BELIEF CHANGE FOR INTROSPECTIVE AGENTS*

Sten Lindström and Wlodek Rabinowicz

ABSTRACT

The modern development in *doxastic* (and *epistemic*) *logic* started with Jaakko Hintikka's seminal book *Knowledge and Belief* (1962). In a doxastic logic of Hintikka-type, with a modal operator **B** standing for "the agent believes that", it is possible to represent and reason about the *static* aspects of an agent's beliefs about the world. Such logic studies various constraints that a rational agent or a set of rational agents should satisfy.

A Hintikka-type logic cannot, however, be used to reason about doxastic change, i.e., 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*.

The theories of belief change developed within the AGM-tradition are not *logics* in the strict 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 – ordinary mathematical English – to characterize the mathematical structures that are under study. Recently, however, various authors such as Johan van Benthem and Maarten de Rijke have suggested representing doxastic change within a formal

---

* In this paper we give an informal presentation of some ideas that we have discussed more formally in Lindström and Rabinowicz (1997) and (1999). We dedicate it to Peter Gärdenfors on his 50th birthday, together with our warmest congratulations and best wishes. Were it not for him, it could never have been written! We wish to thank John Cantwell, Sven Ove Hansson, Tor Sandqvist, and last but not least, Krister Segerberg, for inspiration and advice.

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). This framework may be seen as an extension of standard Hintikka-style doxastic logic, with dynamic operators, such as $[*\varphi]$ or $[-\varphi]$, representing various kinds of transformations of the agent's doxastic state: 'after revising with φ, it would be the case that …', 'after contracting with α, it would be the case that …', etc.. Thus, formulas such as $[*\varphi]\mathbf{B}\psi$ can be used to describe the agent's doxastic state after various transformations.

The kind of DDL that has been mainly studied so far – *basic DDL* – describes an agent who has opinions about the *external world* and an ability to change these opinions in the light of new information. Such an agent is *non-introspective* in the sense of lacking opinions about his own belief state (his own beliefs and his own dispositions for belief change). The agent's own belief state is not a part of the reality that he has beliefs about.

In our paper, we discuss various possibilities for developing a dynamic doxastic logic for *introspective agents*: agents who have the ability both to form higher-order beliefs and to reflect upon and change their minds about their own beliefs. The project of constructing such a logic – *full DDL* or *DDL unlimited* – is ridden with difficulties due to the fact that the agent's own doxastic state (his beliefs as well as his doxastic dispositions) now becomes a part of the reality he is trying to explore. When an introspective agent learns more about the world (and himself) then the reality he holds beliefs about undergoes a change. But then his introspective (higher-order) beliefs have to be adjusted accordingly.

In the paper, we consider various ways of solving this problem. In particular, we outline *a two-dimensional semantics* for belief change. When an introspective agent gets new information, his doxastic state changes. Thereby the total state, of which his belief state is a part, changes as well. What are then his beliefs about? Is it the original state or the new one? One would like to say that he has beliefs about both. In general, therefore, we have to distinguish between the state in which beliefs are held (the *point of evaluation*) and the state about which certain things are believed (the *point of reference*). Within this two-dimensional approach, one can easily characterize reasonable constraints on the agent's new beliefs about the original state. Essentially, they correspond to the AGM-axioms. It is more difficult to determine what the agent should believe about the new state, in particular which of the beliefs about the old state should also extend to the new one and which should be appropriately modified. We make some suggestions as to what such transfer principles might look like, but there is more work to be done in this direction.

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 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.

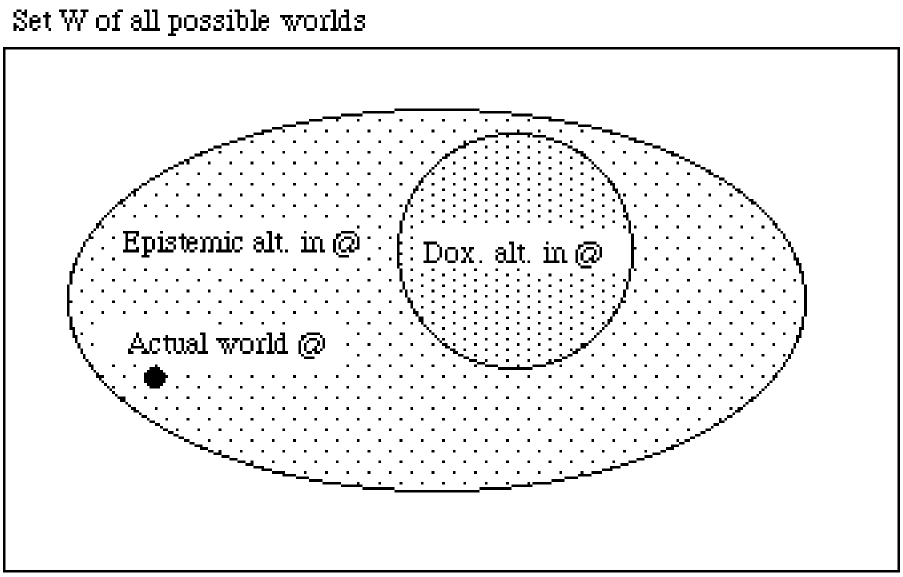
Similarly, the concept of belief appeals to the idea of a set of "doxastically" possible worlds (the agent's *doxastic alternatives*). The corresponding principle is:

> 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:

$\mathbf{K}\alpha$	for	"the agent knows that α".
$\mathbf{B}\alpha$	for	"the agent believes that α".

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) $\models \mathbf{K}(\alpha \to \beta) \to (\mathbf{K}\alpha \to \mathbf{K}\beta)$

(2) $\models \mathbf{B}(\alpha \to \beta) \to (\mathbf{B}\alpha \to \mathbf{B}\beta)$

(3) $\models \mathbf{K}\alpha \to \alpha$ *(Veridicality of Knowledge)*

(4) $\models \mathbf{K}\alpha \to \mathbf{B}\alpha$

(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. 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 doxastic alternativeness relation, one can ensure that some or all of the following principles are also satisfied:

(Cons) $\models \neg\mathbf{B}\bot$ *(Consistency)*

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

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

(PI) $\models$ $\mathbf{B}\alpha \rightarrow \mathbf{BB}\alpha$ *(Positive Introspection)*

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

(Cons), for example, says that a logical contradiction (symbolised by $\bot$) is never believed. (NI) 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.

2. AGM-type theories of belief change

In a doxastic logic of Hintikka-type, with a modal operator **B** standing for "the agent believes that", it is possible to represent and reason about the *static* aspects of an agent's beliefs about the world. Such a logic studies various constraints that a rational agent or a set of rational agents should satisfy. A Hintikka-type logic cannot, however, be used to reason about doxastic change, i.e., 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*.[1] 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 $G+\alpha$ denotes the set of beliefs that results from *expanding* G with α. To expand is dangerous, since $G+\alpha$ might very well be logically inconsistent; and inconsistency is something that we should try to avoid in our beliefs.

[1] Cf. Alchourrón, Gärdenfors, Makinson (1985) and Gärdenfors (1988).

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 $G–\alpha$ 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 classical consequence operation Cn is assumed and the operation of expansion + is defined by

$$G+\alpha = Cn(G \cup \{\alpha\}).$$

By contrast, the operations of contraction and revision are characterized only axiomatically. Thus, the operation of revision is assumed to satisfy the axioms:

(R1)	$Cn(G) = G$	*(Logical Closure)*
(R2)	$\alpha \in G*\alpha$	*(Success)*
(R3)	$G*\alpha \subseteq G+\alpha$	*(Inclusion)*
(R4)	if $\neg\alpha \notin G$, then $G \subseteq G*\alpha$	*(Preservation)*
(R5)	if $\bot \notin Cn(\{\alpha\})$, then $\bot \notin G*\alpha$	*(Consistency)*
(R6)	if $Cn(\{\alpha\}) = Cn(\{\beta\})$, then $G*\alpha = G*\beta$	*(Congruence)*
(R7)	$G*(\alpha \wedge \beta) \subseteq (G * \alpha)+\beta$	
(R8)	if $\neg \beta \notin G*\alpha$, then $(G * \alpha) \subseteq G*(\alpha \wedge \beta)$.	

The first four axioms imply:

if $\neg\alpha \notin G$, then $G*\alpha = G+\alpha$, *(Expansion)* i.e., if the new information α is consistent with G, then $G*\alpha$ is simply the expansion of G with α. Consistency says that if α is consistent, then $G*\alpha$ is also consistent. According to Congruence, if α and β are logically equivalent, then revising G with α yields the same result as revising G with β. In view of (R1) and (R2), the last two axioms yield:

$\neg \beta \notin G*\alpha$, then $G*(\alpha \wedge \beta) = (G*\alpha)+\beta$ *(Revision by Conjunction)*

i.e., if β is consistent with $G*\alpha$, then revising G with $(\alpha \wedge \beta)$ yields the same result as first revising G with α and then expanding the result with β.

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 $\neg\alpha$.

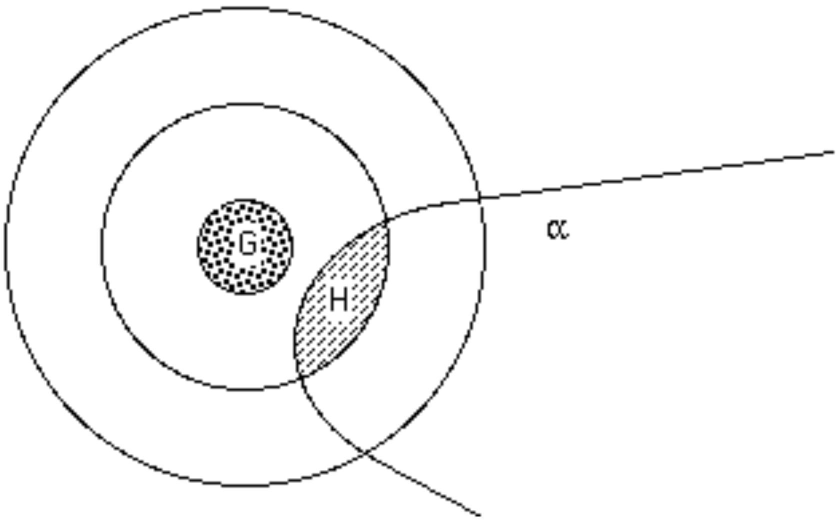
Grove's (1988) possible worlds modelling for AGM

In his (1988) paper, Grove presents two closely related possible worlds modellings 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. 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.

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. 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 propositions as sets of possible worlds, and also representing theories as such sets (rather than as sets of propositions), 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., linearly 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).



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 and Rabinowicz: relational belief revision

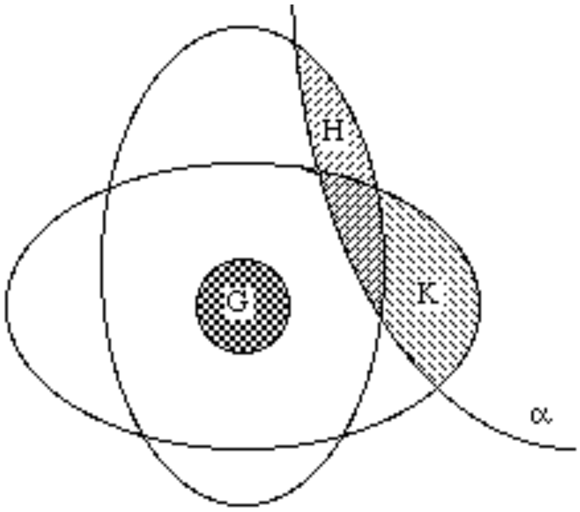
In a series of papers, Lindström and Rabinowicz have proposed a generalization *LR* of the AGM approach according to which belief revision was treated as a *relation* $GR_{\alpha}H$ between theories (belief sets) rather than as a function on theories.[2] The idea was to allow for there being several equally reasonable revisions of a theory with a given proposition. Thus, $GR_{\alpha}H$ means that H is one of those reasonable revisions 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, in LR the family of fall-backs around a given theory will no longer be nested. It will no longer be a family of spheres but rather a family of “ellipses”. This change opens up for the possibility of 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.

[2] Cf. Lindström and Rabinowicz (1989), (1990), (1992) and Rabinowicz and Lindström (1994).

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).[3] 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 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,

---

[3] Cf. Segerberg (1995), (1996), (1997), (1999a) and (1999b).

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

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

In DDL, we think of the agent's belief state as being represented by what Segerberg calls a *hypertheory*, or in the terminology of Lindström and Rabinowicz a system of *fallbacks*. A hypertheory encodes the agent's current beliefs as well as his dispositions for belief change.

Basic DDL

The kind of DDL that has been mainly studied so far – *basic DDL* – describes an agent who has opinions about the *external world* and an ability to change these opinions in the light of new information. Such an agent is *non-introspective* in the sense of lacking opinions about his own belief state (his own beliefs and his own dispositions for belief change). In the semantics, there is a space U of points representing different states of the external world, where the agent's beliefs and doxastic dispositions are not part of the world that he has beliefs about. Consequently, the doxastic actions of the agent do not affect the world.

The object language for basic DDL can be described as follows. We define the sets **Term**, **BForm** and **Form** of *terms*, *Boolean formulas* and *formulas* to be the smallest sets satisfying the following conditions:

(i) for any $n < \omega$, the propositional letter P_n belongs to **BForm**

(ii) $\bot \in$ **BForm**

(iii) if $\alpha, \beta \in$ **BForm**, then $(\alpha \rightarrow \beta) \in$ **BForm**

(iv) if $\alpha, \beta \in$ **Form**, then $(\alpha \rightarrow \beta) \in$ **Form**

(v) if $\alpha \in$ **BForm**, then $\alpha \in$ **Form**

(vi) if $\alpha \in$ **BForm**, then **B**$\alpha \in$ **Form**.

(vii) if $\alpha \in$ **BForm**, then $+\alpha, -\alpha, *\alpha \in$ **Term**.

(viii) 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.

As is easily seen, basic DDL is severely limited in its expressive power. To begin with, the belief operator **B** only operates on Boolean formulas. Thus introspection is disallowed, i.e., formulas such as **B**$\neg$**B**α or **B**$[*\alpha]\beta$ are not well-formed. Secondly, the formula α that we contract, revise, or expand with, must always be Boolean. Thus, formulas such as $[*\neg$**B**$\alpha]\beta$ are not well-formed either. The reason for these limitations is obvious. Since the agent only holds beliefs about the world that his doxastic state is not a part of, he has no "higher order" beliefs. And since he only receives information that concerns the external world, he cannot revise his beliefs with propositions about his own doxastic state.

DDL Unlimited (Full DDL)

The language of unlimited (or full) DDL is obtained from that of basic DDL by removing the syntactic limitations just mentioned.

That is in unlimited DDL, we let the belief operator **B** take any formula α as argument. Moreover, for any formula α, $[+\alpha]$, $[-\alpha]$, $[*\alpha]$ are allowed as doxastic operators. Thus the sets of terms and formulas of full DDL can be defined in the following simple way:

(i) the propositional letters 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**.

Once we have provided the language of full DDL with a suitable semantics, we should be able to describe agents who

(i) have beliefs about their own doxastic states as well as about the external world.

(ii) can change their beliefs in the light of new information about their own belief states.

We say that such agents are *fully introspective*.

4. Paradoxes of introspective belief change

The project of constructing a dynamic doxastic logic for a fully introspective agent is faced with difficulties due to the fact that the agent's own doxastic state (his beliefs as well as his doxastic dispositions) now becomes a part of the reality that he is trying to explore.

> *When an introspective agent learns more about the world (and himself) then the reality that he holds beliefs about undergoes change. But then his introspective (higher-order) beliefs have to be adjusted accordingly.*

We illustrate the problems that arise with two stories.

Story 1.[4]

We consider an agent who can only receive worldly propositions as his doxastic inputs, but who is able to hold higher-order beliefs about his own current beliefs. We also assume that the following principles are satisfied:

[4] The paradox presented in Story 1 is closely related to Fuhrmann's (1989) "paradox of serious possibility". Cf. also Levi (1988).

(N) the operators **B** and [$*\alpha$], where α is Boolean, are normal modal operators

(*Normality*)

(S) $[*\alpha]\mathbf{B}\alpha$ (*Success*)

For any worldly proposition α, there is a way for the agent of revising his beliefs by α:

(PR) $\langle *\alpha \rangle T$, where T is any tautology (*Possibility of Revision*)

A consistent agent cannot become inconsistent simply by learning a true fact about the world:

(C) $\neg\mathbf{B}\bot \rightarrow (\alpha \rightarrow [*\alpha]\neg\mathbf{B}\bot)$

Then, we have a principle saying that no agent can believe both α and that he does not believe α, unless, of course, his beliefs are inconsistent:

(M) $\neg\mathbf{B}\bot \rightarrow \neg\mathbf{B}(\alpha \wedge \neg\mathbf{B}\alpha)$ (*Moore´s Principle*)

The Principle (M) is logically equivalent (relative to the base logic **K**) to:

$$\neg\mathbf{B}\bot \rightarrow (\mathbf{B}\alpha \rightarrow \neg B\neg\mathbf{B}\alpha),$$

which says that if a consistent agent believes α, then it is consistent with his beliefs that he believes α. This in turn is nothing but the principle that we have earlier called *Veridicality of Negative Introspection*:

$$\neg\mathbf{B}\bot \rightarrow (\mathbf{B}\neg\mathbf{B}\alpha \rightarrow \neg\mathbf{B}\alpha).$$

The next principle, finally, is one of the axioms of the AGM approach to belief revision:

(P) $\neg\mathbf{B}\neg\alpha \rightarrow (\mathbf{B}\beta \rightarrow [*\alpha]\mathbf{B}\beta)$ (*Preservation)*

It says that if the new information α is compatible with the agent's beliefs, then the agent retains all his old beliefs when revising with α.

We are now going to show that these principles cannot all be maintained, i.e., together they lead to an absurdity.

Suppose that our agent does not have an opinion one way or another whether:

α: It is raining in Uppsala

That is,

(1) $\neg\mathbf{B}\alpha$ and $\neg\mathbf{B}\neg\alpha$.

The agent also believes (correctly) that he does not believe that it is raining in Uppsala:

(2) $\mathbf{B}\neg\mathbf{B}\alpha$

Suppose also that as a matter of fact it is raining in Uppsala.

Consider now what happens when our agent learns that α is true. Then, he revises his beliefs by α after which he comes to believe α (by Success):

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

From the fact that (1) $\neg\mathbf{B}\neg\alpha$ and (2) $\mathbf{B}\neg\mathbf{B}\alpha$ it follows by Preservation that:

(4) $[*\alpha]\mathbf{B}\neg\mathbf{B}\alpha$.

But then (since **B** and $[*\alpha]$ are normal modal operators)

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

which in turn yields:

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

i.e., after learning α, the agent believes that: α and that he does not believe α. But by Moore´s Principle and the normality of $[*\alpha]$ we have that:

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

Thus,

(8) $[*\alpha]\mathbf{B}\bot$.

That is, the agent becomes inconsistent simply by learning the true proposition α. But since the agent was not inconnsistent to start with, we get by (C) that:

(9) $[*\alpha]\neg\mathbf{B}\bot$.

(8) and (9) yield $[*\alpha]\bot$, which is contrary to the Possibility of Revision-principle.

The principle to blame clearly seems to be Preservation. From Normality, Success and Preservation alone we get:

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

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 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.

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 if I 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. But as we shall see from our next story, giving up preservation does not solve all our problems.

Story 2.

We are now considering an agent who has higher order beliefs about his own current beliefs but is also able to revise his beliefs with doxastic propositions.

Consider the following story: as in our previous example, α is a true proposition which the agent has no opinion about. In particular, then, it is true that

(1) $\alpha \wedge \neg\mathbf{B}\alpha$.

The agent is now informed that (1) holds; he has received true information. Since (1) is true, it is clearly a consistent proposition. We would therefore expect that revision with (1) will not lead the agent to an inconsistent belief state (Principle (C)). In particular, then, it should be the case that

(2) $[*(\alpha \wedge \neg\mathbf{B}\alpha)]\neg\mathbf{B}\bot$.

But we also know that revision is supposed to satisfy Success. Thus, upon the revision with (1), the agent must believe that (1) holds:

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

But Normality together with Moore´s Principle yield:

(4) $[*(\alpha \wedge \neg\mathbf{B}\alpha)]\neg\mathbf{B}\bot \rightarrow [*(\alpha \wedge \neg\mathbf{B}\alpha)]\neg\mathbf{B}(\alpha \wedge \neg\mathbf{B}\alpha)$

From (2) and (4) we get by modus ponens:

(5) $[*(\alpha \wedge \neg\mathbf{B}\alpha)]\neg\mathbf{B}(\alpha \wedge \neg\mathbf{B}\alpha)$.

By Normality, (3) and (5) yield:

(6) $[*(\alpha \wedge \neg\mathbf{B}\alpha)]\bot$.

Contrary to what we should expect, revision with a true proposition such as (1) turns out to be impossible!

Lemma. Suppose that **B** and $[*\alpha]$ are normal modal operators in unlimited DDL that satisfy the following principles:

(S) $[*\alpha]\mathbf{B}\alpha$

(M) $\neg\mathbf{B}\bot \rightarrow \neg\mathbf{B}(\alpha \wedge \neg\mathbf{B}\alpha)$

(PR) $\neg[*\alpha]\bot$

(C) $\neg\mathbf{B}\bot \rightarrow (\alpha \rightarrow [*\alpha]\neg\mathbf{B}\bot)$

then the agent believes every true proposition, i.e.,

$$\alpha \rightarrow \mathbf{B}\alpha,$$

which in turn implies that the agent is either inconsistent or completely accurate in his beliefs, i.e.,

$$\neg\mathbf{B}\bot \rightarrow (\mathbf{B}\alpha \leftrightarrow \alpha).$$

Proof: Suppose the opposite, i.e., that there is a proposition α such that

$$\alpha \wedge \neg\mathbf{B}\alpha$$

and go through the above argument. Q.E.D.

5. Toward a solution: A two-dimensional Semantics for Full DDL

We are now going to sketch a semantics for unlimited (full) DDL which is intended to resolve the above paradoxes.

When an introspective agent gets new information, his doxastic state undergoes a change. Thereby the total state changes as well. What are then his beliefs about? The original state or the new one? One would like to say that he has beliefs about the old state as well as about the new one. In general, therefore, we have to distinguish between the state in which beliefs are held (the *point of evaluation*) and the state about which certain things are believed (the *point of reference*).

We are considering a semantics where each model **M** contains a set U of (total) *states*. Each state x has two components w(x), the *world component*, and d (x), the *doxastic component*. We write x = (w(x), d(x)).

The doxastic component of the different states jointly determine two kinds accessibility relations between states:[5]

(1) The accessibility relation that represents the agent´s beliefs: b(x, y, z) iff *z is compatible with what the agent believes in x about y*.

(2) the agent´s dispositions to revise his beliefs is represented by a relation R(∗α):

R(∗α)(x, y, z) iff *z is a state which the agent may reach from x by revising his beliefs about y with* α.

Since revision is a purely *doxastic action*, i.e., an action that does not change the external world, the world-component is not affected from a move from x to z by means of R(∗α), i.e., if R(∗α)(x, y, z), then w(x) = w(z).

The idea of having beliefs in one state about another gives rise to a *two-dimensional semantics*:

---

[5] To be more precise, we assume that the doxastic component d(x) of a state x is a function which to every state y assigns a doxastic state d(x)(y) that specifies the agent's views *in* the evaluation point x *about* the reference point y. We may speak of d(x)(y) as the agent's *doxastic state in x about y*. In a Segerberg-style semantics, we can identify each such d(x)(y) with a hypertheory. The hypertheory d(x)(y) specifies both which worlds are compatible with the agent's beliefs in x about y and the agent's dispositions to change his belief state x when he recevies new information about the state y. (For more formal details, see Lindström & Rabinowicz (1999)).

A formula α is, in general, *true at a point* x ("the point of evaluation") *with respect to a point* y ("the point of reference"). We write this as:

$$x, y \models \alpha.$$

For each α and each $y \in U$,

$$||\alpha||_y = \{x \in U: x, y \models \alpha\}$$

is the *proposition* expressed by α with respect to the point of reference y.

A proposition $P \subseteq U$ is *worldly* if it is closed under world-equivalence: whenever $x \in P$ and $w(x) = w(y)$, then $y \in P$.

Relative to a model **M**, we have the following semantic clauses:

(i) if α is atomic, then α is assigned a worldly proposition $||\alpha||$ as its semantic value; and for all $x, y \in U$,

$x, y \models \alpha$ iff $x \in ||\alpha||$.

(ii) It is not the case that $x, y \models \bot$.

(iii) $x, y \models (\alpha \to \beta)$ iff it is either the case that not: $x, y \models \alpha$ or it is the case that $x, y \models \beta$.

(iv) $x, y \models \mathbf{B}\alpha$ iff

for all z such that $b(x, y, z)$, $z, y \models \alpha$.

(v) $x, y \models [*\alpha]\beta$ iff

for all z such that $R(*\alpha)(x, y, z)$, $z, y \models \beta$.

We also extend the language with a new operator † that takes the current point of evaluation and makes it the point of reference:

(vi) $x, y \models \dagger\alpha$ iff $x, x \models \alpha$.

We introduce † in order to be able to distinguish between an agent's *posterior beliefs about his original state* (the one he is in before performing a doxastic action):

(1) $x, x \models [*\alpha]\mathbf{B}\beta$

and his *posterior beliefs about the posterior state* (the one he is in after the action):

(2) $x, x \models [*\alpha]\dagger\mathbf{B}\beta$.

We say that a formula α is *ordinary* if its truth or falsity does not depend on the point of reference, i.e., if for all x, y, z:

$$x, y \models \alpha \text{ iff } x, z \models \alpha.$$

A formula is *special*, if it is not ordinary.

An ordinary formula α expresses one and the same proposition (written, $||\alpha||$) with reference to every point of reference. It is easily seen that:

(a) Boolean formulas are ordinary.

(b) For any formula α, $\dagger\alpha$ is ordinary.

If α is ordinary, then

$$x, y \models \dagger\alpha \text{ iff } x, y \models \alpha.$$

Thus, for ordinary α,

$$||\dagger\alpha|| = ||\alpha||.$$

The proposition

$$||\dagger\alpha|| = \{x \in U\text{: } x, x \models \alpha\},$$

we call the *diagonal proposition* corresponding to α.

We say that a formula α is *true at the point* x if and only if $x, x \models \alpha$. In other words, α is true at x if and only if the proposition $||\alpha||_X$ expressed by α with reference to x is true at the point x itself.

We say that a formula α is *valid* (or *weakly valid*) in the model **M** (in symbols, $\mathbf{M} \models \alpha$) if and only if α is true at every point in **M**.

Let us say that a pair <x, y> of points in U is *normal* if x = y. We have defined truth at a point x as truth relative to the normal pair <x, x>, and we have defined validity in a model **M** as truth relative to all *normal pairs* in **M**.

There is another notion of validity in a model: We say that α is *strongly valid* in the model **M** if and only if, for every pair of points <x, y> in **M**, $x, y \models \alpha$. Of course, if α is strongly valid in **M**, then α is valid in **M**. The converse does not hold in general. Consider, for example,

$$\alpha \leftrightarrow \dagger\alpha.$$

Every instance of this schema is weakly valid in every model. However, if α is a special formula, then $\alpha \leftrightarrow \dagger\alpha$ is not strongly valid.

Notice, however, that for every model **M**, α is weakly valid in **M** if and only if $\dagger\alpha$ is strongly valid in **M**.

6. How the two-dimensional semantics handles the paradoxes

Consider a model **M** that strongly validates the following conditions:

(P) $\neg\mathbf{B}\neg\alpha \rightarrow (\mathbf{B}\beta \rightarrow [*\alpha]\mathbf{B}\beta)$ *(Preservation)*

(S) $[*\alpha]\mathbf{B}\alpha$ *(Success)*

From the strong validity of (P) and (S) in M, we can infer that also the following formula is strongly valid in M:

(1) $\neg\mathbf{B}\neg\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha \rightarrow [*\alpha](\mathbf{B}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha)$,

This formula was previously considered paradoxical. However, the *meaning* of the formula has changed from the old semantics to the new one: it is no longer paradoxical. To see this, one should compare (1) with:

(2) $\neg\mathbf{B}\neg\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha \rightarrow [*\alpha]\dagger(\mathbf{B}\alpha \wedge \mathbf{B}\neg\mathbf{B}\alpha)$,

which is indeed paradoxical. However, (2) is, of course, not even weakly valid.

While (1) is about what the agent, after having learned α, would believe about the state prior to the revision, (2) is about what he then would believe about the state obtaining after the revision. There is no reason to suppose that (2) would hold.

Consider now the case in which:

(3) $x, x \models \alpha \wedge \neg\mathbf{B}\alpha$,

The agent then learns (3) and revises his beliefs about the point x with this information. By Success:

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

but there is nothing paradoxical about (4), since the beliefs that are referred to in the formula following the revision operator are all about the prior state x and not about the one posterior to the revision. In contrast to (2), the following situation would be paradoxical:

(5) $x, x \models [*(\alpha \wedge \neg \mathbf{B}\alpha)]\dagger\mathbf{B}(\alpha \wedge \neg \mathbf{B}\alpha)$.

But the formula occurring in (5) is of course not (even weakly) valid.

But what about the following formula?

(6) $[*\dagger(\alpha \wedge \neg \mathbf{B}\alpha)]\mathbf{B}\dagger(\alpha \wedge \neg \mathbf{B}\alpha)$.

Isn't this formula valid, by Success? Yes, indeed it is. What it says, is that if one revises one's original beliefs with the diagonal proposition $\dagger(\alpha \wedge \neg \mathbf{B}\alpha)$, then, in the posterior state, one will have the belief about the prior state that $\dagger(\alpha \wedge \neg \mathbf{B}\alpha)$ was true then. In our example, however, this posterior belief about the prior state is in fact true. Hence, there is nothing paradoxical about it.

The formula (6) may be contrasted with:

(7) $[*\dagger(\alpha \wedge \neg \mathbf{B}\alpha)]\dagger\mathbf{B}(\alpha \wedge \neg \mathbf{B}\alpha)$,

which says that the agent after revision with $\dagger(\alpha \wedge \neg \mathbf{B}\alpha)$ would believe $\alpha \wedge \neg \mathbf{B}\alpha$ about his *posterior* state. This would indeed be paradoxical. But this formula is not even weakly valid, so no paradox is forthcoming.

In the two-dimensional semantics, we can impose various introspection principles, like (PI), (NI), (VPI), (VNI). These principles do not lead to trouble as long as we only assume them to be weakly, rather than strongly, valid.

Our conclusion is that the two-dimensional semantics avoids the original paradoxes, without – as far as we can see – creating new ones. This semantics has one serious drawback, however: it only determines the agent's posterior beliefs about the *prior* state:

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

What we would like to infer, however, are posterior beliefs about the posterior state:

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

Thus, we would like to have some *transfer principles*: at least for all the Boolean formulas β, we would like it to be valid that

(3) $[*\alpha](\mathbf{B}\beta \rightarrow \dagger\mathbf{B}\beta)$,

(4) $[*\alpha](\neg\mathbf{B}\beta \rightarrow \dagger\neg\mathbf{B}\beta)$.

Such principles would allow us to infer from one of $[*\alpha]\mathbf{B}\beta$, $<*\alpha>\mathbf{B}\beta$, $[*\alpha]\neg\mathbf{B}\beta$, $<*\alpha>\neg\mathbf{B}\beta$, to the corresponding statements about the posterior state: $[*\alpha]\dagger\mathbf{B}\beta$, $<*\alpha>\dagger\mathbf{B}\beta$, $[*\alpha]\dagger\neg\mathbf{B}\beta$, $<*\alpha>\dagger\neg\mathbf{B}\beta$, respectively.

In fact, even with respect to beliefs in doxastic propositions, there should be a large measure of agreement between posterior beliefs of this kind regarding the prior and the posterior state. If I initially believe that I believe the earth to be round, then after the revision with some information about, say, the weather in Sweden, I will keep my beliefs about what I believe to be the shape of the earth both with regard to my prior state and with regard to the posterior state. But transfer principles for posterior beliefs in doxastic propositions are much more difficult to formulate: many posterior beliefs about doxastic propositions are *not* transferable, as we have seen.

This shows that there is work that remains to be done. Still, we have at least made some first steps towards the development of a two-dimensional semantics for unlimited DDL. It is to be hoped that this project can be further developed.

7. Forward-looking revision

The main intuition behind the two-dimensional approach is that the new information that we revise with is a proposition about an antecedently specified state. We learn something new about the given state and revise our other beliefs about that state accordingly. Since we thereby move to a new state, we also acquire various beliefs about the state we have reached, but this latter change is, in a sense, secondary, as compared with the modification of our beliefs about the state we have originally considered. Thus, this view of revision is essentially *backward-looking*. Normally, the antecedently specified state will simply be the state in which we were originally situated. A belief change caused by perception or introspection seems to be a typical example of backward-looking revision.

A different conception of revision would see this doxastic action as a *forward-looking* task: for a given proposition α, the agent is asked to change his beliefs in such a way that he comes to believe α about the state that obtains *after* the revision. (An apt example might be a computer that is given a task to re-organize its "beliefs" in such a way as to accommodate α.)

This kind of forward-looking revision cannot be expressed using the revision operator introduced above, in section 5. As we remember, the truth condition for that operator is as follows:

(v) $x, y \models [*\alpha]\beta$ iff for all z such that $R(*\alpha)(x, y, z)$, $z, y \models \beta$.

In this condition, the state y about which we receive new information α is fixed beforehand, along with the prior state x. It cannot be required that y coincides with a posterior state z, since the appropriate candidates for z can only be determined after y has been fixed: z is any state such that $R(*\alpha)(x, y, z)$, and it would only be a happy coincidence if z turned out to coincide with y.

Let us refer to the forward-looking revision as $\oplus$. For the revision operator that corresponds to this doxastic action, we would need another truth condition:

(v') $x, y \models [\oplus\alpha]\beta$ iff for all z such that $R(*\alpha)(x, y, z)$, $z, z \models \beta$.

Note that, in this truth condition, the reference to y is idle. Since the new information is supposed to be about the posterior state, the need for an antecedently specified point of reference disappears. Thus, if we only want to discuss forward-looking revision, we might just as well give up the two-dimensional approach altogether and go back to the standard one-dimensional semantics. Instead of (v'), we then have a simpler condition:

$x \models [\oplus\alpha]\beta$ iff for all z such that $R(*\alpha)(x, z)$, $z \models \beta$.

Similarly, we might replace the two-dimensional truth condition for the belief operator with its standard one-dimensional version:

$x \models \mathbf{B}\alpha$ iff for all z such that $b(x, z)$, $z \models \alpha$.

In this condition, it is implicitly assumed that the beliefs held in a given state are about the state in which they are held.

It should be clear that forward-looking revision need not obey Preservation: Moving to a state in which one believes α may well necessitate adjustments in various higher-order beliefs. This is the lesson of Story 1 above. On the other hand, it seems reasonable to demand that forward-looking revision should obey Success. But then, as Story 2 shows, we must accept that revising with a consistent proposition sometimes will be impossible on pain of inconsistency . A Moorean proposition $\alpha \wedge \neg\mathbf{B}\alpha$ is consistent, but if I get as a task to move to a state in which I believe that proposition about that state itself, I will have to end up with inconsistent beliefs. In this sense, a proposition like this is a doxastic blindspot: It can be true, but it cannot be believed without inconsistency.

Some suggestions as to how to characterize this kind of forward-looking revision within unlimited DDL are to be found in Lindström & Rabinowicz (1999) (cf. our discussion of what we call "cautious revision"). But we are fully aware of the fact that the suggestions offered are still very preliminary.

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, Cambridge, MA: MIT Press.

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

Levi, I.: 1988, 'Iteration of conditionals and the Ramsey test', Synthese **76**, 49-81.

Lewis, D.: 1973, *Counterfactuals.* Oxford: Blackwells.

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.

Lindström, S., and Rabinowicz, W.: 1997, 'Extending dynamic doxastic logic: accommodating iterated beliefs and Ramsey conditionals within DDL', in Lindahl, Needham and Sliwinski (eds.) *For Good Measure*, Uppsala Philosophical Studies 46, 126-153.

Lindström, S. and Rabinowicz, W.: 1999, 'DDL Unlimited: Dynamic Doxastic Logic for Introspective Agents', Erkenntnis **50**, 353-385.

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.

de Rijke, M.: 1994, 'Meeting some neighbours: a dynamic modal logic meets theories of change and knowledge representation'. In *Logic and Information Flow*, edited by Jan van Eijck and Albert Visser, 170-196. Cambridge, MA: MIT Press.

Segerberg, K.: 1973, 'Two-dimensional modal logic', *Journal of Philosophical Logic* **2**, pp. 77-96.

Segerberg, K.: 1995, 'Belief revision from the point of view of doxastic logic'. *Bulletin of the IGPL*, vol. 3, 535-553.

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

Segerberg, K.: 1997, 'Proposal for a theory of belief revision along the lines of Lindström and Rabinowicz', *Fundamenta Informaticae* **32**, 183-191.

Segerberg, K.: 1999a, 'The basic dynamic doxastic logic of AGM'. In *The Goldblatt variations, pp. 76-107.* Uppsala Prints and Preprints in Philosophy, 1999:1.

Segerberg, K.: 1999b, 'A completeness proof in full DDL'. In Sliwinski, R., *Philosophical crumbs*. Essays dedicated to Ann-Mari Henschen-Dahlquist on the occasion of her seventy-fifth birthday, Uppsala Philosophical Studies 49, 195-207.