the class of rationally permissible doxastic states. Those features, then, that are common to this class, i.e., the valid formulas

8 See e.g., Stalnaker (1984), Lewis (1974), and Lewis (1999). Both Stalnaker and Lewis, however, recognize that there must be a sense in which agents may have contradictory beliefs or fail to have beliefs that are closed under logical consequence. Lewis (2000) argues that we may think of an agent's doxastic state as consisting of various fragments, where each fragment may be represented by a possible worlds model. An agent, then, may have inconsistent beliefs by having fragments that disagree, and the agent may have beliefs that fail to be logically closed by believing, say, ϕ relative to one fragment and $\phi \rightarrow \psi$ relative to another, but not believing ψ relative to any fragment.

9 See, for example, Shoemaker (1996a, 1996b) for arguments that, at least in certain cases, positive introspection should hold as a constitutive matter.

10 The arguments in Williamson (2000), for example, put serious pressure on the idea that the transparency principles (B_4) and (B_5) will hold for actual agents.

given the class of models, represent, on this view, rational requirements on any doxastic state.

Now this view may seem to be a mere notational variant on the first interpretation. However, concluding this would, I think, be a mistake. What an actual agent ought to believe and what an idealized agent would believe are not the same thing. Here's a somewhat facile, but I think sufficiently instructive example that illustrates this point. An idealized agent would, plausibly, believe that they are idealized. However, a rational agent, who is not an idealized agent, should not be rationally required to believe that they are idealized. Thus, a representation of what an idealized doxastic state would look like is not, thereby, a representation of what doxastic features a rational agent ought to have.

The view that our doxastic models serve to codify rational requirements on doxastic states, again, makes it a substantive question which class of Kripke models, if any, we should take as the appropriate class for formulating our doxastic logic. One may, for example, maintain that doxastic states ought to be such that they're consistent and closed under logical entailment, but deny that doxastic states ought to be transparent on pain of irrationality. Once again, the issues here are subtle and we will simply content ourselves with flagging the issues, without making any attempt to resolve them.

Having noted these three possible roles that our doxastic models may play, it is worth highlighting that different classes of models might, in fact, play different roles. So, for example, one might maintain that the class $\mathcal{K}$ is the smallest class that serves to characterize how a rational agent's doxastic state ought to be. But one might still find it profitable to investigate what features are exhibited by the sorts of idealized doxastic states characterized by, say, $\mathcal{K} \cap \mathcal{D} \cap \mathcal{K}_4$.

In what follows, I will often continue to speak as if the Kripke models, as well as other models we'll introduce, are meant to represent idealized doxastic states. However, in certain cases a normative or descriptive interpretation may seem more natural and so I will sometimes talk as if the models in question are meant to describe such facts. In each case, though, it is worth bearing in mind the alternative interpretations that are available.

2 BELIEF REVISION

So far we've seen how to represent certain features of idealized doxastic states. In particular, we've seen how to represent the beliefs of an idealized doxastic agent, including beliefs about that agent's beliefs, as well as beliefs about other agents' beliefs. The models that we've looked at, however, say nothing about how idealized doxastic states should change given new

information. In this section we'll look at an influential account of belief revision called AGM.[11]

The basic theory of AGM consists of a set of formal postulates that serve to codify rational constraints on belief revision, expansion and contraction. In addition to such formal postulates, however, various authors have provided models of how functions meeting these constraints may be determined. We'll begin, in Section 2.1, by considering the basic postulates of AGM. In Section 2.2, we'll then consider some possible connections between rational belief revision, expansion and contraction. In Section 2.3, we'll consider some models for belief revision and contraction. And, finally, in Section 2.4 we'll look at some additional postulates that have been proposed to handle the phenomenon of iterated belief revision.

2.1 *AGM: The Basic Postulates*

Let $\mathcal{L}$ be a set of sentences closed under the standard Boolean operators, and let $\Gamma \subseteq \mathcal{L}$ be a set of such sentences. We'll denote by $Cl(\Gamma)$ the logical closure of Γ. In standard presentations of this theory, it is assumed that rational agents have belief states that can be (at least partially) modeled by logically closed sets of sentences. In this section, we will follow this practice as well. Note that this marks a departure from our treatments of belief states in the previous section, where such states were modeled as sets of possible worlds. We'll let B be a possible belief set, i.e, a set of sentences such that $B = Cl(B)$. We denote the set of belief sets $\mathcal{B}$.

We'll first consider the AGM postulates governing rational belief expansion. We let $+ : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function mapping pairs of belief sets and sentences to belief sets. We'll let B^+_ϕ be the result of applying the function $+$ to the belief set B and sentence ϕ. According to AGM, rational belief expansions must satisfy the following constraints.

(B^+_1) $\phi \in B^+_\phi$.

(B^+_2) $B \subseteq B^+_\phi$.

(B^+_3) If $\phi \in B$, then $B = B^+_\phi$.

(B^+_4) If $A \subseteq B$, then $A^+_\phi \subseteq B^+_\phi$.

(B^+_5) For any operation # satisfying B^+_1–B^+_4, $B^+_\phi \subseteq B^\#_\phi$.

We can think of expansion as an operation that increases the agents belief set B to accommodate belief in ϕ. Then (B^+_1) tells us that, given this

11 This account developed out of Alchourrón et al. (1985). For a comprehensive survey see Gärdenfors (1988). See also Huber (this volume) for a helpful treatment of this and other related material.

operation, ϕ will be a part of the resultant belief set. While (B_2^+) tells us that everything that was believed prior to the operation is believed after the operation. Also (B_3^+) tells us that, if ϕ is already believed, given B, then the expansion operation is trivial. Additionally (B_4^+) tells us that the expansion operation is monotone, i.e., that it preserves the subset relation. And, finally, (B_5^+) tells us that, in a specific sense, expansion is the most conservative operation with the preceding characteristics.

It can be shown that (B_1^+)–(B_5^+) uniquely pin down the expansion operator. Thus:

> THEOREM 8. A function $+$ satisfies (B_1^+)–(B_5^+) just in case $B_\phi^+ = Cl(B \cup \{\phi\})$.[12]

Next, we consider belief revision. We let $* : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function mapping pairs of belief sets and sentences to belief sets. According to AGM, rational belief revisions must satisfy the following constraints.

(B_1^*) $\phi \in B_\phi^*$.

(B_2^*) $B_\phi^* \subseteq B_\phi^+$.

(B_3^*) If $\neg\phi \notin B$, then $B_\phi^+ \subseteq B_\phi^*$.

(B_4^*) $B_\phi^* = \mathcal{L}$ just in case $\models \neg\phi$.

(B_5^*) If $\models \phi \leftrightarrow \psi$, then $B_\phi^* = B_\psi^*$.

(B_6^*) $B_{\phi\wedge\psi}^* \subseteq (B_\phi^*)_\psi^+$.

(B_7^*) If $\neg\psi \notin B_\phi^*$, then $(B_\phi^*)_\psi^+ \subseteq B_{\phi\wedge\psi}^*$.

Like expansion, we can think of revision as an operation that changes an agent's belief set B to accommodate belief in ϕ. Unlike expansion, though, in belief revision certain beliefs may be discarded to accommodate ϕ.

Constraint (B_1^*) tells us that, given this operation, ϕ will be a part of the resultant belief set. While (B_2^*) tells us that everything that is believed given belief revision will be believed given belief expansion. Also (B_3^*) tells us that the reverse is also true, and so expansion and revision deliver the same output, when ϕ is logically compatible with B. And (B_4^*) tell us that that the result of belief revision will be a consistent belief set just in case ϕ is itself logically consistent. Constraint (B_5^*) tells us that logically equivalent sentences induce the same revision operation on a belief set. And (B_6^*) tells us that everything that is believed after revising a belief set given a conjunction $\phi \wedge \psi$, will be believed after first revising the same belief set given ϕ and then expanding the resultant belief set given ψ.

12 See Gärdenfors (1988).

Finally, (B_7^*) tell us that the reverse is true, and so revising given ϕ and then expanding given ψ delivers the same output as revising given $\phi \wedge \psi$, when ψ is consistent with the result of revision given ϕ.

Unlike with the constraints on $+$, these constraints do *not* suffice to uniquely determine the function $*$. Instead, there is a non-empty, non-singleton, set of functions that satisfy (B_1^*)–(B_7^*).

Finally, we consider belief contraction. We let $- : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function mapping pairs of belief sets and sentences to belief sets. According to AGM, rational belief contractions must satisfy the following constraints.

(B_1^-) $B_\phi^- \subseteq B$.

(B_2^-) If $\phi \notin B$, then $B_\phi^- = B$.

(B_3^-) If $\not\models \phi$, then $\phi \notin B_\phi^-$.

(B_4^-) If $\phi \in B$, then $B \subseteq (B_\phi^-)_\phi^+$.

(B_5^-) If $\models \phi \leftrightarrow \psi$, then $B_\phi^- = B_\psi^-$.

(B_6^-) $B_\phi^- \cap B_\psi^- \subseteq B_{\phi \wedge \psi}^-$.

(B_7^-) If $\phi \notin B_{\phi \wedge \psi}^-$, then $B_{\phi \wedge \psi}^- \subseteq B_\phi^-$.

We can think of contraction as an operation that changes an agent's belief set B to accommodate the removal of a belief ϕ.

Constraint (B_1^-) tells us that belief contraction does not introduce any new beliefs. While (B_2^-) tells us that if the agent does not already believe ϕ, then the result of contracting by ϕ leaves the agent's belief set unchanged. Then (B_3^-) tells us that if ϕ is not a logical truth, then ϕ will not be in any belief set that is contracted by ϕ. And (B_4^-) tells us that if a belief set B contains ϕ, then the result of contracting this belief set by ϕ and then expanding the resulting set by ϕ will contain everything that is in B. Next (B_5^-) tells us that logically equivalent sentences induce the same contraction operation on a belief set. And (B_6^-) tells us that every belief that remains when a belief set is contracted by ϕ and by ψ will remain when the belief set is contracted by $\phi \wedge \psi$. Finally, (B_7^-) tells us that if ϕ is not in the belief set that results from contracting a belief set B by $\phi \wedge \psi$, then everything that results from contracting B by $\phi \wedge \psi$ will be in the set that result from contracting B by ϕ.

As with the constraints on $*$, these constraints do not uniquely determine the function $-$. Again, there is a non-empty, non-singleton set of functions that satisfy each of (B_1^-)–(B_7^-).

2.2 *Relations Between Operations*

Given these rational constraints on contraction, revision and expansion, it is natural to ask what connections there might be between these three operations. In this section, we'll consider some possible options.

So far we've been talking as if agents adopt *distinct* policies of rational belief revision, contraction and expansion. Following Levi (1977), however, one might maintain that agents only really adopt policies of contraction and expansion, and that a policy of revision is determined by the latter two policies in the following manner.

> CONSTITUTIVE LEVI IDENTITY. $B^*_\phi =_{\text{df}} (B^-_{\neg\phi})^+_\phi$.

If the claim that the adoption of a rational revision policy simply consists in the adoption of rational contraction and expansion policies is to be at all plausible, it must be the case that, given the putative analysis, it is ensured that the resulting revision policy will indeed be rational given that the contraction and expansion policies are. The following result shows that, given the CONSTITUTIVE LEVI IDENTITY, this is so.

> THEOREM 9. If $-$ satisfies (B^-_1)–(B^-_3) and (B^-_5)–(B^-_7), and $+$ satisfies (B^+_1)–(B^+_5), then, given the CONSTITUTIVE LEVI IDENTITY, $*$ satisfies (B^*_1)–(B^*_7).[13]

Now even if one wants to reject the claim that the adoption of a belief revision policy simply consists in the adoption of expansion and contraction policies, one should, I think, nonetheless hold that there are important rational constraints governing which policies of belief revision, expansion and contraction an agent may simultaneously adopt. In particular, whether one thinks that rational belief revision should be *analyzed* in terms of rational belief contraction and expansion, one should, I think, endorse the following normative constraint.

> NORMATIVE LEVI IDENTITY. If $*$ is an agent's revision policy, $-$ her contraction policy, and $+$ her expansion policy, then the agent ought to be such that $B^*_\phi = (B^-_{\neg\phi})^+_\phi$.

Two points are worth mentioning here.

First, the NORMATIVE LEVI IDENTITY does, indeed, impose a substantive constraint in addition to those imposed by (B^+_1)–(B^+_5), (B^*_1)–(B^*_7), and (B^-_1)–(B^-_7). For there are functions $+$, $*$ and $-$ that satisfy (B^+_1)–(B^+_5), (B^*_1)–(B^*_7), and (B^-_1)–(B^-_7), respectively, but that fail to jointly satisfy the condition that $B^*_\phi = (B^-_{\neg\phi})^+_\phi$.[14]

13 See Gärdenfors (1988), ch. 3.6.

14 This follows from the fact that there is a unique function $+$ satisfying conditions (B^+_1)–(B^+_5), together with the fact that there are multiple functions $*$ and $-$ satisfying (B^*_1)–(B^*_7), and (B^-_1)–(B^-_7) respectively.

Second, THEOREM 9 guarantees that the constraints imposed by the NORMATIVE LEVI IDENTITY are, indeed, consistent with the constraints imposed by (B_1^+)–(B_5^+), (B_1^-)–(B_7^-), and (B_1^*)–(B_7^*). For there are functions $-$ and $+$ that satisfy the constraints imposed by (B_1^-)–(B_7^-), and (B_1^+)–(B_5^+), and it follows from this fact, together with THEOREM 9 that, given such functions, there is a function $*$ that satisfies the constraints imposed by (B_1^*)–(B_7^*), and, in addition, is such that $B_\phi^* = (B_{\neg\phi}^-)_\phi^+$.

Another option, following Harper (1976), is to maintain that agents only really adopt policies of revision and expansion. In particular, one may maintain that an agent's policy of contraction is determined by her policy of revision as follows.

> CONSTITUTIVE HARPER IDENTITY. $B_\phi^- =_{\text{df}} B \cap B_{\neg\phi}^*$.

If the claim that the adoption of a rational contraction policy simply consists in the adoption of a rational revision policy is to be at all plausible, it must the case that, given the putative analysis, it is ensured that the resulting contraction policy will indeed be rational given that the revision policy is. The following result shows that, given the CONSTITUTIVE HARPER IDENTITY, this is so.

> THEOREM 10. If $*$ satisfies (B_1^*)–(B_7^*), then, given the CONSTITUTIVE HARPER IDENTITY, $-$ satisfies (B_1^-)–(B_7^-).[15]

Now, again, even if one wants to reject the claim that the adoption of a belief contraction policy simply consists in the adoption of a revision policy, one should, I think, still hold that there are important rational constraints governing which contraction and revisions policies an agent may simultaneously adopt. In particular, whether one thinks that rational belief contraction should be analyzed in terms of rational belief revision, one should, I think, endorse the following normative constraint:

> NORMATIVE HARPER IDENTITY. If $*$ is an agent's belief revision policy, and $-$ her belief contraction policy, then the agent ought to be such that $B_\phi^- = B \cap B_{\neg\phi}^*$.

Two points are, again, worth mentioning here.

First, the NORMATIVE HARPER IDENTITY provides a substantive constraint in addition to (B_1^*)–(B_7^*) and (B_1^-)–(B_7^-). For there are functions $*$ and $-$ that satisfy the latter constraints but for which the identity $B_\phi^- = B \cap B_{\neg\phi}^*$ fails to hold.[16]

Second, THEOREM 10 guarantees that the constraint imposed by the NORMATIVE HARPER IDENTITY is consistent with the constraints imposed

15 See Gärdenfors (1988) ch. 3.6.

16 This follows from the fact that there are multiple functions $*$ and $-$ satisfying (B_1^*)–(B_7^*), and (B_1^-)–(B_7^-) respectively.

by (B_1^*)–(B_7^*) and (B_1^-)–(B_7^-). For there is some function $*$ satisfying (B_1^*)–(B_7^*), and so, given THEOREM 10, it follows that there is some other function $-$ satisfying (B_1^-)–(B_7^-), such that $B_\phi^- = B \cap B_{\neg\phi}^*$.

2.3 *AGM: Models*

Constraints (B_1^-)–(B_7^-) determine a class of rational contraction functions, while (B_1^*)–(B_7^*) determine a class of rational revision functions. There are, however, other ways of characterizing the classes determined by (B_1^-)–(B_7^-) and (B_1^*)–(B_7^*). In particular, we can provide models of possible features of an agent's doxastic state that might serve to determine which rational revision or contraction policy she adopts, and we can show that, given certain rational constraints on such features, the classes of rationally permissible revision or contraction functions are the same classes as those determined by (B_1^-)–(B_7^-) and (B_1^*)–(B_7^*).

2.3.1 *Sphere Systems*

We first consider a model of how an agent's doxastic state might serve to determine a rational revision policy.[17] In rough outline, we may think of an agent's doxastic state, in addition to determining a belief set, as also determining, for each possible belief set B, an ordering of plausibility amongst various maximal consistent descriptions of how the world might be. Given a doxastic state with such structure, we may think of the agent as adopting a revision policy such that, given a belief set B and a sentence ϕ, the resulting revised belief set is just the intersection of the most plausible maximal consistent descriptions of how the world might be that contain ϕ.

A little more pedantically: Let $\mathcal{W}$ be the set of maximal consistent subsets of $\mathcal{L}$. Following the literature, we'll refer to these as *possible worlds*. It is, however, worth keeping in mind that, as sets of of sentences, these differ from the possible worlds considered in the previous section. For any belief set B, we let $[\![B]\!] = \{w \in \mathcal{W} : B \subseteq w\}$. And for any $P \subseteq \mathcal{W}$, we let $B_P = \cap\{w : w \in P\}$.

Let $\mathbf{S}$ be a set of subsets of $\mathcal{W}$. We call $\mathbf{S}$ a *system of spheres centered on* B just in case $\mathbf{S}$ satisfies the following conditions.

> ORDERING. For every $S, S' \in \mathbf{S}$, either $S \subseteq S'$ or $S' \subseteq S$.
>
> CENTERING. $[\![B]\!] \in \mathbf{S}$, and for every $S \in \mathbf{S}$, $[\![B]\!] \subseteq S$.
>
> UNIVERSALITY. $\mathcal{W} \in \mathbf{S}$.
>
> LIMIT ASSUMPTION. Let ϕ be a sentence. If there is some $S \in \mathbf{S}$ such that $S \cap [\![\phi]\!] \neq \emptyset$, then there is some $S \in \mathbf{S}$ such that (i)

17 See Grove (1988) for the initial development of this model. The model is, in certain respects, notably similar to the semantic theory for counterfactuals developed in Lewis (1973).

$S \cap [\![\phi]\!] \neq \emptyset$, and (ii) for every $S' \in \mathbf{S}$ such that $S' \cap [\![\phi]\!] \neq \emptyset$, $S \subseteq S'$.

ORDERING tells us that the members of $\mathbf{S}$ can be totally ordered by the subset relation. CENTERING tells us that the set of worlds that are compatible with the belief set B is a member of $\mathbf{S}$ that is minimal with respect to the subset ordering on $\mathbf{S}$. UNIVERSALITY tells us that the set of all worlds is itself a member of $\mathbf{S}$. And, finally, the LIMIT ASSUMPTION tells us that for any sentence ϕ if the set of $S \in \mathbf{S}$ such that ϕ is true at some world in S is non-empty, then this set has a least element relative to the subset ordering on $\mathbf{S}$.

If $\not\models \neg\phi$, then we let S_ϕ be the smallest sphere intersecting ϕ. Given UNIVERSALITY and the LIMIT ASSUMPTION, such a sphere is guaranteed to exist. And if $\models \neg\phi$, then we let $S_\phi = \mathcal{W}$. We let $C_S(\phi) = [\![\phi]\!] \cap S_\phi$.

Let $\mathcal{S}$ be a function that maps each $B \in \mathcal{B}$ to a system of spheres centered on B. Call this a *sphere function*. We denote the sphere system determined by a sphere function $\mathcal{S}$, for some belief set B, by $\mathcal{S}_B$. Finally, let $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function such that there is some sphere function $\mathcal{S}$ such that for every $B \in \mathcal{B}$ and $\phi \in \mathcal{L}$, $f(B, \phi) = \cap C_{\mathcal{S}_B}(\phi)$. We call this a *sphere revision function*.

We can think of a sphere revision function, determined by some sphere function $\mathcal{S}$, as mapping a belief state B and a sentence ϕ to the belief state that is determined by the worlds at which ϕ holds that, according to $\mathcal{S}$, are closest to B.

More precisely, we can think of a sphere revision function, determined by some sphere function $\mathcal{S}$, as operating in the following manner. Given a belief set B and a sentence ϕ, we first look at the sphere system centered on B determined by $\mathcal{S}$. Next we find the smallest sphere that is compatible with ϕ, and consider the set of worlds within this sphere at which ϕ holds. The sphere revision function, then, returns as a belief set the set of sentences that are true at every world within this set.

The following theorem shows that the adoption of a rational revision function can always be modeled in terms of the adoption of a sphere revision function determined by some sphere function $\mathcal{S}$, and that, conversely, the adoption of a sphere revision function determined by some sphere function $\mathcal{S}$ will always correspond to the adoption of a rational revision function.

THEOREM 11. Function f is a sphere revision function just in case f satisfies (B_1^*)–(B_7^*).[18]

We can support the claim that a system of spheres may be thought of as encoding a relation of doxastic plausibility as follows. Call a relation $\leq$ over $\mathcal{W}$ with the following properties, a *B-plausibility ordering*.

18 See Grove (1988).

CONNECTIVITY. For every $w, w' \in \mathcal{W}$, either $w \leq w'$ or $w' \leq w$.

TRANSITIVITY. If $w \leq w'$ and $w' \leq w''$, then $w \leq w''$.

ϕ-MINIMALITY. If $[\![\phi]\!] \neq \emptyset$, then $\{w \in [\![\phi]\!] : w \leq w' \text{ for all } w' \in [\![\phi]\!]\} \neq \emptyset$.

B-MINIMALITY. $w \leq w'$ for all $w' \in \mathcal{W}$ just in case $w \in [\![B]\!]$.

Given a B-plausibility ordering $\leq$, let $S_w = \{w' \in \mathcal{W} : w' \leq w\}$. We let $\mathcal{S}_{\leq} = \{S_w : w \in \mathcal{W}\}$. It can be shown that:

THEOREM 12. For any B-plausibility ordering $\leq$, $\mathcal{S}_{\leq}$ is a system of spheres centered on B, and for any system of spheres $\mathcal{S}$ centered on B, there is a unique B-plausibility ordering $\leq$, such that $\mathcal{S} = \mathcal{S}_{\leq}$.[19]

A system of spheres, thus, encodes an ordering over possible worlds, and it is this plausibility ordering, that, according to this model, serves to determine how a rational agent will revise her belief set B given some sentence ϕ.

Given THEOREM 10, we can also use this model to provide a model of how an agent might adopt a rational contraction function. Given the adoption of a sphere revision function determined by some $\mathcal{S}$, the agent would adopt a policy of contracting a belief set B, given some sentence ϕ, so that her new belief set is given by the set of sentences that are true in all and only the worlds in $C_S(\neg\phi) \cup [\![B]\!]$. That is, such an agent will contract her belief set by ϕ by adding to the set of worlds representing this set, the most plausible $\neg\phi$ worlds. Such a policy will be guaranteed to satisfy (B_1^-)–(B_7^-).

2.3.2 *Epistemic Entrenchment*

Next, we consider a model of how an agent's doxastic state might serve to determine a rational belief contraction policy.[20] In rough outline, we may think of an agent's doxastic state, in addition to determining a belief set, as also determining, for each belief set B, a binary relation on the set of sentences of $\mathcal{L}$ that encodes information about how epistemically entrenched such sentences are, given the belief set B. While the notion of epistemic entrenchment is best thought of as being functionally defined via its role in the following account of belief contraction, one can think of an epistemic entrenchment ordering, roughly, as corresponding to an ordering representing how committed an agent is to retaining certain beliefs, given that they have a belief set B.

19 See Gärdenfors (1988).
20 See Gärdenfors and Makinson (1988) for this type of model.

Given a doxastic state that determines an entrenchment ordering for each belief set B, we may think of the agent as adopting a contraction policy such that, given a belief set B and a sentence ϕ, the agent restricts B to the subset of elements ψ such that either ψ is a theorem or $\psi \vee \phi$ is more epistemically entrenched than ϕ.

A little more pedantically: let $\leq$ be a binary relation over $\mathcal{L}$. We let $\phi < \psi =_{\text{df}} \phi \leq \psi \wedge \psi \not\leq \psi$. We call $\leq$ a *B-entrenchment relation* just in case it satisfies the following postulates.

E1. If $\phi \leq \psi$ and $\psi \leq \xi$, then $\phi \leq \xi$.

E2. If $\phi \models \psi$, then $\phi \leq \psi$.

E3. For all ϕ, ψ, either $\phi \leq \phi \wedge \psi$ or $\psi \leq \phi \wedge \psi$.

E4. If $B \neq \mathcal{L}$, then $\phi \in B$ just in case $\phi \leq \psi$, for all ψ.

E5. If $\psi \leq \phi$, for all ψ, then $\models \phi$.

Let $\preceq$: $\mathcal{B} \to \mathcal{P}(\mathcal{L} \times \mathcal{L})$ be a function that maps each $B \in \mathcal{B}$ to a binary relation over $\mathcal{L}$. We denote each such relation by $\preceq_B$. If each $\preceq_B$ is a B-entrenchment relation, we'll call $\preceq$ an *entrenchment function*. Let $C_{\preceq}$: $\mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function such that $C_{\leq}(B, \phi) = \{\psi : \psi \in B$ and either $\phi < \phi \vee \psi$ or $\models \psi\}$. Let $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function such that there is some entrenchment function $\preceq$ such that for every $B \in \mathcal{B}$ and $\phi \in \mathcal{L}$, $f(B, \phi) = C_{\leq}(B, \phi)$. We call this an *entrenchment contraction function*.

The following theorem shows that the adoption of a rational contraction function can always be modeled in terms of the adoption of an entrenchment contraction function determined by some entrenchment function $\preceq$, and that, conversely, the adoption of an entrenchment contraction function determined by some some entrenchment function $\preceq$ will always correspond to the adoption of a rational contraction function.

THEOREM 13. Function f is an entrenchment contraction function just in case f satisfies (B_1^-)–(B_7^-).[21]

Given THEOREM 9, we can also use this model to provide a model of how an agent might adopt a rational revision function. Given the adoption of an entrenchment contraction function determined by some $\preceq$, the agent would adopt the policy of revising a belief set B given ϕ, by first contracting B to the subset of elements $\psi \in B$ such that either ψ is a theorem or $\psi \vee \neg\phi$ is more epistemically entrenched that $\neg\phi$, and then expanding the resulting set by ϕ. Such a policy will be guaranteed to satisfy (B_1^*)–(B_7^*).

21 See Gärdenfors and Makinson (1988).

2.4 *Iterated Belief Revision*

If an agent has adopted a revision policy $*$ satisfying (B_1^*)–(B_7^*), then not only is it determined how the agent should revise her current belief set B, given some information ϕ, but it is also determined how the agent should revise this new belief set, given additional information ψ. For a revision policy satisfying (B_1^*)–(B_7^*) determines how any belief set should be revised given any piece of information. It has, however, been suggested that the AGM postulates provide implausible results when we consider which patterns of iterated belief revision they count as rationally permissible and rationally mandated.[22] In response to these putative problems, various emendations of, or additions to, the AGM postulates have been suggested. In this section, we'll consider some putative problems with iterated belief revision that arise for AGM and look at a few solutions that have been suggested.

2.4.1 *Problems with Iterated Belief Revision in AGM*

There are two types of problems that the AGM revision postulates have been thought to have with iterated belief revision. On the one hand, the AGM revision postulates have been thought to be too permissive, vindicating as rational certain patterns of iterated belief revision that would seem to be irrational. On the other hand, the AGM revision postulates have been thought to be too restrictive, ruling out as irrational certain patterns of iterated belief revision that would seem to be rational. Let me say a bit more about each of these worries in turn.

The first problem stems from the fact that (B_1^*)–(B_7^*) put very few constraints on iterated belief revision. To see this, let b be some particular belief set. We'll, then, let (b_1^*)–(b_7^*) be the postulates that result from (B_1^*)–(B_7^*) by saturating the variable B ranging over elements of $\mathcal{B}$ with the particular element $b \in \mathcal{B}$. Then (b_1^*)–(b_7^*) provide constraints on a function $b* : \mathcal{L} \to \mathcal{B}$ mapping sentences to belief sets. In particular, they provide constraints on functions that tell us how the belief set b should be revised given new information. Call such a function a *b-revision function*.

> DEF. For any function $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$, let f^b be the function such that $< l_x, b_y > \in f^b \leftrightarrow < b, l_x, b_y > \in f$.
>
> DEF. For any function $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$, and any function $g : \mathcal{L} \to \mathcal{B}$, let $f/g_b : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be the function such that if $b_x \neq b$, then $< b_x, l_y, b_z > \in f \leftrightarrow < b_x, l_y, b_z > \in f/g_b$, while $< b, l_y, b_z > \in f/g_b \leftrightarrow < l_y, b_z > \in g$.

22 See, for example, Boutilier (1996), Darwiche and Pearl (1997), and Stalnaker (2009).

We can think of f^b as the b-revision function determined by the revision function f. And we can think of f/g_b as the revision function that results from swapping g_b for the b-revision function determined by f. We, then, have the following results.

- For any f satisfying (B_1^*)–(B_7^*) and any $b \in \mathcal{B}$, f^b will satisfy (b_1^*)–(b_7^*).
- For any g satisfying (b_1^*)–(b_7^*), and any f satisfying (B_1^*)–(B_7^*), f/g_b will also satisfy (B_1^*)–(B_7^*).

What this shows is that functions satisfying (B_1^*)–(B_7^*) can be thought of as the result of freely choosing, for each $b \in \mathcal{B}$, some function satisfying (b_1^*)–(b_7^*). Thus (B_1^*)–(B_7^*) allow us to mix-and-match b-revision functions as we like.

This degree of freedom, however, has problematic consequences. For a b-revision function $f : \mathcal{L} \to \mathcal{B}$ can be seen as encoding *conditional beliefs*. We'll say that the conditional belief $\phi|\psi$ is accepted by f just in case $\phi \in f(\psi)$. The fact that (B_1^*)–(B_7^*) allow for arbitrary mixing and matching of b-revision functions, shows that (B_1^*)–(B_7^*) impose almost no constraints on which conditional beliefs an agent should maintain or give up when changing her belief set in light of some new information.

Here's an example that illustrates the problem.[23]

> Flying Bird. You initially believe of some animal in the distance that it's neither a bird, $\neg B$, nor can it fly, $\neg F$. You, however, have the conditional belief that it can fly, given that it's a bird, $F|B$. That is, you are disposed to come to believe that the animal can fly, were you to learn that it's a bird. Now you learn that the animal can indeed fly, and as a result you give up the conditional belief $F|B$, and form the conditional belief $\neg F|B$.

This sort of transition will seem to many to be irrational. Learning that the consequent of a conditional belief is true would seem to provide no evidence against that conditional belief. However, this transition will be sanctioned as rationally permissible given (B_1^*)–(B_7^*). Examples such as this have convinced a number of authors that further constraints, in addition to (B_1^*)–(B_7^*), are required to adequately constrain rational revision functions.

To see why one might think that (B_1^*)–(B_7^*) are not only too permissive but also too restrictive, note that if an agent adopts a revision policy satisfying (B_1^*)–(B_7^*), then which belief set she should have, given some information ϕ, is a function of her current belief set. This, however, has the following consequence. Given the adoption of a revision policy satisfying (B_1^*)–(B_7^*), an agent who starts out with a belief set B and who then receives

23 This type of example and others may be found in Darwiche and Pearl (1997).

a series of information $\psi_1, \psi_2, \ldots, \psi_3$ that she uses to successively revise her beliefs and who, as a result, winds up, again, with belief set B, is rationally required to revise this belief set given some information ϕ in exactly the same manner as she would have revised this belief set, given ϕ, prior to receiving the series of information $\psi_1, \psi_2, \ldots, \psi_3$. One might, however, think that a series of information that would ultimately leave an agent's belief set unchanged could rationally lead to a change in the agent's conditional beliefs.[24] This sort of change, however, is ruled out as irrational by the AGM postulates.

2.4.2 *Iterated Belief Revision Functions*

We'll first consider an amendment to the AGM account of rational revision that is meant to address the first problem. We'll, then, consider an alternative amendment that addresses both the first and the second problem.

In response to worries about the excessive permissiveness of AGM, Boutilier (1996) proposes a much more restricted account of rational belief revision. Perhaps the simplest way to present the account is by appeal to Grove's model in which a revision policy is represented by a set of total pre-orders over the space of possible worlds.

Let $\preceq$ be a function mapping each B to a B-plausibility ordering $\preceq_B$. As we noted earlier, each function satisfying (B_1^*)–(B_7^*) may be represented by some such function $\preceq$. For each B, let $C(B, \phi) = \{w : \phi \in w \text{ and } w \preceq_B w', \text{ for each } w' \text{ such that } \phi \in w'\}$. Boutilier (1996) suggests that in order for $\preceq$ to represent a rational revision function, in addition to each $\preceq_B$ being a B-plausibility ordering, it must also satisfy the following.

> Boutilier's Constraint. $\preceq_{C(B,\phi)}$ must be such that for all $w, w' \notin C(B, \phi)$, $w \preceq_{C(B,\phi)} w'$ just in case $w \preceq_B w'$.

The idea here is that a rational agent, in revising her beliefs in response to some information ϕ, should adjust her plausibility ordering over worlds in such a way that the most plausible ϕ-worlds are ranked highest in her new plausibility ordering but otherwise leaves the plausibility ordering amongst worlds untouched. An agent who adjusts her belief state in this manner will effectively make the minimal adjustments to her conditional beliefs as is necessary in order to accommodate ϕ.

It is easy enough to see that the constraints that Boutilier (1996) suggests rule out as irrational the problematic case of revision in Flying Bird. It has been argued, however, that Boutilier's account of belief revision demands that too many conditional beliefs be preserved, and that this has undesirable consequences about when an agent may be rationally required

24 For worries in this vicinity see Levi (1988) and Darwiche and Pearl (1997).

to give up certain beliefs. Darwiche and Pearl (1997) give the following example.

> SEQUENTIAL RED BIRD. You are initially uncertain about whether a certain animal is a red, R, or is a bird, B. You then get information that the animal is a bird, B. Then you get information, from a different source, that the animal is red, R. However, further consultation with an expert indicates that, in fact, the first piece of evidence was wrong and, in fact, the animal is not a bird, $\neg B$. As a result, you wind up believing that the animal is not a bird, $\neg B$, but that the animal is red, R.

Intuitively, this process of revision would seem to be perfectly rational. The constraints on revision proposed by Boutilier (1996), however, deem this process of revision irrational. To see this, consider the following model. Let $w_1 = R \wedge B$, $w_2 = \neg R \wedge B$, $w_3 = R \wedge \neg B$ and $w_4 = \neg R \wedge \neg B$. At t_1, your plausibility ordering is such that:

$$(t_1) \quad w_1 = w_2 = w_3 = w_4.$$

Upon getting the information B, at t_2 you minimally adjust your ordering, in accord with the Boutillier model, so that:

$$(t_2) \quad w_1 = w_2 < w_3 = w_4.$$

Then, at t_3, upon getting the information R, you again minimally adjust your ordering so that:

$$(t_3) \quad w_1 < w_2 < w_3 = w_4.$$

Finally, upon getting the information $\neg B$, you once again minimally adjust your plausibility ordering, so that at t_4 we have:

$$(t_4) \quad w_3 = w_4 < w_1 < w_2.$$

And so, what we find is that, upon making these minimal adjustments, you will fail to believe R, since this is a proposition that is false at some world that is amongst the most plausible according to your plausibility ordering at t_4.

The problem may be diagnosed as follows. When you start out uncertain about R and B, you lack the conditional belief $R|\neg B$. For, in your state at the time, you would not come to believe that the animal is red if you were to learn that the animal is not a bird. On the Boutillier model, however, when you get information ϕ, you should minimally adjust your conditional beliefs, i.e., you should only change your conditional beliefs insofar as such a change is forced on you by taking what were previously the most plausible ϕ-worlds to now be the most plausible worlds tout court. Since

coming to believe B and then R does not force one to accept the conditional belief $R|\neg B$, the Boutillier model requires that you continue to lack the conditional belief $R|\neg B$. And so, when you come to believe $\neg B$, since you lack the appropriate conditional belief that would sanction your continuing to believe R, you must give up this belief.

The point that would seem to clearly emerge from this type of example is that if we want to allow that an agent may rationally preserve certain beliefs that she forms over time, we need to allow the agent to adjust her conditional beliefs in certain non-minimal ways that are precluded given the Boutillier model.

In response to the problems of iterated revision faced, on the one hand, by Alchourrón et al. (1985), and, on the other hand, by Boutilier (1996), Darwiche and Pearl (1997) offer an alternative account of iterated revision. Their account is intended to offer stricter constraints on iterated belief revision than those imposed by Alchourrón et al. (1985), while allowing for certain permissible variations in how an agent's conditional beliefs may be updated over time that are ruled out by Boutilier (1996). In addition, their theory is designed to accommodate the second worry about iterated belief revision for AGM considered in Section 2.4.1.

According to Darwiche and Pearl (1997), belief revision should not be thought of, fundamentally, in terms mapping one belief *set* to another. Instead, belief revision should be thought of as mapping one belief *state* to another, where a belief state here is something that determines a belief set but, in addition, encodes information about the agent's conditional beliefs. We can model such a state as a plausibility ordering over the set of possible worlds.[25]

Let $\mathcal{G}$ be the set of belief states. For each $G \in \mathcal{G}$, we'll let $Bel(G)$ be the belief set determined by G. Let $\circ : \mathcal{G} \times \mathcal{L} \to \mathcal{G}$, be a function mapping pairs of belief states and sentences to belief states. Paralleling the AGM postulates, Darwiche and Pearl suggest the following constraints for such a function.

(G_1°) $\phi \in Bel(G_\phi^\circ)$.

(G_2°) $Bel(G_\phi^\circ) \subseteq Bel(G)_\phi^+$.

(G_3°) If $\neg\phi \notin Bel(G)$, then $Bel(G)_\phi^+ \subseteq Bel(G_\phi^\circ)$.

(G_4°) $Bel(B_\phi^\circ) = \mathcal{L}$ just in case $\models \neg\phi$.

(G_5°) If $\models \phi \leftrightarrow \psi$, then $G_\phi^\circ = G_\psi^\circ$.

25 We can, of course, think of the acceptance of a belief revision function on the AGM picture as adopting a policy for mapping one belief state to another. However, importantly, on the AGM account all that matters for this mapping is what the belief set looks like.

(G°_6) $Bel(G^\circ_{\phi\wedge\psi}) \subseteq Bel(G^\circ_\phi)^+_\psi$.

(G°_7) If $\neg\psi \notin Bel(G^\circ_\phi)$, then $Bel(G^\circ_\phi)^+_\psi \subseteq Bel(G^\circ_{\phi\wedge\psi})$.

In addition, however, Darwiche and Pearl also propose the following constraints.

(G°_8) If $\phi \models \psi$, then $Bel((G^\circ_\psi)^\circ_\phi) = Bel(G^\circ_\phi)$.

(G°_9) If $\phi \models \neg\psi$, then $Bel((G^\circ_\psi)^\circ_\phi) = Bel(G^\circ_\phi)$.

(G°_{10}) If $Bel(G^\circ_\phi) \models \psi$, then $Bel((G^\circ_\psi)^\circ_\phi) \models \psi$.

(G°_{11}) If $Bel(G^\circ_\phi) \not\models \neg\psi$, then $Bel((G^\circ_\psi)^\circ_\phi) \not\models \neg\psi$.

Darwiche and Pearl (1997) then show how revision functions satisfying these constraints may be modeled. Let $\preceq$ now be a function that maps each *belief state* G to a total pre-order on the set of possible worlds $\preceq_G$. (Again we'll think of possible worlds as maximal consistent sets of $\mathcal{L}$.) We let $w_1 \prec_G w_2 =_{df} w_1 \preceq_G w_2$ and $w_2 \not\preceq_G w_1$.

DEF. We say that $\preceq$ is a *faithful assignment* just in case:

(i) if $Bel(G) \subseteq w_1$ and $Bel(G) \subseteq w_2$, then $w_1 \preceq_G w_2$ and $w_2 \preceq_G w_1$;

(ii) $Bel(G) \subseteq w_1, Bel(G) \not\subseteq w_2$, then $w_1 \prec_G w_2$.

Let $f : \mathcal{G} \times \mathcal{L} \to \mathcal{G}$. We again let $B_P = \cap\{w : w \in P\}$, given a set of worlds P. And for each $G \in \mathcal{G}$, and each $\preceq$ we let $C_\preceq(G, \phi) = \{w : \phi \in w$ and $w \preceq_G w'$, for each w' such that $\phi \in w'\}$. Darwiche and Pearl (1997) show:

THEOREM 14. Function f satisfies (G°_1)–(G°_7) just in case there exists a faithful assignment $\preceq$ such that $Bel(G^f_\phi) = B_{C_\preceq(G,\phi)}$.

In addition, Darwiche and Pearl (1997) show:

THEOREM 15. If f satisfies (G°_1)–(G°_7), then f satisfies (G°_8)–(G°_{11}) just in case f and any corresponding faithful assignment $\preceq$ such that $Bel(G^f_\phi) = B_{C_\preceq(G,\phi)}$ satisfy:

(iii) if $\phi \in w_1$ and $\phi \in w_2$, then $w_1 \preceq_G w_2$ if and only if $w_1 \preceq_{G^f_\phi} w_2$;

(iv) if $\neg\phi \in w_1$ and $\neg\phi \in w_2$, then $w_1 \preceq_G w_2$ if and only if $w_1 \preceq_{G^f_\phi} w_2$;

(v) if $\phi \in w_1$ and $\neg\phi \in w_2$, then if $w_1 \prec_G w_2$, then $w_1 \prec_{G^f_\phi} w_2$;

(vi) if $\phi \in w_1$ and $\neg\phi \in w_2$, then if $w_1 \preceq_G w_2$, then $w_1 \preceq_{G^f_\phi} w_2$.

Given THEOREM 14 and THEOREM 15, we can think of an agent's belief state as being representable by a total pre-order over the space of possible worlds, while the agent's rational revision policy may be represented as a function mapping a pair of such a pre-order and a sentence to another pre-order. A rational revision policy will be representable by a function, $\circ$, that maps each such pre-order, G, and each sentence, ϕ, to a pre-order G°_{ϕ}, such that:

- the minimal worlds in G°_{ϕ} are the minimal ϕ-worlds in G;
- the ordering amongst the ϕ-worlds in G°_{ϕ} is exactly the ordering of the ϕ-worlds in G;
- the ordering amongst the $\neg\phi$-worlds in G°_{ϕ} is exactly the ordering of the $\neg\phi$-worlds in G;
- any strict or weak preference for a ϕ-world w_1 over a $\neg\phi$-world w_2 in G is preserved in G°_{ϕ}.

Note, however, that the Darwiche and Pearl's account does not require that strict or weak preferences for $\neg\phi$-worlds over ϕ-worlds, given G, be preserved in G°_{ϕ}. More specifically, unlike on the Boutillier model, the Darwiche and Pearl account allows that a ϕ-world w_1, which is non-minimal in G and which may not be strictly preferable to some $\neg\phi$-world w_2 may be strictly preferable to w_2 relative to G°_{ϕ}. And this allows Darwiche and Pearl to deal with the problematic case SEQUENTIAL RED BIRD.

Again let $w_1 = R \wedge B$, $w_2 = \neg R \wedge B$, $w_3 = R \wedge \neg B$ and $w_4 = \neg R \wedge \neg B$. At t_1, your plausibility ordering is such that:

(t'_1) $\quad w_1 = w_2 = w_3 = w_4$.

And, again, given the Darwiche and Pearl model, upon getting the information B, at t_2 you will adjust your ordering so that:

(t'_2) $\quad w_1 = w_2 < w_3 = w_4$.

At t_3, however, upon getting the information R, the Darwiche and Pearl model allows you to adjust your ordering such that:

(t'_3) $\quad w_1 < w_2 < w_3 < w_4$.

Compare this to the ordering that is required by the Boutillier model: $w_1 < w_2 < w_3 = w_4$. The key difference here is that, upon getting the information R, Boutilier (1996) requires that you only promote the most plausible R-world. Darwiche and Pearl (1997), however, allows that each of the R-worlds may be promoted. And, given the ordering $w_1 < w_2 < w_3 < w_4$, upon getting the information $\neg B$ at t_4 you will adjust your ordering so that:

(t_4') $\quad w_3 < w_4 < w_1 < w_2$.

And so what we find is that at the end of this process you will believe $\neg B$ and you will believe R.

The postulates proposed in Darwiche and Pearl (1997), however, are not without problems. In particular, (G_9°) would seem to be subject to potential counterexamples. Thus consider the following case:[26]

> CONJUNCTIVE RED BIRD. You are initially uncertain about whether a certain animal is a red, R, or is a bird, B. Moreover you start out assuming that information about whether or not the animal is a bird, gives you no information about the animal's color. In particular, you do not have the conditional belief $R|\neg B$. You then get information that the animal is a red bird, $R \wedge B$. However, further consultation with an expert indicates that, in fact, the animal is not a bird, $\neg B$. As a result, you wind up believing that the animal is not a bird, $\neg B$, but that the animal is red, R.

Intuitively this would seem to be a rational progression of belief revision. This progression, however, is ruled out as irrational given (G_9°). To see this, note that since $\neg B$ is incompatible with $R \wedge B$, (G_9°) requires that the result of your revising, given $\neg B$, the belief state you have after incorporating $R \wedge B$, be the same as the belief state that would have resulted had you first gotten the information $\neg B$. But since you start out lacking the conditional belief $R|\neg B$, (G_9°), then, precludes your continuing to believe R once you accept $\neg B$.

The problem here would seem to be that upon getting some information, say $R \wedge B$, it may be rational for one to take various parts of that information to be independent of others in the sense that one takes it that one part is true, conditional on some other part turning out to be false. But (G_9°) precludes assuming this sort of independence. It's hard to see, however, why such assumptions of independence should be rationally precluded.

3 DYNAMIC DOXASTIC LOGIC

The models developed in Section 1 allowed us to represent an agent's beliefs, including various higher-order beliefs about the agent's own beliefs. Those models, however, failed to represent important features of an agent's doxastic state. In particular, they failed to provide any representation of an agent's conditional beliefs. The AGM models, on the other hand, allowed us to capture this feature of an agent's doxastic state. However, the AGM

26 See, e.g., Stalnaker (2009) for this type of example.

models failed to provide any representation of an agent's higher-order beliefs.

In this section, we'll begin by presenting models, in the style of Hintikka, that allow us to represent both an agent's unconditional beliefs and her conditional beliefs, and also allow us to represent various higher-order conditional and unconditional beliefs. We'll, then, consider how to add to the language dynamic operators that serve to express how an agent's beliefs, both conditional and unconditional, would be revised in light of new information.

3.1 *Doxastic Plausibility Models*

Let $\mathcal{L}$ be a propositional language including the standard Boolean connectives. In addition, we'll assume that $\mathcal{L}$ contains a binary operator $B_\alpha(\cdot,\cdot)$. As a notational simplification, we will write the second argument as a superscript, so that $B_\alpha(\phi,\psi) =_{\text{df}} B_\alpha^\psi\phi$. The intuitive gloss of $B_\alpha^\psi\phi$ will be "Alpha believes ϕ, conditional on ψ."

A *plausibility model* for $\mathcal{L}$ is a tuple $M = \langle W, \leq, [\![\cdot]\!]\rangle$. W, as before, is a set of worlds, and $[\![\cdot]\!]$ is the *interpretation function* mapping propositional letters to sets of possible worlds. $\leq$ is a *ternary* relation on W. We write this as: $w_1 \leq_w w_2$. The intuitive gloss on this is that, relative to Alpha's plausibility ordering in w, w_1 is at least as plausible as w_2. We assume that, for each w, $\leq_w$ is connected, transitive and satisfies ϕ-minimality. For ease of reference, we list these conditions again.

> CONNECTIVITY. For every $w', w'' \in \mathcal{W}$, either $w' \leq_w w''$ or $w'' \leq_w w'$.
>
> TRANSITIVITY. If $w' \leq_w w''$ and $w'' \leq_w w'''$, then $w' \leq_w w'''$.
>
> ϕ-MINIMALITY. For each $Q \subseteq W$ such that $Q \neq \emptyset$, $\{w' \in Q : w' \leq_w w'',$ for all $w'' \in Q\} \neq \emptyset$.

The truth of a sentence ϕ at a world w in a plausibility model M may be defined inductively in the standard manner. Here we simply give the condition for $B_\alpha^\psi(\phi)$.

> DEF. $[\![\phi]\!]_m =_{\text{df}} \{w : [\![\phi]\!]_m^w = 1\}$.
>
> DEF. For each $Q \subseteq W$, we let $Min_{\leq_w}(Q) =_{\text{df}} \{w' \in Q : w' \leq_w w'',$ for all $w'' \in Q\}$

We then say:

> $[\![B_\alpha^\psi\phi]\!]_m^w = 1$ just in case $Min_{\leq_w}([\![\psi]\!]_m) \subseteq [\![\phi]\!]_m$.

We'll take the notion of unconditional belief to be defined in terms of conditional belief as follows:

DEF. $B_\alpha \phi = B_\alpha^\top \phi$.

Our models here, of course, look quite a lot like the Grove models from Section 2.3.1. There are, however, some differences that are worth highlighting. One, not terribly important, difference is that, in these models, we once again take possible worlds to be primitive entities, instead of maximally consistent sets of sentences. Another, more significant, difference is that our doxastic plausibility models, unlike the Grove models, are defined for a language with an iterable operator that expresses conditional belief. Our models, then, are able to represent the conditional and unconditional beliefs of an agent who has conditional and unconditional beliefs about her own conditional and unconditional beliefs.

We can provide an axiomatic theory for our language $\mathcal{L}$ that is sound and complete with respect to the class of plausibility models so characterized. We'll call this theory C.

AXIOMS OF C

(C_1) $B_\alpha^\phi \phi$.

(C_2) $(B_\alpha^\phi \psi \wedge B_\alpha^\psi \phi) \rightarrow (B_\alpha^\phi \xi \leftrightarrow B_\alpha^\psi \xi)$.

(C_3) $(B_\alpha^{\phi \vee \psi} \phi) \vee (B_\alpha^{\phi \vee \psi} \psi) \vee (B_\alpha^{\phi \vee \psi} \xi \leftrightarrow (B_\alpha^\phi \xi \wedge B_\alpha^\psi \xi))$.

INFERENCE RULES OF C

(TI) If $(\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow \psi$ is a tautology, then $\vdash_c \phi_1 \wedge \ldots \wedge \phi_n \Rightarrow \vdash_c \psi$.[27]

(DWC) If $\vdash_c (\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow \psi$ then $\vdash_c (B_\alpha^\xi \phi_1 \wedge \ldots \wedge B_\alpha^\xi \phi_n) \rightarrow B_\alpha^\xi \psi$.

Axiom (C_1) tells us that, for every ϕ, Alpha believes ϕ conditional on ϕ. Axiom (C_2) tells us that if Alpha believes ψ conditional on ϕ, and ϕ conditional on ψ, then, for any ξ, Alpha believes ξ conditional on ϕ just in case they believe ξ conditional on ψ. Axiom (C_3) tells us that Alpha is such that, conditional on $\phi \vee \psi$, either they believe ϕ, or they believe ψ, or, for every ξ, they believe ξ just in case they believe ξ conditional on ϕ and conditional on ψ. Rule (TI) tells us that the system C is closed under logical entailment. And, finally, (DWC) tells us that Alpha's conditional beliefs are closed under entailment given C.

Let $\mathcal{P}$ be the set of plausibility models satisfying the constraints we've laid down. We, then, have the following result.

THEOREM 16. $\vdash_c \phi \Leftrightarrow \models_{\mathcal{P}} \phi$.[28]

27 It is assumed, for both rules, that if $n = 0$, then $(\phi_1 \wedge \ldots \wedge \phi_n) = \top$, and so the conditional $(\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow \psi$ is equivalent to ψ.

28 For a proof of this result see Lewis (1971). Lewis' proof concerns the logic of conditionals, but the same proof applies when we replace the conditional $\phi > \psi$ with $B_\alpha^\phi(\psi)$.

As with the case of our earlier Kripke models, we can characterize subsets of $\mathcal{P}$ by imposing constraints on the plausibility relation $\leq$.

> DEF. We say that $\leq$ is *minimally homogeneous* just in case for every w, every $z \in Min_{\leq_w}(W)$ is such that $\leq_w = \leq_z$. DEF. We say that $\leq$ is *minimally weakly homogeneous* just in case for every w, every $z \in Min_{\leq_w}(W)$ is such that if $w_1 \not<_w w_2$ then $w_1 \not<_z w_2$.

Let $\mathcal{P}_H$ be the members of $\mathcal{P}$ such that $\leq$ is minimally homogeneous. And let $\mathcal{P}_{WH}$ be the members of $\mathcal{P}$ such that $\leq$ is minimally weakly homogeneous. Now consider the following positive and negative introspection principles.

$(C_4) \quad B^{\phi}_{\alpha}(\psi) \rightarrow B_{\alpha}(B^{\phi}_{\alpha}(\psi))$.

$(C_5) \quad \neg B^{\phi}_{\alpha}(\psi) \rightarrow B_{\alpha}(\neg B^{\phi}_{\alpha}(\psi))$.

Principle (C_4) tells us that if Alpha believes ψ conditional on ϕ, then Alpha believes that they believe ψ conditional on ϕ. And (C_5) tells us that if Alpha does not believe ψ conditional on ϕ, then Alpha believes that they do not believe ψ conditional on ϕ.

We can show:

> THEOREM 17. Principle (C_5) is valid relative to the class $\mathcal{P}_{WH}$.

To see why this result holds, note that for there to be conditional beliefs in z that are not in w there need to be strict preferences amongst worlds in z that are not strict preferences in w. That is, if two worlds differ only in that certain strict preferences, $w_1 <_w w_2$, relative to w, are weak preferences, $w_1 \leq_z w_2$ and $w_2 \leq_z w_1$, relative to z, then while there will be certain conditional beliefs had at w that will not be had at z there will be no additional conditional beliefs at z. Thus, if, in accordance with the condition of minimal weak homogeneity, each of the most plausible worlds z, relative to w, imposes no strict preferences that are not imposed in w, then any conditional belief that Alpha fails to have in w, will also be such that Alpha fails to have it in z. And so, if $\leq$ is minimally weakly homogeneous, then, for any w, if Alpha fails to believe some ϕ conditional on ψ, then she will also fail to believe ϕ conditional on ψ relative to the most plausible worlds, given w, and so Alpha, at w, will believe that she fails to believe ϕ conditional on ψ.

We can also show the following.

> THEOREM 18. Principles (C_4) and (C_5) are valid relative to the class $\mathcal{P}_H$.

This result should be obvious, since any two worlds w and z, such that $\leq_w = \leq_z$, will agree about all conditional belief facts. It's worth, however, pointing out that there is no weaker condition on $\leq$ that will ensure the validity of (C_4). The reason for this is that if two worlds w and z are such that $\leq_w \neq \leq_z$, then there will be some possible assignment such that the conditional beliefs relative to w will differ from those at z. To see this, assume that we have $w_1 \leq_w w_2$ and $w_1 \not\leq_z w_2$, and so $w_2 <_z w_1$. Let ϕ and ψ be atomic sentences such that $I(\phi) = \{w_1, w_2\}$ and $I(\psi) = \{w_2\}$. Then, given our assumptions about $\leq$, relative to these assignments, we will have $[\![B^{\phi}_{\alpha}\psi]\!]^z_m = 1$ and $[\![B^{\phi}_{\alpha}\psi]\!]^w_m = 0$. Thus, minimal homogeneity is the weakest condition on $\leq$ that will ensure that, for every w, every conditional belief in w is a conditional belief in each of the minimal worlds (relative to w).

Our plausibility models can clearly be generalized to the multi-agent setting. And in such models, various conditional generalizations of the group doxastic properties of common and distributed belief can be represented. We won't, however, consider such models here. Instead, we'll move on to consider how certain dynamic operators, representing facts about how an agent's conditional beliefs would be revised given new information, may be added to our language.

3.2 *Dynamic Operators*

Let us add to our language $\mathcal{L}$ the following binary operator $[^*_\alpha]$. The rough intuitive reading of $[^*_\alpha\phi]\psi$ will be "ψ holds after Alpha revises its belief state given information ϕ." A model for our augmented language will still be a tuple $M = \langle W, \leq, [\![\cdot]\!]\rangle$, with $\leq$ subject to the same constraints. In order to characterize truth-at-a-world for formulas of the form $[^*_\alpha\phi]\psi$, however, we first need to characterize an operation on the set of models $\mathcal{P}$.

For illustrative purposes, we'll assume that rational belief revision for idealized agents works in the manner described in Boutilier (1996). That is, given an agent with a plausibility ordering over the space of worlds, such an agent will revise their belief state, given information ϕ, by minimally adjusting their plausibility ordering so that the order remains the same except that the most plausible ϕ worlds are now the most plausible worlds tout court.

> DEF. For each $w \in W$ and $Q \subseteq W$, let $C(\leq_w, Q) = \{z \in W : z \in Q$ and for all $x \in Q$, $z \leq_w x\}$.
>
> DEF. For each $w \in W$ let $\leq^{*q}_w$ be the binary relation on W such that (i) for every $z \in C(\leq_w, Q)$ and every $x \in W$, $z \leq^{*q}_w x$, and (ii) for every $x, z \in W - C(\leq_w, Q)$ $z \leq^{*q}_w x$ iff $z \leq_w x$.[29]

29 Note that given that $\leq_w$ is transitive, connected and satisfies ϕ-minimality, so too will $\leq^{*q}_w$.

We can now characterize the truth of a sentence $[^*_\alpha\phi]\psi$ at a world w in a model $M = \langle W, \leq, [\![\cdot]\!]\rangle$. We say:

> $[\![[^*_\alpha\phi]\psi]\!]^w_m = 1$ iff $[\![\psi]\!]^w_{m'} = 1$, where $M' = \langle W, \leq', [\![\cdot]\!]\rangle$ is such that for each $z \in W$ $\leq'_z = \leq^{*[\![\phi]\!]}_z$.

Operators such as B_α in our earlier Kripke models, or B^ϕ_α in our current plausibility models, function by shifting the world of evaluation. Operator $[^*_\alpha\phi]$ is, however, quite different in nature. Instead of shifting the world parameter of evaluation, $[^*_\alpha\phi]$ shifts the *model* of evaluation. We'll call operators that have the semantic function of shifting models of evaluation in this manner *dynamic operators*.

It has been shown, in van Bentham (2007), that if we add to C the following so-called reduction axioms, we get an axiomatic system that is sound and complete relative to the class of models $\mathcal{P}$ given this semantics.

(C_6) $[^*_\alpha\phi]\psi$ for each atomic ψ.

(C_7) $[^*_\alpha\phi]\neg\psi \leftrightarrow \neg[^*_\alpha\phi]\psi$.

(C_8) $[^*_\alpha\phi]\psi \wedge \xi \leftrightarrow [^*_\alpha\phi]\psi \wedge [^*_\alpha\phi]\xi$.

(C_9) $[^*_\alpha\phi]B^\psi_\alpha(\xi) \leftrightarrow [(B^\phi_\alpha\neg[^*_\alpha\phi]\psi) \wedge (B^{[^*_\alpha\phi](\psi)}_\alpha[^*_\alpha\phi]\xi)] \vee [(\neg B^\phi_\alpha\neg[^*_\alpha\phi]\psi) \wedge (B^{\phi\wedge[^*_\alpha\phi]\psi}_\alpha[^*_\alpha\phi]\xi)]$.

Axiom (C_6) tells us that atomic statements will not change their truth-value when Alpha revises its belief state given information ϕ. Axiom (C_7) tells us $\neg\psi$ will hold when Alpha revises its belief state given information ϕ just in case ψ does not hold when Alpha revises its belief state given information ϕ. Axiom (C_8) tells us that a conjunction will hold when Alpha revises its belief state given information ϕ just in case both of the conjuncts hold. These latter conditions are all fairly intuitive. Unfortunately, axiom (C_9) is much more unwieldy and lacks a simple intuitive gloss. This principle tells us that at least one of the following two conditions must obtain.

(i) Alpha believes ξ, conditional on ψ, when Alpha revises its belief state given information ϕ just in case, (a) conditional on ϕ, Alpha believes that it's not the case that if they revise their belief state given ϕ, then ψ will hold, and (b) conditional on ψ holding, if Alpha revises its belief state given ϕ, then Alpha believes that if they revise their belief state given ϕ, then ξ holds.

(ii) (a) It is not the case that, conditional on ϕ, Alpha believes that it's not the case that, if Alpha revises their beliefs given ϕ, then ψ, and (b) Alpha believes, conditional on the conjunction of ϕ and the claim that if Alpha revises their beliefs given ϕ, then ψ will hold, that if Alpha revises their beliefs given ϕ, then ξ will hold.

One thing that these reduction axioms highlight is that the addition of $[^*_\alpha]$ to our language $\mathcal{L}$ in fact adds no real expressive power. For the reduction axioms show that any formula involving such an operator is equivalent to some formula that doesn't contain this operator. The equivalent $[^*_\alpha]$-free formulas may, however, be extremely complex. The introduction of $[^*_\alpha]$, then, provides a way of expressing, in a concise manner, claims that might otherwise lack a simple expression.

I said earlier that the rough gloss on $[^*_\alpha\phi]\psi$ will be "ψ holds after Alpha revises its belief state given information ϕ." However, the semantics for $[^*_\alpha]$ encodes certain idealizing assumptions about what happens when an agent revises her beliefs given new information. In particular, the semantics we've outlined entails that if ϕ is an atomic sentence, then, when an agent gets new information ϕ, not only will the agent come to believe ϕ, but the agent will believe that they believe ϕ, and believe that they believe that they believe ϕ, and so on. Let B^n_α abbreviate n iterations of B_α. Then, if ϕ is atomic, we have that, for any n, $[\![[^*_\alpha\phi]B^n_\alpha\phi]\!]^w_m = 1$, for all worlds w and models M. New information, at least when it concerns some atomic proposition, will, on this model, be transparent to an agent.

The reason that such formulas are valid, given our semantics, is that an evaluation of the truth of $[^*_\alpha\phi]\psi$ in a model M at a world w, requires us to assess ψ at w relative to a model M' which differs from M in that the best ϕ-worlds relative to $\leq_w$ are the best worlds relative to each $\leq_z$. On this semantic theory, $[^*_\alpha\phi]$ effects a global shift on the plausibility ordering. In order to avoid the assumption that new information will not only be believed, but also believed to be believed etc., one would need to take the operator $[^*_\alpha\phi]$ to simply shift the the model M to a model M' whose plausibility ordering only differs relative to the world of evaluation w. We won't, however, look at these alternative semantic treatments here.

We've seen, then, how we can introduce a dynamic operator into our language that, in a certain sense, corresponds to the belief revision policy of Boutilier (1996). As noted earlier, though, any belief revision policy satisfying the AGM postulates can be thought of as a function mapping a belief state encoding conditional beliefs to another such state. Given any such policy f, then, we can introduce a dynamic operator $[^f]$. A formula of the form $[^f\phi]\psi$ will be true in a model M at a world w just in case ψ is true in a model M' at w, where M' is the model which shifts each $\leq_z$ to $f(\leq_z)$.

The application of dynamic operators in doxastic and epistemic logic is a rapidly developing area of study. In addition to expressing the sorts of revision that AGM was concerned with, dynamic operators can also be used to express other sorts of doxastic and epistemic changes, such as the doxastic results of so-called public announcements, which make certain pieces of information common knowledge amongst a group of agents.

The literature here is vast and growing, and a thorough survey is beyond the scope of this work. Our goal here has, instead, been to simply give a sense of how such dynamic operators function. The following, though, provides a small sample of work in this tradition: Baltag, Moss, and Solecki (1998), Segerberg (1998), Segerberg (2001), van Ditmarsche (2005), Baltag and Smets (2006a), Baltag and Smets (2006b), Rott (2006), Leitgeb and Segerberg (2007), van Bentham (2007), van Ditmarsche, van der Hoek, and Kooi (2008), Beltag and Smets (2008), van Bentham (2011) Girard and Rott (2014).

4 DOXASTIC PARADOXES

In this final section, we'll look at two doxastic paradoxes and consider, on the one hand, how some of the tools developed in the previous sections may be brought to bear to analyze these cases, and, on the other hand, how such paradoxes may serve to call into question certain assumptions made earlier about the principles governing the doxastic states of idealized agents.

4.1 *Moore's Paradox*

As Moore (1942) famously noted, there is something decidedly odd about the sentence 'ϕ and I don't believe ϕ'. What is puzzling about the case is that, while claiming that ϕ and I don't believe ϕ would seem, in some way, to be incoherent, the claim itself is perfectly consistent. There is nothing that prevents it from being true that ϕ and I don't believe ϕ.

Hintikka (1962) argued that the oddity of Moore paradoxical sentences such as $\phi \wedge \neg B_\alpha \phi$ can be explained by the fact that such claims are unbelievable for agents whose doxastic states meet certain constraints. Thus, let us assume that Alpha is an agent whose doxastic state is consistent, closed under logical consequence, and satisfies positive introspection. Given these assumptions, we can show that the following can never be true $B_\alpha(\phi \wedge \neg B_\alpha \phi)$. For assume that it is. Then since Alpha's doxastic state is closed under logical consequence we have $B_\alpha \phi$ and $B_\alpha \neg B_\alpha \phi$. And, since Alpha's doxastic state satisfies positive introspection, we have $B_\alpha B_\alpha \phi$. But, then, contrary to our assumption, Alpha's doxastic state is inconsistent.

Besides being unbelievable for certain agents, Moore paradoxical sentences have other odd features. Let us assume that our agent Alpha's idealized doxastic state is consistent, logically closed, and satisfies positive and negative introspection. Then such an agent's doxastic state may be represented by a KD45 model. More specifically, we can represent the agent's

doxastic state, as well as other facts about the world, by a particular point w in a some KD45 model $M = \langle W, R_\alpha, [\![\cdot]\!] \rangle$.

Now given an idealized agent such as Alpha, whose doxastic state may be represented by a particular point w in a some KD45 model M, we can represent the change in such an agent's doxastic state that results from getting some true information ϕ by the point w in a model M_ϕ. In particular, let $W^\phi = W \cap [\![\phi]\!]_m$, $R_\alpha^\phi = R_\alpha \cap W^\phi \times W^\phi$, and $[\![\cdot]\!]^\phi = [\![\cdot]\!] \cap W^\phi$. Then the model representing Alpha's doxastic state, after Alpha has received some true information ϕ will be $M_\phi = \langle W^\phi, R_\alpha^\phi, [\![\cdot]\!]^\phi \rangle$. We can think of M_ϕ as the model that results when one removes from M all of the worlds in which ϕ is false and then minimally adjusts the accessibility relation and valuation function.

Interestingly, there are certain sentences ϕ that, while true relative to w and M, may be false relative to w and M_ϕ. Indeed, there are certain sentences ϕ that are *guaranteed* to be false relative to w and M_ϕ. Call such sentences *self-refuting*. If ϕ is an atomic sentence, then the Moore paradoxical sentence $\phi \wedge \neg B_\alpha \phi$ is a paradigmatic case of a self-refuting sentence.

Let M be a KD45 model and w such that $[\![\phi \wedge \neg B_\alpha \phi]\!]_m^w = 1$. Let $M' = M_{\phi \wedge \neg B_\alpha \phi}$. Then it's guaranteed that $[\![\phi \wedge \neg B_\alpha \phi]\!]_{m'}^w = 0$. For, since any world in M in which ϕ is false makes $\phi \wedge \neg B_\alpha \phi$ false, each $\neg\phi$-world in M will be removed from M'. But, then, since ϕ is atomic, it follows that ϕ must be true for every world in M'. But this guarantees that we have $[\![B_\alpha \phi]\!]_{m'}^w = 1$ and so $[\![\phi \wedge \neg B_\alpha \phi]\!]_{m'}^w = 0$. Indeed, we can see that this reasoning establishes that, for each $w' \in W'$, $[\![\phi \wedge \neg B_\alpha \phi]\!]_{m'}^{w'} = 0$.

Moore paradoxical sentences, then, are not only unbelievable for certain idealized agents, they are also such that if they are true and learned to be so by such an agent then they become false.[30] While Moore paradoxical sentences may be true, their truth is, in a particular manner, unstable.

The fact that a Moore paradoxical sentence $\phi \wedge \neg B_\alpha \phi$ fails to hold for each point in the model $M_{\phi \wedge \neg B_\alpha \phi}$ is relevant for the assessment of certain principles of belief revision. For recall that, in AGM, it is assumed that $\phi \in B_\phi^*$. That is, upon revising their belief set by ϕ, an ideal agent will believe ϕ. This, of course, seems prima facie quite plausible, but Moore paradoxical sentences would seem to provide a counterexample to this claim. For, given that one's doxastic state upon learning $\phi \wedge \neg B_\alpha \phi$ is represented by $M_{\phi \wedge \neg B_\alpha \phi}$, we've seen that, upon revising one's belief set given $\phi \wedge \neg B_\alpha \phi$, this sentence will not be believed.

Now there are a few ways of responding to this worry.

First, one could grant that this is a counter-example to (B_1^*) formulated as an unrestricted principle governing belief revision. However, one could

30 Holliday and Icard (2010) show that, in a certain sense, for introspective agents all self-refuting formulas are Moore paradoxical in character.

claim that this principle, properly understood, should only apply to sentences that don't contain any belief operators. And, indeed, we can show that any sentence ϕ that doesn't contain such operators will be guaranteed to hold at any point w in M_ϕ, if it held at w in M.[31] And, furthermore, as we earlier noted, the languages that the proponents of AGM initially considered simply had no resources for talking about particular agent's beliefs or revision policies.

Another response, though, would be to argue that, properly construed, belief revision should concern *propositions*. The correct principle in the vicinity, then, is that if one learns some proposition ϕ, then one's revised belief state, in light of this, should include that proposition. One may argue, then, that if we think of the objects of belief as propositions and so revision policies as concerning which propositions one should believe, given new information, the problem for (B_1^*), so construed, disappears. For while Moore paradoxical sentences are self-refuting, Moore-paradoxical *propositions* are not.[32] For a Moore-paradoxical proposition will be time-indexed. But, if one learns between t_1 and t_2, that ϕ held at t_1 but one did not believe ϕ, this claim will remain true and may be consistently believed at t_2.

The AGM account of belief revision was formulated on the assumption that the objects of belief are sentences. Moorean phenomena, however, make it apparent that, if one wants to maintain one of the most basic principles of the theory, then the correct formulation of this theory should, instead, take the objects of belief to be propositions.

4.2 *The Burge-Buridan Paradox*

So far we've assumed that an idealized agent will have beliefs that are consistent, logically omniscient, closed under logical consequence, and that satisfy positive introspection. A close cousin of Moore's paradox, however, would seem to show that these constraints cannot be jointly satisfied if the expressive power of the language over which our doxastic models are defined is enriched in a certain manner.

31 Indeed the class of formulas with this property is larger than the class of formulas lacking any belief operators. See Holliday and Icard (2010). So there is room to enlarge the scope of this principle to certain sentences containing belief operators.

32 To be clear, by 'proposition' I mean an *eternal* proposition, i.e., something that determines a function from worlds to truth-values. The present points don't hold if one thinks that the objects of belief are temporal propositions, i.e., things that only serve to determine a function from world, time pairs to truth-values.

We'll call sentences such as the following *Burge-Buridan sentences*: "I don't believe that this sentence is true."[33] If we consider an agent who can entertain the proposition expressed by a sentence such as this, we can show, given plausible auxiliary assumptions, that this agent cannot satisfy all of the constraints we've imposed on idealized doxastic states.

So far, we have been working with propositional languages. To treat the Burge-Buridan sentence in a formal setting, however, we need to add to our language $\mathcal{L}$ a single predicate $T(\cdot)$, as well as a single term β. The intuitive interpretation of $T(\beta)$ will be that the sentence referred to by β is true. Being a sentence of our predicate language $\mathcal{L}$ may be defined in the standard manner.

We will stipulate that in our language $\mathcal{L}$ the term β refers to the sentence $\neg B_\alpha T(\beta)$. As an instance of the T-schema, then, we have:

$$\text{(T)} \quad T(\beta) \leftrightarrow \neg B_\alpha T(\beta)$$

Given a conception of logic on which the valid principles governing truth count as logical truths, it is quite plausible that (T) will count as a logical truth.[34] Assume, then, that our idealized agent Alpha satisfies logical omniscience. Then, we have $B_\alpha(T(\beta) \leftrightarrow \neg B_\alpha T(\beta))$. Now we can show that Alpha's doxastic state cannot also be consistent, logically closed and satisfy positive transparency. Our proof will proceed by cases. First, assume $\neg B_\alpha T(\beta)$. Then, it follows by closure that $B_\alpha T(\beta)$ which, of course, contradicts our assumption. Next, assume $B_\alpha T(\beta)$. Then, by closure we have $B_\alpha \neg B_\alpha T(\beta)$. But, by positive introspection, we also have $B_\alpha B_\alpha T(\beta)$. And so Alpha fails to have a consistent doxastic state.

Although Alpha cannot have a doxastic state that is consistent, logically omniscient, logically closed and positively transparent, there is no problem, in principle, with Alpha having a doxastic state that satisfies only the first three constraints. To do so, we provide a model in which these properties will be satisfied.

A doxastic model for $\mathcal{L}$ is a tuple $M = \langle W, R_\alpha, D, [\![\cdot]\!] \rangle$. W and R_α are the same as in our earlier propositional doxastic models, while D is a set of objects, and $[\![\cdot]\!]$ is a function which assigns, to propositional letters, subsets of W, to singular terms, elements of D, and, to unary predicates, functions mapping elements of w to subsets of D. Truth in such a model is defined in the obvious way.

---

33 This type of sentence was first discussed in the modern literature in Burge (1978), who attributes the paradox it raises to Buridan. For other discussion see, e.g., Burge (1984), Conee (1987), Sorensen (1988), Caie (2011), and Caie (2012).

34 Note that this instance of the T-schema is compatible with classical logic. This is established by the model given below. Even, then, if one thinks that cases such as the Liar paradox should lead us to reject certain instances of the T-schema as invalid, we don't have similar reason to reject *this* instance of the T-schema.

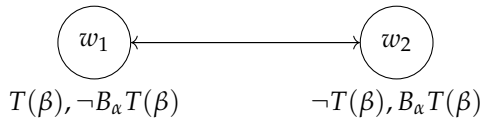

Figure 2: Modeling the Burge-Buridan sentence

Now, let $W = \{w_1, w_2, \}$, let $R_\alpha = \{\langle w_1, w_2\rangle, \langle w_2, w_1\rangle\}$ and let $[\![T]\!] = \{\langle w_1, \{\beta\}\rangle, \langle w_2, \emptyset\rangle\}$. Then we have $[\![T(\beta)]\!]_m^{w_1} = [\![\neg B_\alpha T(\beta)]\!]_m^{w_1} = 1$ and $[\![\neg T(\beta)]\!]_m^{w_2} = [\![B_\alpha T(\beta)]\!]_m^{w_2} = 1$. We may picture this model as in Figure 2.

In the model under consideration, $[\![\beta]\!] = \neg B_\alpha T(\beta)$. This corresponds to our stipulation that the sentence of $\mathcal{L}$, $\neg B_\alpha T(\beta)$, will be the denotation of the term β. Moreover, for each world w, $[\![T(\beta)]\!]_m^w = 1$ just in case $[\![\neg B_\alpha T(\beta)]\!]_m^w = 1$. This corresponds to the assumption that the T-schema for β is believed by Alpha to hold. Alpha moreover will believe all propositional logical truths, as well as any other logical truth that follows from the assumption that the T-schema holds. Since the relation R_α is serial, it follows that Alpha's doxastic state is consistent. And, as with any possible worlds doxastic model, Alpha's beliefs will be closed under logical consequence.

Thus, while idealized agents can be consistent, logically omniscient, and have beliefs that are closed under logical consequence, they cannot always be, in addition, positively transparent.

Now, there are certain ways around this result. For example, if we weaken our background logic governing the Boolean connectives, then we can show that the Burge-Buridan sentences do not preclude an idealized agent from satisfying positive and negative transparency, in addition to consistency, omniscience, and logical closure.[35] However, in order for this to be a non-ad hoc move, the weakening of the background logic would need to be sufficiently independently motivated. And whether this is so is a controversial matter.

We've considered two classes of sentences, the Moore-paradoxical and the Burge-Buridan sentences. It's worth noting, however, that the latter class is really a subclass of the former. In general, a Moore-paradoxical sentence is one that has the following form $\phi \wedge \neg B\phi$. A Burge-Buridan sentence, on the other hand, has the form $\neg BT(\beta)$, where β refers to that very sentence. On the surface, of course, this does not seem to have the form of a Moore-paradoxical sentence. However, given the plausible assumption that $T(\beta)$ and $\neg BT(\beta)$ are logically equivalent, then we get that $\neg BT(\beta)$ is, in fact, equivalent to $T(\beta) \wedge \neg BT(\beta)$. Thus a Burge-Buridan sentence, while not having the overt form of a Moore-paradoxical sentence, is equivalent to a Moore-paradoxical sentence. This sub-class of the

35 See Caie (2012) for a proof of this.

Moore-paradoxical sentences, however, have striking consequences that other members of the class of Moore-paradoxical sentences lack. For it's only with these degenerate cases of Moore-paradoxicality that we find that transparency assumptions come into conflict with other plausible principles governing idealized doxastic states.

REFERENCES

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, *50*, 510–530.

Aumann, R. (1976). Agreeing to disagree. *The Annals of Statistics*, 1236–1239.

Baltag, A., Moss, L., & Solecki, S. (1998). The logic of piblic announcements, common knowledge and private suspicions. In *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge* (pp. 43–56). Morgan Kaufmann Publishers.

Baltag, A. & Smets, S. (2006a). Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, *165*, 5–21.

Baltag, A. & Smets, S. (2006b). Dynamic belief revision over multi-agent plausibility models. *Proceedings of LOFT*, *6*, 11–24.

Barwise, J. (1988). Three views of common knowledge. In *Proceedings of the 2nd conference on theoretical aspects of reasoning about knowledge* (pp. 365–379). Morgan Kaufmann Publishers Inc.

Beltag, A. & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. *Logic and the Foundations of Game and Decision Theory*, *3*, 9–58.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal logic*. Cambridge University Press.

Bonanno, G. & Nehring, K. (2000). Common belief with the logic of individual belief. *Mathematical Logic Quarterly*, *46*(1), 49–52.

Boutilier, C. (1996). Iterated revision and minimal revision of conditional beliefs. *Journal of Philosophical Logic*, *25*, 262–305.

Burge, T. (1978). Buridan and epistemic paradox. *Philosophical Studies*, *34*, 21–35.

Burge, T. (1984). Epistemic paradox. *Journal of Philosophy*, *81*(1), 5–29.

Caie, M. (2011). *Paradox and belief* (Doctoral dissertation, University of California, Berkeley).

Caie, M. (2012). Belief and indeterminacy. *The Philosophical Review*, *121*(1), 1–54.

Chellas, B. (1980). *Modal logic: An introduction*. Cambridge University Press.

Colombetti, M. (1993). Formal semantics for mutual belief. *Artificial Intelligence*, *63*(341-353).

Conee, E. (1987). Evident, but rationally unacceptable. *Australasian Journal of Philosophy*, *65*, 316–326.

Darwiche, A. & Pearl, J. (1997). On the logic of iterated belief revision. *Artificial Intelligence*, *89*, 1–29.

Fagin, R., Halpern, J., & Vardi, M. (1991). A model-theoretic analysis fo knolwedge. *Journal of ACM*, *38*(2), 382–428.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. MIT Press.

Gärdenfors, P. & Makinson, D. (1988). Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of the 2nd conference on theoretical aspects of reasoning about knowledge* (pp. 83–95). San Francisco: Morgan Kaufmann.

Girard, P. & Rott, H. (2014). *Belief revision and dynamic logic*. (ms.)

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, *17*, 157–170.

Halpern, J. & Moses, Y. (1984). Knowledge and common knowledge in a distributed environment. In *Proceedings of the 3rd acm confernece on principles of distributed computing* (pp. 50–61).

Halpern, J. & Moses, Y. (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, *54*(319-379).

Halpern, J., Moses, Y., & Vardi, M. (1995). *Reasoning about knowledge*. MIT Press.

Harper, W. (1976). Rational conceptual change. *PSA*, *2*, 462–494.

Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Cornell University Press.

Holliday, W. H. & Icard, T. F. (2010). Moorean phenomena in epistemic logic. In L. Beklemishev, V. Goranko, & V. Shehtman (Eds.), *Advances in modal logic* (Vol. 8, pp. 178–199).

Huber, F. (2019). Ranking theory. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

Hughes, G. E. & Cresswell, M. J. (1996). *A new introduction to modal logic*. Psychology Press.

Leitgeb, H. & Segerberg, K. (2007). Dynamic doxastic logic: Why, how and where to? *Synthese*, *155*, 167–190.

Levi, I. (1977). Subjunctives, dispositions and chances. *Synthese*, *34*, 423–455.

Levi, I. (1988). Iteration of conditionals and the ramsey test. *Synthese*, *76*, 49–81.

Lewis, D. (1969). *Convention*. Cambridge University Press.

Lewis, D. (1971). Completeness and decidability of three logics of counterfactual conditionals. *Theoria*, *37*(1), 74–85.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1974). Radical interpretation. *Synthese*, *23*, 331–44.

Lewis, D. (1999). Reduction of mind. In *Papers on metaphysics and epistemology*. Cambridge University Press.

Lewis, D. (2000). Logic for equivocators. In *Papers in philosophical logic*. Cambridge University Press.

Lismont, L. & Mongin, P. (1994). On the logic of common belief and common knowledge. *Theory and Decision*, *37*, 75–106.

Moore, G. E. (1942). A reply to my critics. In P. Schilpp (Ed.), *The philosophy of g.e. moore* (Vol. 4, pp. 535–677). The Libraray of Living Philosophers. Northwester University.

Rott, H. (2006). Shifting priorities: Simple representations of twenty-seven iterated theory change operations. In H. Lagerlund, S. Lindstrom, & R. Sliwinski (Eds.), *Modality matters* (Vol. 53, pp. 359–385). Uppsala Philosophical Studies.

Segerberg, K. (1998). Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic*, *39*, 287–306.

Segerberg, K. (2001). The basic dynamic doxastic logic of agm. In M. Williams & H. Rott (Eds.), *Frontiers in belief revision* (pp. 57–84). Dordrecht: Kluwer.

Shoemaker, S. (1996a). Moore's paradox and self-knowledge. In *The first-person perspecive and other essays* (pp. 74–96). Cambridge University Press.

Shoemaker, S. (1996b). On knowing one's own mind. In *The first-person perspecive and other essays* (pp. 25–49). Cambridge University Press.

Sorensen, R. (1988). *Blindspots*. Oxford University Press.

Stalnaker, R. (1984). *Inquiry*. MIT Press.

Stalnaker, R. (2009). Iterated belief revision. *Erkenntnis*, *70*(2), 189–209.

van Bentham, J. (2007). Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, *17*(2), 129–155.

van Bentham, J. (2011). *Logical dynamics of information and interaction*. Cambridge University Press.

van Ditmarsche, H. (2005). Prolegomena to dynamic logic for belief revision. *Synthese*, *147*, 229–275.

van Ditmarsche, H., van der Hoek, W., & Kooi, B. (2008). *Dynamic epistemic logic*. Springer.

Williamson, T. (2000). *Knowledge and it's limits*. Oxford University Press.