In Lindahl L., Needham, P., and Sliwinski, R., *Foor Good Measure*, Uppsala Philosophy Studies 46, Uppsala University, Department of Philosophy, Uppsala, Sweden 1997.

EXTENDING DYNAMIC DOXASTIC LOGIC: ACCOMMODATING ITERATED BELIEFS AND RAMSEY CONDITIONALS WITHIN DDL*

Sten Lindström and Wlodek Rabinowicz

In this paper we distinguish between various kinds of *doxastic theories*. One distinction is between *informal and formal* doxastic theories. AGM-type theories of belief change are of the former kind, while Hintikka's logic of knowledge and belief is of the latter. Then we distinguish between *static theories* that study the unchanging beliefs of a certain agent and *dynamic theories* that investigate not only the constraints that can reasonably be imposed on the doxastic states of a rational agent but also rationality constraints on the changes of doxastic state that may occur in such agents. An additional distinction is that between *non-introspective* theories and *introspective* ones. Non-introspective theories investigate agents that have opinions about the external world but no higher-order opinions about their own doxastic states. Standard AGM-type theories as well as the currently existing versions of Segerberg's *dynamic doxastic logic* (DDL) are non-introspective. Hintikka-style doxastic logic is of course introspective but it is a static theory. Thus, the challenge remains to devise doxastic theories that are both dynamic and introspective. We outline the semantics for a *truly introspective* dynamic doxastic logic, i.e., a dynamic doxastic logic that allows us to describe agents who have both the ability to form higher-order beliefs and to reflect upon and change their minds about their own (higher-order) beliefs. This extension of DDL demands that we give up the Preservation condition on revision. We make some suggestions as to how such a non-preservative revision operation can be constructed. We also consider extending DDL with conditionals satisfying the Ramsey test and show that Gärdenfors' well-known impossibility result applies to such a framework. Also in this case, Preservation has to be given up.

1. Static doxastic logic: Hintikka's logic of knowledge and belief

The modern development in *doxastic logic* (the logic of *belief*) and *epistemic logic* (the logic of *knowledge)* started with Jaakko Hintikka's seminal book *Knowledge and Belief* (1962). Hintikka's basic idea was to apply the possible worlds semantics for modal logic to so-called propositional attitude constructions like "believes that" and "knows that". According to Hintikka to ascribe

* Ancestors of this paper have been presented at the Swedish Collegium for Advanced Studies in the Social Sciences (SCASSS), the workshop "Logic and Rational Argumentation" at Uppsala University, March 4-5, 1997 and the "Friday Philosophy Seminar" at Umeå University. We are grateful to the participants for their comments and friendly criticism. We owe special thanks to John Cantwell, Sven Ove Hansson, Ingvar Johansson, Isaac Levi, Tor Sandqvist, and Krister Segerberg.

knowledge to a person x is to invoke the idea of a set of "epistemically" possible worlds (with respect to the person x). These worlds, the person's *epistemic alternatives* are precisely the worlds that are compatible with everything that the person knows (in the actual world). Although they all agree with respect to what the person knows, they still differ in ways that make them incompatible with each other. The analogy with necessity leads to the following principle for knowledge:

> x *knows that* α (in the actual world) if and only if, in every possible world compatible with what x knows it is the case that α.

Similarly, the concept of belief appeals to the idea of a set of "doxastically" possible worlds (the agent's *doxastic alternatives*). The corresponding principle is:

> x *believes that* α if and only if, in every possible world compatible with what x believes it is the case that α.

It is natural to assume:

(i) *Knowledge implies truth.* Hence, the actual world is itself one of the possible worlds that is compatible with everything the agent knows in the actual world. That is, the actual world is one of the epistemic alternatives for the agent in the actual world.

(ii) *Knowledge implies belief.* Hence, if a possible world is compatible with everything the agent believes, then it is compatible with everything he knows. That is, the set of doxastic alternatives is a subset of the set of epistemic alternatives.

In the formal development of doxastic/epistemic logic, Hintikka extends a language of sentential or predicate logic with special operators of knowledge and belief:[1]

Kα	for	"the agent knows that α".
Bα	for	"the agent believes that α".

Writing

> wR_Kv for "the world v is compatible with everything that the agent knows in the world w".

---

[1] The philosophically most interesting aspects of Hintikka's epistemic/doxastic logic are concerned with the interplay between propositional attitudes and *quantifiers*: for instance, Hintikka's analyses of the *de dicto-de re* distinction, *knowing who* constructions, and *interrogatives*. This dimension falls outside the scope of the present paper, since we are concerned with *sentential* doxastic logic only.

wR_Bv for "the world v is compatible with everything that the agent believes in the world w".

one gets the following truth-clauses for these operators:

Kα is true in w iff for every v, if wR_Kv, then α is true in v.
Bα is true in w iff for every v, if wR_Bv, then α is true in v.

For each world w, the sets $K(w) = \{v: wR_Kv\}$ and $B(w) = \{v: wR_Bv\}$ are the sets of the agent's *epistemic* and *doxastic alternatives* in the world w, respectively. We assume that:

(i) for every w, $w \in K(w)$
(ii) for every w, $B(w) \subseteq K(w)$.

(i) says that the world w is itself compatible with everything that the agent knows to be true in that world. This is so, since knowledge implies truth. (ii) says that if a world is compatible with everything the agent believes, then it is also compatible with everything he knows. The motivation for this principle is that the set of all propositions that constitute the agent's knowledge constitutes a (possibly proper) subset of the set of the agent's beliefs. Hence, to be compatible with all that is believed is at least as stringent a requirement on a world as is the one of being compatible with all that is known.

Writing $\models \alpha$ for α being logically valid, i.e., true in every world in every model, one gets the following minimal set of principles for Hintikka-style epistemic/doxastic logic:

(1) $\models \mathbf{K}(\alpha \rightarrow \beta) \rightarrow (\mathbf{K}\alpha \rightarrow \mathbf{K}\beta)$
(2) $\models \mathbf{B}(\alpha \rightarrow \beta) \rightarrow (\mathbf{B}\alpha \rightarrow \mathbf{B}\beta)$
(3) $\models \mathbf{K}\alpha \rightarrow \alpha$ *(Veridicality of Knowledge)*
(4) $\models \mathbf{K}\alpha \rightarrow \mathbf{B}\alpha$
(5) If $\models \alpha$, then $\models \mathbf{K}\alpha$
(6) If $\models \alpha$, then $\models \mathbf{B}\alpha$

From now on, our main subject will be the concept of belief and we shall return to knowledge only occasionally. The principles (2) and (6), although by no means uncontroversial, will constitute our basic logic for the belief operator **B**. By imposing additional requirements on the relation R_B, one can ensure that some or all of the following principles are also satisfied:

(7) $\neg\mathbf{B}\bot$ *(Consistency)*
(8) $\mathbf{BB}\alpha \rightarrow \mathbf{B}\alpha$ *(Veridicality of Positive Introspection)*
(9) $\neg\mathbf{B}\bot \rightarrow (\mathbf{B}\neg\mathbf{B}\alpha \rightarrow \neg\mathbf{B}\alpha)$ *(Veridicality of Negative Introspection)*
(10) $\mathbf{B}\alpha \rightarrow \mathbf{BB}\alpha$ *(Positive Introspection)*

(11) $\neg\mathbf{B}\alpha \rightarrow \mathbf{B}\neg\mathbf{B}\alpha$ *(Negative Introspection)*

(7), for example, says that a logical contradiction (symbolised by $\perp$) is never believed. (11) says that if the agent does not believe that α, then he believes that he does not believe that α. These and other principles for iterated beliefs will be discussed in due course.

In terms of the operators **K** and **B** one can define the dual operators **k** and **b** by means of:

$$\mathbf{k}\alpha =_{df} \neg\mathbf{K}\neg\alpha$$
$$\mathbf{b}\alpha =_{df} \neg\mathbf{B}\neg\alpha.$$

The intended readings of these are given by:

kα: "It is possible, for all that the agent knows, that α".
bα: "It is compatible with everything the agent believes that α".

2. AGM-type theories of belief change

In a doxastic logic of Hintikka-type it is possible to represent and reason about the *static* aspects of an agent's beliefs about the world: it studies the various *constraints* that one might think that a rational agent or a set of rational agents should satisfy. Such a logic cannot, however, be used to reason about doxastic change, i.e., the various kinds of *doxastic actions* that an agent may perform. The agent may, for instance, *revise* his beliefs by adding a new piece of information, while at the same time making adjustments to his stock of beliefs in order to preserve consistency. Or he may *contract* his beliefs by giving up a proposition that he formerly believed. Such operations of doxastic change are studied in the theories of *rational belief change* that started with the work of Alchourrón, Gärdenfors and Makinson in the 80's: the so-called AGM-approach. According to AGM, there are three basic types of doxastic actions:

Expansion: The agent adds a new belief α to his stock of old beliefs without giving up any old beliefs. If G is the set of old beliefs, then G+α denotes the set of beliefs that results from *expanding* G with α. To expand is dangerous, since G+α might very well be logically inconsistent; and inconsistency is something that we should try to avoid in our beliefs.

Contraction: The agent gives up a proposition α that was formerly believed. This requires that he also gives up other propositions that *logically imply* the proposition α. We use G–α to denote the result of contracting α from the old set G of beliefs.

Revision: The revision $G*\alpha$ of the set G with the new information α is the result of adding α to G in such a way that consistency is preserved whenever possible. The idea is that $G*\alpha$ should be a set of beliefs that preserves as much as possible of the information that is contained in G and still contains α. $G*\alpha$ should be a minimal change of G that incorporates α.

The following is an important guiding principle when revising and contracting belief sets:

The Principle of Conservatism: Try not to give up or add information to your original belief set unnecessarily.

Within the AGM approach, the agent's belief state is represented by his *belief set*, i.e., the set G of all sentences α such that the agent believes that α. An underlying deductive logic is assumed and the operation of revision is assumed to satisfy the following axioms:

Gärdenfors' axioms for revision:

(R0) Every belief set G is closed under logical consequence (of the underlying deductive logic).

(R1) $\alpha \in G*\alpha$, that is, the new information α is contained in $G*\alpha$

(R2) If α is consistent with G, then $G*\alpha = G+\alpha$, that is, $G*\alpha$ is the smallest logically closed set Γ such that $G \cup \{\alpha\} \subseteq \Gamma$.

(R3) $G*\alpha$ is consistent if and only if α is consistent.

(R4) If α and β logically equivalent, then $G*\alpha = G*\beta$.

(R5) If β is consistent with $G*\alpha$, then $G*(\alpha \wedge \beta) = (G * \alpha)+\beta$.

AGM also contains axioms for *contraction* (omitted here) as well as the following bridging principles:

$G*\alpha = (G-\!-\neg\alpha)+\alpha$ *(The Levi identity)*

$G-\alpha = (G*\alpha) \cap (G*\neg\alpha)$ *(The Harper identity)*

The Levi identity says that the result of revising the belief set G by the sentence α equals the result of first making room for α by (if necessary) contracting G with $\neg\alpha$ and then expanding the result with α. The Harper identity says that the result of contracting α from G is the common part of G revised with α and G revised with $\neg\alpha$.

Adam Grove's (1988) possible worlds modeling for AGM

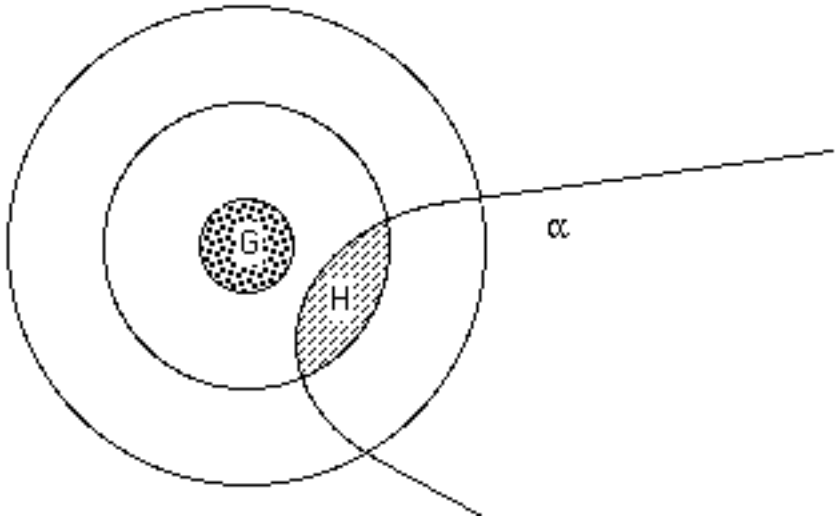
In his (1988) paper, Grove presents two closely related possible worlds modelings of AGM-type belief revision, one in terms of a family of "spheres" around the agent's belief set (or theory) G and the other in terms of an epistemic entrenchment ordering of propositions.[2] Intuitively, a proposition α is at least as entrenched in the agent's belief set as another proposition β if and only if the following holds: provided the agent would have to revise his beliefs so as to falsify the conjunction $\alpha \wedge \beta$, he should do it in such a way as to allow for the falsity of β.

The "sphere"-terminology is natural when one looks upon belief sets (theories) and propositions as being represented by sets of possible worlds. Grove's spheres may be thought of as possible "fallback" theories relative to the agent's original theory: theories that he may reach by deleting propositions that are not "sufficiently" entrenched (according to standards of sufficient entrenchment of varying stringency). To put it differently, fallbacks are theories that are closed upwards under entrenchment: if T is a fallback, α belongs to T, and β is at least as entrenched as α, then β also belongs to T. The entrenchment ordering can be recovered from the family of fallbacks by the definition: α is at least as entrenched as β if and only if α belongs to every fallback to which β belongs.

Representing theories and propositions as sets of possible worlds, the following picture illustrates Grove's family of spheres around a given theory G and his definition of revision. Notice that the spheres around a theory are "nested", i.e., simply ordered. For any two spheres, one is included in the other. Grove's family of spheres closely resembles Lewis' sphere semantics for counterfactuals, the main difference being that Lewis' spheres are "centered" around a single world instead of a theory (a set of worlds).

---

[2] Actually, Grove works with an ordering of epistemic *plausibility*. But as Gärdenfors (1988, sect. 4.8) points out, the notions of plausibility and entrenchment are interdefinable. Thus, a proposition α is at least as plausible as a proposition β given the agent's beliefs if and only if non-β is at least as entrenched as non-α in the agent's belief set. The notion of epistemic entrenchment is primarily defined for the propositions that belong to the agent's belief set: one adopts the convention that propositions that are not believed by the agent are minimally entrenched. On the other hand, the notion of plausibility primarily applies to the propositions that are incompatible with the agent's beliefs (the propositions that are compatible with what he believes are all taken to be equally and maximally plausible). Thus, this is a notion of *conditional* plausibility. α is at least as plausible as β in this sense iff the following holds: on the condition that I would have to revise my beliefs with $\alpha \vee \beta$, I should change them in such a way as to allow for α.



The shaded area H represents the revision of G with a proposition α. The revision of G with α is defined as the strongest α-permitting fallback theory of G expanded with α. In the possible worlds representation, this is the intersection of α with the smallest sphere around G that is compatible with α. (Any revision has to contain the proposition we revise with. Therefore, if α is logically inconsistent, the revision with α is taken to be the inconsistent theory.)

Lindström & Rabinowicz: non-deterministic revision

In a series of papers, Lindström and Rabinowicz proposed a generalization of the AGM approach according to which belief revision was treated as a relation $GR_{\alpha}H$ between theories (belief sets) rather than as a function on theories.[3] The idea was to allow for there being several equally reasonable revisions of a theory with a given proposition. Thus, $GR_{\alpha}H$ means that H is one of those reasonable revision of the theory G with the new information α. AGM, of course, assumes that belief revision is functional (or deterministic), that is,

If $GR_{\alpha}H$ and $GR_{\alpha}H'$, then H = H'.

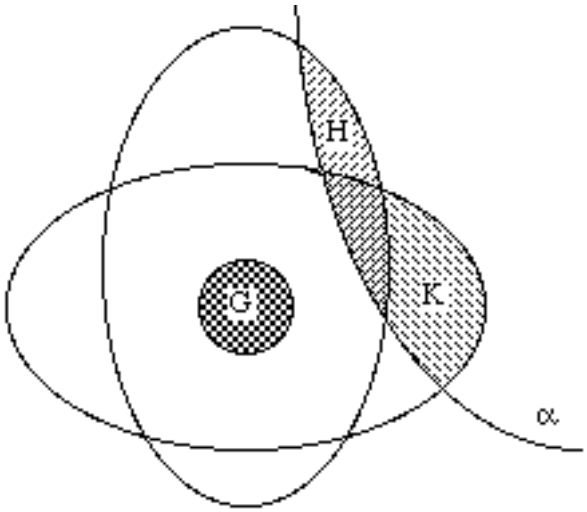
Given this assumption, one can define:

$G*\alpha$ = the theory H such that $GR_{\alpha}H$.

The relational notion of belief revision results from weakening epistemic entrenchment by not assuming it to be *connected.* In other words, we allow that some propositions may be incomparable with respect to epistemic entrenchment. As a result, the family of fallbacks around a given theory will no longer have to be nested. It will no longer be a family of spheres but rather a family of

[3] Cf. Lindström and Rabinowicz (1989), (1990), (1992) and Rabinowicz and Lindström (1984).

"ellipses". This change opens up the possibility for several different ways of revising a theory with a given proposition.



In this figure, the two ellipses represent two different fallback theories for G, each of which is a strongest α-permitting fallback. Consequently, there are two possible revisions of G with α: each one of H and K is the intersection of α with a strongest α-permitting fallback.

3. Dynamic doxastic logic

The theories of belief change developed within the AGM-tradition are not doxastic logics in the formal sense, but rather *informal axiomatic theories* of belief change. Instead of characterizing the models of belief and belief change in a formalized object language, the AGM-approach uses a natural language — like ordinary mathematical English — to characterize the mathematical structures that are under study. Recently, however, various authors such as van Benthem and Maarten de Rijke have suggested representing epistemic change within a formal logical language: a dynamic modal logic. Inspired by these suggestions Krister Segerberg has developed a very general logical framework for reasoning about doxastic change: *dynamic doxastic logic* (DDL).[4] This framework may be seen as an extension of standard Hintikka-style doxastic logic with dynamic operators representing various kinds of transformations of the agent's doxastic state. Segerberg writes $+\alpha$, $*\alpha$, and $-\alpha$, respectively, for the *doxastic actions* of *expanding*, *revising* and *contracting* the agent's beliefs with (the information contained in) the sentence α. Hence, $+\alpha$ denotes the action of simply adding α

4 Cf. Segerberg (1995), (1996a) and (1996b).

to the stock of beliefs (without checking for consistency). $*\alpha$ is the action of adding α, while at the same time modifying the belief state in such a way that consistency is preserved, whenever possible. $-\alpha$, finally, means that the agent changes his belief state in such a way that any belief that α is given up. In DDL, one uses the following notation with the following informal meaning:

$[+\alpha]\beta$ "If the agent were to *expand* his beliefs with α, then it would be the case that β".

$[*\alpha]\beta$ "If the agent were to *revise* his beliefs with α, then it would be the case that β".

$[-\alpha]\beta$ "If the agent were to *contract* his beliefs with α, then it would be the case that β".

As long as the agent's belief state is not part of the world, doxastic actions do not affect the world. Thus, if β expresses a *worldly proposition*, i.e., a proposition that only concerns the (external) world, then we should expect $[+\alpha]\beta \leftrightarrow \beta$ to hold, and similarly for the other doxastic actions. So the interesting case is the one when β contains epistemic operators. In particular, we are interested in statements of the forms: $[+\alpha]\mathbf{B}\beta$, $[*\alpha]\mathbf{B}\beta$, $[-\alpha]\mathbf{B}\beta$. For example,

$$[*\alpha]\mathbf{B}\beta$$

means: if the agent were to revise his beliefs with α, he would believe β. In the AGM approach this kind of statement is expressed as:

$$\beta \in G*\alpha,$$

where G refers to the agent's current belief set, i.e., the set of all sentences σ such that the agent believes σ, and $G*\alpha$ is the belief set that results from revising G by α. In AGM, $[+\alpha]\mathbf{B}\beta$ and $[-\alpha]\mathbf{B}\beta$ correspond to, respectively, $\beta \in G+\alpha$ and $\beta \in G-\alpha$.

DDL allows for the possibility of belief change being *nondeterministic*: there may be many different ways for the agent of revising his beliefs with α (Cf. Lindström & Rabinowicz above). Hence, we must distinguish between:

$[*\alpha]\mathbf{B}\beta$ "If the agent were to revise his beliefs with α, he *would* believe that β".

$\langle*\alpha\rangle\mathbf{B}\beta$ "If the agent were to revise his beliefs with α, he *might* believe that β".

$\langle*\alpha\rangle$ is definable in terms of $[*\alpha]$ in the following way:

$$\langle*\alpha\rangle\beta = \neg[*\alpha]\neg\beta.$$

In the same way, one can define $<+\alpha>$ and $<-\alpha>$. For theories like the original AGM-theory in which belief change is deterministic, one would have $<*\alpha>\beta \leftrightarrow [*\alpha]\beta$, and similarly for contraction. Expansion, is of course always deterministic, i.e., $<+\alpha>\beta \leftrightarrow [+\alpha]\beta$.

4. The various types of doxastic agents

In this section, we are going to distinguish between five types of doxastic agents and accordingly between five types of doxastic theories. This categorization is based on the following distinctions: First, we distinguish between *static* and *dynamic* agents. A static agent has no capacity to change his beliefs. A dynamic agent can change his belief state, for example, in response to new information. Another distinction is that between a *non-introspective* and an *introspective* agent. An agent is non-introspective, if he can only form beliefs about the external world and has no capacity to form beliefs about his own belief states. An introspective agent is an agent who can also form beliefs about his own belief states. Finally, we distinguish between introspective dynamic agents that can only receive informational inputs that concern the external world and such agents that can also receive inputs about their own belief states. In terms of these distinctions, we can define:

> *Type 1 agents:* Non-introspective static agents. (Examples: a fixed data base that you can query for information. An industrial robot that cannot learn from experience.)
> *Corresponding theories:* Hintikka-type doxastic logic that does not allow for the nesting of belief-operators.
>
> *Type 2 agents:* Introspective static agents.
> *Corresponding theories:* Hintikka-type doxastic and epistemic logics.
>
> *Type 3 agents:* Non-introspective dynamic agents. (Examples: Dynamic data bases. Neural networks that can learn from experience.)
> *Corresponding theories*: (Most) theories of Belief Revision within the AGM-tradition. Segerberg's dynamic doxastic logic, where the dynamic action statements of the form $[O\alpha]\mathbf{B}\beta$, where O represents some kind of doxastic action, are restricted in such a way that the sentences α and β are required to be Boolean (i.e., express worldly propositions).[5]

[5] A sentence is said to be *Boolean* if it is built up from propositional letters by means of classical ("Boolean") sentential connectives only.

Type 4 agents: Introspective dynamic agents whose doxastic inputs are limited to propositions about the (external) world.
Corresponding theories. Dynamic doxastic logics, where the dynamic action statements of the form [Oα]**B**β are restricted in such a way that the sentences α is required to be Boolean (i.e., express worldly propositions). β is allowed to be any sentence, also one involving doxastic and dynamic operators.

Type 5 agents: Introspective dynamic agents whose doxastic inputs may also be propositions about the agent's own doxastic states.
Corresponding theories: Dynamic doxastic logics, where the dynamic action statements of the form [Oα]**B**β are unlimited: both α and β are allowed to be any sentences.

Type 1 agents (Non-introspective static agents)
The most primitive kind of agent that we shall consider is a static doxastic agent that lacks any power of introspection. Such an agent we say is of *type 1*. A type 1 agent has beliefs about the external world but no ability to change his beliefs or to form higher-order beliefs about his own beliefs. The total belief state of a type 1 agent can be represented by a *belief set*, i.e., the set of all propositions that are believed by the agent:

The belief set of the agent = {P: the agent believes that P}.

All the propositions of an agent of type 1 are what we might call *worldly propositions:* they concern the external world only and are logically independent of the doxastic state of the agent (the agent might have beliefs about the doxastic states of *other* agents in so far as these are considered to be part of the external world.) Roughly speaking, a type 1 agent can have a belief of type "Snow is white" but not one of the kind "I believe that snow is white". When the theoretician describes such an agent he will make use of statements of the type:

Jones believes that snow is white

but not:

Jones believes that he believes that snow is white.

A doxastic logic for type 1 agents.
We consider a propositional language with *Boolean formulas* built up from a non-empty set of propositional letters by means of the usual sentential connec-

tives $\neg$, $\wedge$, $\vee$, $\rightarrow$. The set of formulas is the smallest set satisfying the conditions:

(i) Boolean formulas are formulas.
(ii) If α is a Boolean formula, then $\mathbf{B}\alpha$ is a formula.
(iii) If α, β are formulas, then $\neg\alpha$, $(\alpha \wedge \beta)$, $(\alpha \vee \beta)$, $(\alpha \rightarrow \beta)$ are formulas.

Observe that this language does not allow for the iteration of the belief operator **B**. Intuitively, Boolean formulas express worldly propositions and the belief operator (or belief predicate) should apply to worldly propositions only.

The models are structures of the form $\mathfrak{M}$ = <W, Bel, V>, where:

(i) W is a non-empty set (of possible worlds). W represents all the possible states the world could be in.
(ii) Bel $\subseteq$ W. Intuitively, Bel is the set of possible worlds that are compatible with everything that the agent (actually) believes about the world.
(iii) V is a valuation function assigning a subset V(p) of W to every propositional letter p.

We define the notion of *truth* of a formula α relative a model $\mathfrak{M}$ and a world w (in $\mathfrak{M}$) as follows (we suppress the reference to M):

(i) $w \models p$ if and only if $w \in V(p)$
(ii) $w \models \neg\alpha$ iff not: $w \models \alpha$. (and so on, for the other sentential connectives)
(iii) $w \models \mathbf{B}\alpha$ if and only if Bel $\subseteq \|\alpha\|$, where $\|\alpha\| = \{w: w \models \alpha\}$. That is, $w \models \mathbf{B}\alpha$ if and only if α is true in every possible world that is compatible with the agent's beliefs.

A formula is *valid* if it is true in every world in every model. The set of valid formulas can easily be proven to be equal to the smallest set Σ satisfying the following conditions:

(T) All substitution instances of sentential tautologies.
(K) All instances of $\mathbf{B}(\alpha \rightarrow \beta) \rightarrow (\mathbf{B}\alpha \rightarrow \mathbf{B}\beta)$ are in Σ.
(MP) If $\alpha \in \Sigma$ and $(\alpha \rightarrow \beta) \in \Sigma$, then $\beta \in \Sigma$.
(N) If $\alpha \in \Sigma$, then $\mathbf{B}\alpha \in \Sigma$.

Type 2 agents (Introspective static agents).
These are static agents that can have (and in fact have) beliefs about their own beliefs. We shall only consider type 2 agents of level ω (omega), that is, agent's

that have not only beliefs about the world, but also beliefs about his own beliefs about the world, beliefs about such beliefs, etc. For any n=1, 2,... such an agent has beliefs of order n. The *belief state* of a type 2 agent can also be described by a belief set:

{P: the agent believes P}.

But observe that the belief set of a type 2 agent will contain propositions (or statements) that themselves concern the agents beliefs. For instance, the belief set may contain the proposition:

Believes(Snow is white) and ¬Believes (Grass is green).

This proposition is an element of the belief set just in case:

The agent believes that: he believes that snow is white and does not believe that Grass is green.

Moore's paradox: Consider, an agent of type 2 whose belief set contains

(1) It is raining but I don't believe it.

It is true about such an agent that:

(2) The agent believes that: Snow is white and the agent does not believe that snow is white.

We may ask whether such an agent is irrational. Is he inconsistent?
Remark: Compare the following two situations:

(1) Jones believes: It is raining and **I** don't believe it.
(2) Smith believes: It is raining and **Smith** does not believe it

It seems that Jones is worse off than Smith. From (1) we may conclude:

(3) Jones believes that it is raining.
(4) Jones believes that he (himself) does not believe that it is raining.

It seems odd — if not outright inconsistent — for a self-reflective agent to be in a doxastic state characterized by (3) and (4). Consider now Smith, the amnesiac, who does not know that he is Smith. Couldn't he be in the situation (2) without being irrational or inconsistent? Finally, suppose that Jones is an amnesiac (who believes that he is Smith). Would that change anything in the case (1)? In our opinion, the answer is: No.

Hintikka-style doxastic logic may be viewed as a logic for doxastic agents of type 2.

Type 3 agents (Non-introspective dynamic agents)
Like a type 1 agent, a type 3 agent has beliefs about the world but no beliefs about his own beliefs. However, a type 3 agent also has dispositions to change his beliefs about the world in response to new information. The theories developed within the AGM-tradition can be viewed as theories that concern type 3 agents: they study the doxastic states and the transformations of doxastic states of an agent that is, so to speak, placed "outside of the world". The agent's beliefs are about a constant external world that is not affected by the changes in the agent's doxastic state. All propositions believed and all epistemic inputs concern this "mind independent reality". Such an agent has no introspective capacities: he has no beliefs about his own beliefs or his own doxastic dispositions.

Type 4 and type 5 agents (Introspective dynamic agents)
Gärdenfors' axioms appear to be reasonable as long as the sentences of the formal language are taken to represent *worldly propositions*. However, in the presence of doxastic operators in the object language, these axioms will lead to paradoxical results.[6] Suppose that α is a proposition (say "It is raining in Lund at this moment") of which I have no firm belief with respect to its truth:

Hence,

(1) $\alpha \notin G$ and $\neg\alpha \notin G$.

Suppose also that I correctly believe that I do not believe α, i.e.,

(2) $\neg\mathbf{B}\alpha \in G$.

Then, by (R2), the revision of G with α is just the result of expanding G with α, so:

(3) $\neg\mathbf{B}\alpha \in G{*}\alpha$.

But, by (R1),

(4) $\alpha \in G{*}\alpha$.

But, then by (R0), i.e., logical closure,

(5) $\alpha \wedge \neg\mathbf{B}\alpha \in G{*}\alpha$,

that is in the belief state represented by $G{*}\alpha$ it is true that:

The agent believes that: α and that he does not believe α.

---

6 That this is so was pointed out by Levi (1988) and Fuhrmann (1989).

This is highly counterintuitive (Moore's Paradox). In case the underlying logic for **B** satisfies positive introspection, we get:

(6) $\mathbf{B}\alpha \wedge \neg\mathbf{B}\alpha \in G{*}\alpha$,

which is contrary to (R3).

The culprit seems to be (R2). The part of (R2) that we actually used was the following principle,

Preservation:
If α is consistent with G, then if $\beta \in G$, $\beta \in G{*}\alpha$.

It is the application of this principle to doxastic propositions β that seems to create the paradoxical result. Hence, the following weakening of Preservation suggests itself.

Weak Preservation.
If α is consistent with G, β is Boolean and $\beta \in G$, then $\beta \in G{*}\alpha$.

Our conclusion is that Gärdenfors' axioms have to be modified as soon as one considers dynamic doxastic theories for *introspective agents*. If revision is not to be paradoxical for such agents, certain beliefs about one's own beliefs need to be abandoned when one receives new information. For instance, when revising with α one has to give up the belief that one does not believe α. Otherwise, one will automatically acquire a false belief in the process. Hence, full expansion has to be abandoned for introspective agents.

5. Dynamic doxastic logic unlimited: DDL for introspective agents

As developed so far, DDL has been a theory for agents without introspection. Here we wish to remove this limitation. Suppose we have a language with a belief operator **B** and with different kinds of doxastic dynamic operators of the form [A], where A stands for a type of doxastic action. The dynamic operators that we are especially interested in are revisions, expansions and contractions: $[{*}\alpha]$ for revision, $[+\alpha]$ for expansion, and $[-\alpha]$ for contraction, where α is sentence that we revise with, expand with or contract with, respectively. Suppose now that both **B** and the dynamic operators are allowed to operate on arbitrary well-formed formulas, *without restriction*. In particular, we allow formulas expressing iterated beliefs (such as $\mathbf{BB}\alpha$ and $\mathbf{B}\neg\mathbf{B}\alpha$) or beliefs concerning results of (potential) doxastic action (such as $\mathbf{B}[{*}\alpha]\beta$ or $\mathbf{B}\neg[-\alpha]\beta$). We also allow revisions, expansions and contractions with any well-formed formulas.

Thus, for example, expressions such as $[*\mathbf{B}\alpha]\beta$ or $[-[+\alpha]\beta]\gamma$ will be well-formed.

The formal language

Let us now describe our object language L. We define the sets **Term** and **Form** of *terms* and *formulas* to be the smallest sets satisfying the following conditions:

(i) for any $n < \omega$, the propositional letter P_n belongs to **Form**.
(ii) $\bot \in$ **Form**.
(iii) If $\alpha, \beta \in$ **Form**, then $(\alpha \rightarrow \beta) \in$ **Form**.
(iv) If $\alpha \in$ **Form**, then $\mathbf{B}\alpha \in$ **Form**.
(v) If $\alpha \in$ **Form**, then $+\alpha, -\alpha, *\alpha \in$ **Term**.
(vi) If $\tau \in$ **Term** and $\alpha \in$ **Form**, then $[\tau]\alpha \in$ **Form**.

The Boolean connectives $\neg\alpha$, $(\alpha \wedge \beta)$, etc. are defined from $\bot$ and $\rightarrow$ in the usual way. We define: $<\tau>\alpha$ as $\neg[\tau]\neg\alpha$ and $\mathbf{b}\alpha$ as $\neg\mathbf{B}\neg\alpha$.

Semantics for DDL unlimited

Here is a sketch of a *semantics* appropriate for a dynamic doxastic logic (DDL) for fully introspective agents.

In general terms, a semantic model for such a logic should contain the following components:

(1) A set U = {x, y, z, ...} of total *states*. A *total state* x involves both a *doxastic state* of the agent d(x) and a *possible world* w(x). By a possible world, we here mean a state of the world that is external to the agent.
(2) Hence, we have functions w and d that to each state x in U assign the world and the doxastic state, respectively, that obtain in x. We let W = {w(x): $x \in$ U} and D = {d(x): $x \in$ U} be the sets of all possible worlds and all doxastic states, respectively.
(3) A *valuation function* V that assigns subsets of U to atomic formulas. For each atomic formula p, V(p) is the *proposition* expressed in the model.

For each x, d(x) specifies the agent's doxastic state in x (the totality of the agent's beliefs together with his policies-for-belief change), while w(x) specifies the "external world" of x. It is therefore reasonable to assume that d(x) and w(x) together fully characterize the total state x. Thus, we may assume that:

(i) For all x, y in U, if d(x) = d(y) and w(x) = w(y), then x = y.

This restriction might suggest a simplification: why not *identify* elements of U with ordered pairs of doxastic states and worlds? Here, however, we prefer not to make this identification, for the following reason: In more specialized versions of the semantics for DDL, doxastic states will be interpreted as set-theoretic constructs built up from states in U. Given such a construction, reduction of states to doxastic-states-cum-worlds becomes impossible, unless we are prepared to allow sets that are not well-founded.

In addition to the elements that have already been mentioned, a model should contain special components that correspond to the different doxastic operators: either accessibility relations between states, or — what amounts to the same — functions from states to sets of states. These components should be made dependent on the d-function. Thus, if we let b be the function that to each state x assigns the set of states that are compatible with what is believed in x (i.e., if b is to be the component of the model that corresponds to **B**), then we should impose the following restriction on b:

(ii) If $d(x) = d(y)$, then $b(x) = b(y)$.

For doxastic dynamic operators, the dependence relationships are somewhat more complex. Let R^τ be the accessibility relation on states that corresponds to the operator $[\tau]$. Since we take τ to be a *purely* doxastic action, that only modifies the doxastic state but does not "touch" the (external) world, we must assume that:

(iii) If $R^\tau(x, y)$, then $w(x) = w(y)$.

(This means that we do not consider such impurely doxastic operators as the operator of *updating*. For updating, we would need to consider actions that transform a state into another state in which the world has changed and the doxastic state has registered that world-change and thus has itself been changed accordingly.)

On the other hand, since the potential results of action τ as far as the doxastic state is concerned are supposed to be determined by the original doxastic state, we need to assume that the following holds:

(iv) If $R^\tau(x, y)$ and $d(x) = d(x')$, then for some y', $d(y') = d(y)$ and $R^\tau(x', y')$.

Note that by assumption (ii), $w(y') = w(x')$, and by the present assumption $d(y') = d(y)$. This means, given assumption (i), that there is exactly one y' that corresponds to y in this way. If x is R^τ-related to a state y and $d(x) = d(x')$, then x' is

R^{τ}-related to the state y' characterized by the doxastic state d(y) and the external world w(x').

The above assumptions seem to be sufficient as far as a general semantics for DDL is concerned. Thus, we define a model $\mathfrak{M}$ to be a structure <U, w, d, b, R^{+}, R^{-}, R^{*}, V>, where w, d, b, and V are as described above. R^{+} is a function which for every formula α yields an accessibility relation $R^{+\alpha} \subseteq U \times U$, and similarly for R^{-} and R^{*}. The notion of a formula α being true in a model $\mathfrak{M}$ at a state x (in symbols, $\mathfrak{M}, x \vDash \alpha$) is defined recursively as follows:

(i) $\mathfrak{M}, x \vDash p$ iff $x \in V(p)$.

(ii) It is not the case that $\mathfrak{M}, x \vDash \bot$.

(iii) $\mathfrak{M}, x \vDash (\alpha \rightarrow \beta)$ iff it is either the case that not: $\mathfrak{M}, x \vDash \alpha$ or it is the case that $\mathfrak{M}, x \vDash \beta$.

(iv) $\mathfrak{M}, x \vDash \mathbf{B}\alpha$ iff $\forall y$(if $y \in b(x) \rightarrow \mathfrak{M}, y \vDash \alpha$)

(v) If τ is a term, then
$\mathfrak{M}, x \vDash [\tau]\alpha$ iff for all y such that $R^{\tau}(x, y)$, $\mathfrak{M}, y \vDash \alpha$.

In the following we usually suppress the reference to $\mathfrak{M}$ and write $x \vDash \alpha$ instead of $\mathfrak{M}, x \vDash \alpha$.

Let X be a class of models. We then define the notions of X-consequence and X-validity in the expected way. α is an *X-consequence* of a set of formulas Γ (in symbols, $\Gamma \vDash_X \alpha$) if and only if, for any model $\mathfrak{M}$ in X and any state x in $\mathfrak{M}$, if $\mathfrak{M}, x \vDash \beta$ for every β in Γ, then $\mathfrak{M}, x \vDash \alpha$. α is *X-vali*d (in symbols, $\vDash_X \alpha$) if and only if, for every model $\mathfrak{M}$ in X and every state x in $\mathfrak{M}$, $\mathfrak{M}, x \vDash \alpha$.

A *proposition* is represented as a set of states.[7] A *worldly* proposition is any proposition P such that, if $x \in P$ and $w(x) = w(y)$, then $y \in P$. If we wish, we can assume that for every atomic formula p, V(p) is a worldly proposition. This will validate such formulae as:

$$p \rightarrow [A]p,$$

where p is an atomic formula and [A] is a purely doxastic dynamic operator. As we remember, doxastic action does not touch the world; it only modifies the doxastic component of a given state.

Worldly propositions may be contrasted with *doxastic* propositions P for which the following holds: $x \in P$ and $d(x) = d(y)$, then $y \in P$. Clearly, we also have a third category: "mixed" propositions that are neither worldly nor doxastic.

---

[7] Should we have a family of propositions as an additional element of the models or allow all sets of states as propositions? For the time being, we follow the latter alternative.

For every formula α, $\|\alpha\| = \{x \in U: x \models \alpha\}$ is the proposition expressed by α, i.e., the set of states at which α is true. The truth-clause for **B** can be written as:

$x \models \mathbf{B}\alpha$ iff $b(x) \subseteq \|\alpha\|$.

In a *Segerberg-style* semantics for DDL, we let doxastic states be *hypertheories*, i.e. families of subsets of U that satisfy appropriate conditions (cf. Segerberg). We then identify b(x) — the set of states that are compatible with what is believed in x — with $\cap d(x)$, the intersection of all the subsets of U that belong to d(x).

$x \models \mathbf{B}\alpha$ iff $\cap d(x) \subseteq \|\alpha\|$.

The LRS-version of DDL is based on Lindström and Rabinowicz' *relational* approach to belief change, which has been augmented by Segerberg with a specific recipe for belief change *iteration.* This recipe allows for an explicit definition of the accessibility relations corresponding to the doxastic dynamic operators. In his approach, however, there is no room for iterations of the **B**-operator nor is it possible to revise or contract beliefs with propositions other than worldly ones. These restrictions are removed in the present paper.

In the LRS-variant of DDL, we can define the accessibility operations for the operators of expansion, contraction and revision ($[+\alpha]$, $[-\alpha]$ and $[*\alpha]$, with α being an arbitrary formula) as follows:

$R^{+\alpha}(x, y)$ iff
(i) $w(x) = w(y)$, and (ii) $d(y) = d(x) + \|\alpha\| = d(x) \cup \{X \cap \|\alpha\|: X \in d(x)\}$.

$R^{-\alpha}(x, y)$ iff
(i) $w(x) = w(y)$, and
(ii) $d(y) = d(x) \mid Z = \{X \in d(x): Z \subseteq X\}$, for some element Z that is minimal in $\{X \in d(x): X - \|\alpha\| \neq \varnothing\}$.

$R^{*\alpha}(x, y)$ iff
(i) $w(x) = w(y)$, and
(ii) $d(y) = (d(x) \mid Z) + \|\alpha\|$ for some element Z that is minimal in $\{X \in d(x): X \cap \|\alpha\| \neq \varnothing\}$.

Note that $R^{*\alpha}$, when defined in this way, is the relative product of $R^{-(\neg\alpha)}$ and $R^{+\alpha}$. Consequently, $[*\alpha]$ is explicitly definable as $[-(\neg\alpha)][+\alpha]$. However, as we shall argue, this Levi-style definition of $*$ will have to be given up in view of the problem that is presented next.

Doxastic dynamics for introspective agents

Whether we choose to accept this particular LRS-modelling for DDL or prefer to work with the general model, we encounter the following difficulty:

Let us say that revision $*$ is *strongly paradoxical*, if for every state x and every formula α, the following formula is true in x:

(StrongParadox) $\quad \mathbf{b}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha \rightarrow [*\alpha]\mathbf{B}\bot$.

The opposite of strong paradoxicality just requires that there should be a state x and a formula α such that $\mathbf{b}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha \wedge \neg[*\alpha]\mathbf{B}\bot$ holds in x. This seems to be a very reasonable requirement on any belief revision operation.

Lemma 1. Suppose that $*$ satisfies *Preservation and Success*, while **B** satisfies *Positive Introspection*:

(P)	$\mathbf{b}\alpha \rightarrow (\mathbf{B}\beta \rightarrow [*\alpha]\mathbf{B}\beta)$	*(Preservation)*
(S)	$[*\alpha]\mathbf{B}\alpha$	*(Success)*
(PI)	$\mathbf{B}\alpha \rightarrow \mathbf{BB}\alpha$.	*(Positive Introspection)*

Then, if the operator $[*\alpha]$ satisfies closure under logical implication:

$$\text{if } \vDash \beta \rightarrow \gamma, \text{ then } \vDash [*\alpha]\beta \rightarrow [*\alpha]\gamma,$$

and both $[*\alpha]$ and **B** satisfy closure under conjunction, $*$ is strongly paradoxical.[8]

Proof: Suppose that $\mathbf{b}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha$ holds in x. Then, by (P),

(1) $\quad [*\alpha]\mathbf{B}\neg\mathbf{B}\alpha$

is true in x. But by Success it is also true in x that:

(2) $\quad [*\alpha]\ \mathbf{B}\alpha$.

If $[*\alpha]$ is closed under logical implication, (2) and (PI) imply that:

(3) $\quad [*\alpha]\ \mathbf{BB}\alpha$.

If, in addition, $[*\alpha]$ and **B** are closed under conjunction, (3) and (1) imply:

(4) $\quad [*\alpha]\mathbf{B}(\mathbf{B}\alpha \wedge \neg\mathbf{B}\alpha)$,

which in turn yields the desired result: $[*\alpha]\mathbf{B}\bot$. Q. E. D.

---

[8] This lemma is closely related to Fuhrmann's (1989) "paradox of serious possibility". In present terms , Fuhrmann proves $\neg\mathbf{B}\neg\alpha \wedge \neg\mathbf{B}\alpha \rightarrow [*\alpha]\mathbf{B}\bot$, but he relies on Negative Introspection in addition to the positive one. Cf. also Levi (1988).

Note that even in the absence of Positive Introspection, Preservation plus Success will yield unacceptable results. Say that $*$ is *paradoxical* if and only if, for every x and α, the following formula is true in x:

(Paradox) $\quad \mathbf{b}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha \rightarrow [*\alpha](\mathbf{B}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha)$.

This means, in particular, that if the agent holds no opinion as regards α and correctly believes that he does not believe α, then, upon revision with α, he will believe that α is true and, at the same time, believe that he does not believe α. But then he has at least one false belief, namely that he does not believe α. The requirement that $*$ should not be paradoxical in this sense seems eminently plausible.

Lemma 2. Suppose that the $*$ satisfies *Preservation and Success*. Then if $[*\alpha]$ is closed under conjunction, $*$ is paradoxical.

Proof: Suppose that $\mathbf{b}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha$ holds in x. Then, by (P) and (S), respectively,

(1) $\quad [*\alpha]\mathbf{B}\neg\mathbf{B}\alpha$

and

(2) $\quad [*\alpha]\ \mathbf{B}\alpha$,

are true in x. But then, if $[*\alpha]$ is closed under conjunction,

(3) $\quad [*\alpha](\mathbf{B}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha)$

is true in x. Q. E. D.

A natural conclusion is that we should give up *Preservation* for $*$: If I originally neither believe nor disbelieve α and am aware of this fact and then learn that α is true, some of my original beliefs must be given up. In particular, I have to give up my original belief that I do not believe α.

Positive and Negative Introspection ($\mathbf{B}\alpha \rightarrow \mathbf{BB}\alpha$ and $\neg\mathbf{B}\alpha \rightarrow \mathbf{B}\neg\mathbf{B}\alpha$, respectively) should also be given up, but for a different reason, having to do with *contraction* rather than with revision. Let us first consider why Negative Introspection is inappropriate as a general requirement. When we originally do not believe α and then contract with $\neg\mathbf{B}\alpha$ (i.e., stop believing that we do not believe α), then $\neg\mathbf{B}\alpha$ should still be true in the contracted state (contracting with $\neg\mathbf{B}\alpha$ should not make us believe α) but it won't be believed any longer: $\mathbf{B}\neg\mathbf{B}\alpha$ will be false. Thus, in this intermediary state, Negative Introspection will be violated.

That *Positive* introspection will also sometimes be violated is less obvious, but think of an agent who originally believes α and believes that he does believe α. Suppose he is invited to contract his beliefs with $\mathbf{B}\alpha$ (i.e., stop believing that he believes α). In the contracted state, it is no longer true that $\mathbf{BB}\alpha$, but we would like to allow that it is still true that $\mathbf{B}\alpha$. This is, however, impossible unless positive introspection is violated after the contraction. If we insisted on positive introspection being valid, we would have to stop believing α just because we stop believing that we believe α. This seems wrong.

While Positive and Negative Introspection should probably be given up, it seems that we instead might still insist on their converses: we might insist that an (ideal) agent's beliefs concerning his own beliefs are never mistaken:

$\mathbf{BB}\alpha \rightarrow \mathbf{B}\alpha$ *(Veridicality of Positive Introspection)*

$\neg\mathbf{B}\bot \rightarrow (\mathbf{B}\neg\mathbf{B}\alpha \rightarrow \neg\mathbf{B}\alpha)$. *(Veridicality of Negative Introspection)*

(The latter requirement is a slightly qualified converse of Negative Introspection: the qualification in the antecedent is added in order to allow states in which the agent holds inconsistent beliefs.) In our general semantic framework, these requirements are validated by the following conditions:

If $y \in b(x)$, there is some $z \in b(x)$ such that $y \in b(z)$.
If $b(x) \neq \varnothing$, there is some $y \in b(x)$ such that $b(y) \subseteq b(x)$.

Note that both conditions would follow from the following restriction on the model:

If $b(x) \neq \varnothing$, there is some $y \in b(x)$ such that $b(y) = b(x)$.

According to this condition, the agent is never mistaken about his beliefs.

For future reference, we may also mention an even stronger condition according to which an agent is never mistaken about his *doxastic state*. Thus, he is not only never mistaken about his beliefs but also about his policies for belief change. He might not be fully informed about his doxastic state (in particular, he might violate positive and negative introspection) but the beliefs he holds about it are never false:

FULL VERIDICALITY OF INTROSPECTION:
If $b(x) \neq \varnothing$, there is some $y \in b(x)$ such that $d(y) = d(x)$.

Let us now return to our problem with revision. If revision is not to be paradoxical, it should not be fully preservative: in particular, certain beliefs about one's own beliefs need to be given up when one receives new information. How to achieve this result? Here is a suggestion. Levi-style revision with α consists

in two steps: we first contract with $\neg\alpha$ and then expand with α. In some cases, the first step is vacuous, $\neg\alpha$ is not believed to begin with. Then revision reduces to expansion. These are precisely the cases for which Preservation is meant to hold: revision has been supposed to be preservative simply because expansion is cumulative: all the old beliefs are kept when we expand with a new belief. Our suggestion is to replace the expansion step in the process of revision with what might be called *cautious expansion:* before we expand with α, we should first make sure that we give up the belief that we do not believe α. Clearly, this belief should not survive our coming to believe that α. Thus, unlike standard expansion, cautious expansion is not fully cumulative: certain beliefs have to be given up when new beliefs are added. This suggests the following definition of the *cautious expansion* operator $[\oplus\alpha]$:

$$[\oplus\alpha]\beta =\text{df } [-(\neg\mathbf{B}\alpha)][+\alpha]\beta.$$

Thus, cautious expansion with α is itself a two-step process: we first contract with $\neg\mathbf{B}\alpha$ and only then expand with α.

We can then define revision with α in a new way — as contraction with $\neg\alpha$ followed by a cautious expansion with α:

$$[*\alpha]\beta =\text{df } [-(\neg\alpha)][\oplus\alpha]\beta.$$

How does this relate to the LRS-semantics? The definitions of the accessibility relations that correspond to contraction and (standard) expansion may be kept unchanged. But the accessibility relation that corresponds to revision will have to be modified. $R^{*\alpha}$ will now be interpreted as the relative product of $R^{-(\neg\alpha)}$ and $R^{\oplus\alpha}$, where $R^{\oplus\alpha}$ will itself be the relative product of $R^{-(\neg\mathbf{B}\alpha)}$ and $R^{+\alpha}$.

Our definition of cautious expansion would not be satisfactory if introspection weren't assumed to be veridical. To see that, suppose that in the original state in which he lacks belief that α, the agent is fully reflective, so that $\neg\mathbf{B}\alpha$, $\mathbf{B}\neg\mathbf{B}\alpha$, $\mathbf{BB}\neg\mathbf{B}\alpha$, etc., are all true. If he then contracts with $\neg\mathbf{B}\alpha$, as the first step in cautious expansion with α, then — given the veridicality of introspection — he will lose not just his belief in $\neg\mathbf{B}\alpha$ but also all his higher order beliefs: not just $\mathbf{B}\neg\mathbf{B}\alpha$, but also $\mathbf{BB}\neg\mathbf{B}\alpha$, etc., will all be false. Otherwise, if he kept one of these higher beliefs, some of his introspective beliefs would not be veridical. Proof: Suppose that n (n > 1) is the lowest number such that $\mathbf{B}^n\neg\mathbf{B}\alpha$ is still true after contraction. Then the agent has a false introspective belief that $\mathbf{B}^{n-1}\neg\mathbf{B}\alpha$. And he would hold on to that false belief after the second step of the cautious expansion. This would clearly be an unwanted result.

In fact, it seems reasonable to accept Full Veridicality of Introspection. Otherwise, when contracting with $\neg\mathbf{B}\alpha$, we might not get rid of some of the original

beliefs concerning outcomes of potential belief change — beliefs that are dependent on our belief in $\neg\mathbf{B}\alpha$ and that would become false when belief in $\neg\mathbf{B}\alpha$ is removed. As long as we demand Full Veridicality of Introspection, this possibility need not worry us.

6. The Ramsey test, Gärdenfors' impossibility theorem and DDL

In this final section we are going to represent Gärdenfors' impossibility theorem for the Ramsey test within the framework of non-introspective DDL (i.e., the original DDL as developed by Segerberg). For this purpose, we introduce two formal languages which we refer to as $L_>$ and $DL_>$, respectively. $L_>$ is a language of sentential conditional logic with formulas built up from propositional letters and $\bot$ by means of the material conditional $\rightarrow$ and a non-Boolean conditional connective $>$. Formulas of the form $(\alpha > \beta)$ are read as "If α, then β". The language $DL_>$ is a version of DDL defined by means of the following grammatical rules:

(i) formulas of $L_>$ are formulas of $DL_>$;
(ii) if α is a formula of $L_>$, then $\mathbf{B}\alpha$ and $\Box\alpha$ are formulas of $DL_>$;
(iii) If α is a formula of $L_>$, then $+\alpha$, $-\alpha$, $*\alpha$ are terms of $DL_>$;
(iv) If τ is a term of L and α is a formula of L, then $[\tau]\alpha$ is a formula of $DL_>$;
(v) If α, β are formulas of $DL_>$, then $(\alpha \rightarrow \beta)$ is a formula of $DL_>$.

A model $\mathfrak{M}$ for $DL_>$ is a structure $<U, w, d, b, R^+, R^-, R^*, \Rightarrow, V>$, where U, w, d, b, R^+, R^-, R^* and V are defined as before, and $\Rightarrow$: $Pow(U) \times Pow(U) \rightarrow Pow(U)$. The semantic clauses for atomic formulas, the Boolean connectives, the doxastic operator **B**, and the dynamic operators are the same as before. However, there are two new semantic clauses corresponding to the symbols $>$ and $\Box$:

$\mathfrak{M}, x \models (\alpha > \beta)$ iff $x \in (\|\alpha\| \Rightarrow \|\beta\|)$.
$\mathfrak{M}, x \models \Box\alpha$ iff for all $y \in U$, $\mathfrak{M}, y \models \alpha$.

By the Ramsey test we understand the condition:

(RT) $\neg\mathbf{B}\bot \rightarrow (\mathbf{B}(\alpha > \beta) \leftrightarrow [*\alpha]\mathbf{B}\beta)$.

The proviso that $\bot$ is not believed is meant to restrict the test to states in which the agents beliefs are consistent. The Ramsey test says that a (consistent) agent believes a conditional $\alpha > \beta$ just in case he would believe its consequent β, if he

were to revise his beliefs with the information that the antecedent α is true. We may say that $>$ is a *Ramsey conditional* if it satisfies the Ramsey test.

We say that a model $\mathfrak{M}$ is *nontrivial*, if it satisfies the following condition: There is a state $x \in U$ and formulas α, β, γ of $L_>$ such that:

$$\|\alpha \wedge \beta\| = \varnothing, \|\alpha \wedge \gamma\| = \varnothing, \|\beta \wedge \gamma\| = \varnothing \text{ and}$$
$$x \models \neg\mathbf{B}\neg\alpha, x \models \neg\mathbf{B}\neg\beta, x \models \neg\mathbf{B}\neg\gamma.$$

Nontriviality says that there is a state x and three disjoint propositions α, β and γ such that the agent's beliefs in the state x are consistent with each one of α, β, and γ.

By adapting Gärdenfors' impossibility result (1988, Chapter 7) to the present framework we get:

Theorem. There is no nontrivial model for $DL_>$ that validates every instance of the following conditions:

(RT)	$\neg\mathbf{B}\bot \rightarrow (\mathbf{B}(\alpha > \beta) \leftrightarrow [*\alpha]\mathbf{B}\beta)$	*(the Ramsey test)*
(P)	$\neg\mathbf{B}\neg\alpha \rightarrow (\mathbf{B}\beta \rightarrow [*\alpha]\mathbf{B}\beta)$	*(Preservation)*
(S)	$[*\alpha]\mathbf{B}\alpha$	*(Success)*
(E1)	$[+\alpha]\mathbf{B}\beta \leftrightarrow \mathbf{B}(\alpha \rightarrow \beta)$	
(E2)	$[+\alpha]\neg\mathbf{B}\beta \leftrightarrow \neg\mathbf{B}(\alpha \rightarrow \beta)$	
(C)	$[*\alpha]\mathbf{B}\bot \rightarrow \Box\neg\alpha$.	*(Consistency)*

Proof:[9] We first notice that the condition:

(1) $\neg\mathbf{B}\neg\alpha \rightarrow ([+\alpha]\mathbf{B}\beta \rightarrow [*\alpha]\mathbf{B}\beta)$

follows from (P), (S) and (E1). From this we get that for any α, β, γ,

(2) $[+\gamma]\neg\mathbf{B}\neg\alpha \rightarrow ([+\gamma][+\alpha]\mathbf{B}\beta \rightarrow [+\gamma][*\alpha]\mathbf{B}\beta)$.

Suppose now that there is a nontrivial model $\mathfrak{M}$ validating the conditions of the theorem. By nontriviality, there are formulas α, β, γ of $L_>$ and a state x such that α, β and γ are disjoint and:

$$x \models \neg\mathbf{B}\neg\alpha, x \models \neg\mathbf{B}\neg\beta, x \models \neg\mathbf{B}\neg\gamma.$$

Then, it must also the case that:

---

[9] This proof is an adaptation to the present framework of the proof of Theorem 7.14 in Appendix E of Gärdenfors (1988). In the proof, we assume that the belief operator **B** and the operators for expansion and revision satisfy closure under logical implication and conjunction (i.e., that they satisfy the closure conditions of the modal system **K**). This of course follows from the relational semantics given to these operators.

$x \models \neg\mathbf{B}\neg(\alpha \vee \beta)$ and $x \models \neg\mathbf{B}\neg(\alpha \vee \gamma)$.

We are going to consider the complex operation:

$[+(\alpha \vee \beta)][*(\beta \vee \gamma)]$.

Claim 1. $x \models [+(\alpha \vee \beta)][*(\beta \vee \gamma)]\mathbf{B}\beta$.

Proof of claim 1: We first show that:

(3) $x \models [+(\alpha \vee \beta)]\neg\mathbf{B}\neg(\beta \vee \gamma)$.

By (E2), it suffices to prove:

$x \models \neg\mathbf{B}((\alpha \vee \beta) \rightarrow \neg(\beta \vee \gamma))$.

Suppose to the contrary that

$x \models \mathbf{B}((\alpha \vee \beta) \rightarrow \neg(\beta \vee \gamma))$.

But this implies that:

$x \models \mathbf{B}\neg\beta$,

contrary to the assumption. Hence, (3) follows by reductio.

(3) yields by the schema (2),

(4) if $x \models [+(\alpha \vee \beta)][+(\beta \vee \gamma)]\mathbf{B}\beta$,
then $x \models [+(\alpha \vee \beta)][*(\beta \vee \gamma)]\mathbf{B}\beta$.

Now,

(5) $x \models [+(\alpha \vee \beta)][+(\beta \vee \gamma)]\mathbf{B}\beta$

holds, just in case,

$x \models \mathbf{B}(([\alpha \vee \beta) \wedge (\beta \vee \gamma)] \rightarrow \beta)$.

But since α, β and γ are pairwise incompatible, it follows that $\|[\alpha \vee \beta) \wedge (\beta \vee \gamma)]\| = \|\beta\|$. Hence, we have that (5), which together with (4) yields:

$x \models [+(\alpha \vee \beta)][*(\beta \vee \gamma)]\mathbf{B}\beta$.

By a completely symmetrical argument, we get:

Claim 2: $x \models [+(\alpha \vee \gamma)][*(\beta \vee \gamma)]\mathbf{B}\gamma$.

Claims 1 and 2 yield, by the Ramsey test (notice that, the agent's belief state after having expanded with $\alpha \vee \beta$ and $\alpha \vee \gamma$, respectively, is consistent):

(6) $x \models [+(\alpha \vee \beta)]\mathbf{B}((\beta \vee \gamma) > \beta)$
(7) $x \models [+(\alpha \vee \gamma)]\mathbf{B}((\beta \vee \gamma) > \gamma)$.

But (6) and (7) yield:

(8) $x \models [+\alpha]\mathbf{B}((\beta \vee \gamma) > \beta)$
(9) $x \models [+\alpha]\mathbf{B}((\beta \vee \gamma) > \gamma)$.

Using the Ramsey test again, we get:

(10) $x \models [+\alpha][*(\beta \vee \gamma)]\mathbf{B}\beta$
(11) $x \models [+\alpha][*(\beta \vee \gamma)]\mathbf{B}\gamma$.

Hence,

(12) $x \models [+\alpha][*(\beta \vee \gamma)]\mathbf{B}(\beta \wedge \gamma)$.

But, since $\| \beta \wedge \gamma \| = \varnothing$, we get:

(13) $x \models [+\alpha][*(\beta \vee \gamma)]\mathbf{B}\bot$.

But this is contrary to the consistency condition (C). Thus, we have obtained a contradiction and proved the theorem. Q.E.D.

The straightforward conclusion seems to be that the only way to accommodate Ramsey conditionals within DDL is to give up the Preservation condition. In this respect, there is an analogy between higher-order beliefs and Ramsey conditionals. Admitting either requires that we give up Preservation. Higher-order beliefs and Ramsey conditionals are alike in that they should sometimes be given up when we add new information to our stock of beliefs — even when the new information is compatible with our old beliefs. It is not surprising that Ramsey conditionals behave like beliefs about beliefs in this respect. After all, what the Ramsey test says is that the agent should accept the conditional "If α, then β" just in case he is disposed to accept β, if he were to learn α. That is, the agent's acceptance of conditionals should *reflect* his conditional dispositions to believe. In the light of new information compatible with what the agent believes, it might very well be rational to relinquish some of these conditional dispositions. But then, according to the Ramsey test, the agent should also cease to believe the corresponding conditionals.

References

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

Fuhrmann, A., (1989) 'Reflective modalities and theory change', Synthese **81**, 115-134.

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

Gärdenfors, P. (1988) *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, Bradford Books, MIT Press.

Hintikka, J., (1962) *Knowledge and Belief*, Ithaca, N. Y.: Cornell University Press.

Hintikka, J., 1969, ‘Semantics for propositional attitudes’, in J. W. Davies *et al.* (ed.), *Philosophical Logic*, pp. 21-45, D. Reidel, Dordrecht. Reprinted in J. Hintikka, *Models for Modalities*, D. Reidel, 1969; and in L. Linsky (ed.), *Reference and Modality*, Oxford University Press, London, 1971.

Levi, I. (1988) ‘Iteration of conditionals and the Ramsey test’, Synthese **76**, 49-81.

Lindström, S., and Rabinowicz, W. (1989) ‘On probabilistic representation of non-probabilistic belief revision’, *Journal of Philosophical Logic* **18**, 69-101.

Lindström, S., and Rabinowicz, W. (1990) ‘Epistemic Entrenchment with Incomparabilities and Relational Belief Revision’, in Fuhrmann and Morreau (eds.), *The Logic of Theory Change*, Lecture Notes in Artificial Intelligence no. **465**, Springer Verlag, 93-126.

Lindström, S., and Rabinowicz, W. (1992) ‘Belief Revision, Epistemic Conditionals and the Ramsey Test, *Synthese* **91**, 195-237.

Rabinowicz, W. and Lindström, S., (1994) ‘How to Model Relational Belief Revision’, in D. Prawitz and D. Westerståhl (eds.), *Logic and Philosophy of Science in Uppsala*, Kluwer, 1994, 69-84.

Segerberg, K., (1995) ‘Belief revision from the point of view of doxastic logic’. *Bulletin of the IGPL*, vol. 3, 535-553.

Segerberg, K., (1996a) ‘Two traditions in the logic of belief: bringing them together’. *Uppsala Prints and Preprints in Philosophy*, 1996, number 9.

Segerberg, K., (1996b) ‘Three Recipes for revision’. *Uppsala Prints and Preprints in Philosophy, 1996*, number 11.