Segerberg on the Paradoxes of Introspective Belief Change

Enqvist, Sebastian; Olsson, Erik J

Published in:
Krister Segerberg on Logic of Action

2013

Segerberg on the Paradoxes of Introspective Belief Change

Sebastian Enqvist Erik J. Olsson

October 31, 2011

1 Introduction

Theories of rational belief change [1, 4, 5] are traditionally presented in a semi-formalized manner. While a formalized language is used to speak about the content of a state of belief, the theory of belief revision is, like most mathematical theories, presented in mathematical English rather than a formal language in the strict sense.

It is possible to formulate axioms of belief change, like the well-known AGM postulates [1], in a fully formalized language. This is the purpose of the socalled *Dynamic Doxastic Logic* (henceforth DDL, or "full" DDL) developed by Krister Segerberg [10, 12], in which epistemic states are modelled using modal operators of belief in the style of Jaakko Hintikka's classic [7], and belief *changes* are modelled using dynamic operators reminiscent of those studied in propositional dynamic logic [6].

Reasoning about belief in a formal language has the advantage of added expressive strength. Rather than just speaking about beliefs about the external world, we can now also reason about *introspective* beliefs, i.e. beliefs that an agent has about her *own state of belief*. For instance, I can believe that the world is round, which presumably means that I don't believe it is flat. Suppose now that someone asks me whether I believe the Earth is round; I answer that I do believe it. In these circumstances I am apparently *aware* that I believe the Earth is round, that is, I *believe that I believe* that the world is round. In the same manner, I might be asked whether I believe the Earth is flat, and I answer that I do not believe that. In this case, I have revealed that I *believe that I do not believe* that the Earth is flat. If r stands

for "the Earth is round" and f for "the Earth is flat", we can formalize these beliefs as

$$BBr$$

and

$$B\neg Bf$$

respectively.

In the case of DDL, where we have the capacity to speak about not only beliefs but also belief *change*, it turns out that this added expressive power comes with a price: given that we adopt the AGM postulate known as *Vacuity*, we arrive at some disturbing *paradoxes* of introspective belief change. These paradoxes are discussed at length by Sten Lindström and Wlodek Rabinowicz in [8], where a modification of the semantics of DDL is presented as a solution to the problem.

In this paper, we present and criticize Krister Segerberg's own solution to this problem. We present three alternative ways that the paradoxes of introspective belief change may be avoided: the first is a solution due to Sten Lindström and Wlodek Rabinowicz, using a *two-dimensional* semantics for DDL. The second solution is found in a logic for belief change suggested by Giacomo Bonanno, in which the operator for belief is replaced by a *class* of operators for belief, each supplied with a temporal index [3]. The third solution we present is a logic for belief change due to Johan van Benthem [14], founded on the method of Dynamic Epistemic Logic where dynamics is modelled by operations on entire models, rather than some structure within the models. We shall argue that, while there are some differences between these approaches, there is a strong structural similarity between them, and the they avoid the paradoxes of DDL in essentially the same way. Furthermore, the way that these logics avoid the paradoxes is both different from and, we think, more natural than Segerberg's own solution.

Throughout the discussion we presuppose familiarity with the AGM model of belief revision, as well as the basics of modal logic.

2 DDL and the paradoxes

We begin by introducing the system DDL and the paradoxes it gives rise to. Throughout the paper, we work with a fixed, countably infinite supply of propositional variables $Prop$. The language of DDL is then defined in

Backus-Naur form as follows, where $p \in Prop$:

$$\mathcal{L}_{DDL} : p \mid \neg\alpha \mid \alpha \vee \alpha \mid B\alpha \mid [*\alpha]\alpha$$

Classical connectives $\wedge, \rightarrow, \leftrightarrow$ are defined as usual. Informally, $B\alpha$ means "the agent believes α", and $[*\alpha]\beta$ means "after revision by α, it will be the case that β".

We now provide semantics for this language. Throughout the paper, given a binary relation R over a set W and given an element $u \in W$, we use the notation

$$R(u) =_{df.} \{v \in W : uRv\}$$

The logic of revision inherent in the semantics will be rather minimal, since the details of belief revision are irrelevant to the problem we address and its solutions. All we shall require of revision in this semantics, and in the other semantics presented in the paper, are the following conditions:

- after revision by α, the agent believes α
- revision by any consistent sentence results in a consistent belief state and
- Some semantic version of the *Vacuity* postulate holds.

We recall that, in the standard AGM framework for belief revision, the *Vacuity* postulate is:

$$\neg\alpha \notin K \Longrightarrow K * \alpha = Cn(K \cup \{\alpha\})$$

where Cn is the logical closure operator of the propositional logic underlying the epistemic states. This postulate says that if some input proposition is consistent with the agent's beliefs, then revision by that proposition amounts to simply adding the proposition to the initial stock of beliefs and forming the logical closure of the results; in other words, no information is *lost* in consistent revision.

Semantics for $\mathcal{L}_{DDL}$ is given as follows.

Definition 1. A *revision model* is a structure

$$\langle W, B, R^*, V\rangle$$

defined as follows: B is a binary relation over W, and $R^* : 2^W \rightarrow 2^{W\times W}$ is a function from subsets of W (sometimes called *propositions*) to relations over W. Furthermore we require that for each $X \subseteq W$, if $vR^*(X)w$ then

1. $B(w) \subseteq X$
2. if $X \neq \emptyset$ then $B(w) \neq \emptyset$
3. if $B(v) \cap X \neq \emptyset$ then $B(w) = B(v) \cap X$

Finally, $V : Prop \to 2^W$ is an evaluation function in the usual sense. A *pointed* revision model is a pair $(\mathfrak{A}, u)$ where $\mathfrak{A}$ is a revision model and $u \in W$.

The reader should note that the last item on the list in this definition is the obvious way to formulate the *Vacuity* postulate in the present framework. The truth definition for formulas of $\mathcal{L}_{DDL}$ in a pointed revision model is given as follows:

- $(\mathfrak{A}, u) \vDash p$ *iff* $u \in V(p)$
- standard clauses for Boolean connectives
- $(\mathfrak{A}, u) \vDash B\alpha$ *iff* $(\mathfrak{A}, v) \vDash \alpha$ for each v such that uBv
- $(\mathfrak{A}, u) \vDash [*\alpha]\beta$ *iff* $(\mathfrak{A}, v) \vDash \beta$ for each v such that $uR^*(\|\alpha\|)v$

Here, $\|\alpha\|$ denotes the set

$$\{w \in W : (\mathfrak{A}, w) \vDash \alpha\}$$

From this semantics we define the consequence relation $\vDash_{DDL}$ over $\mathcal{L}_{DDL}$ by setting, for all sets of formulas $\Gamma \cup \{\alpha\}$, $\Gamma \vDash_{DDL} \alpha$ *iff*

$$(\mathfrak{A}, u) \vDash \Gamma \Longrightarrow (\mathfrak{A}, u) \vDash \alpha$$

for any pointed revision model $(\mathfrak{A}, u)$. Here, $(\mathfrak{A}, u) \vDash \Gamma$ means that $(\mathfrak{A}, u) \vDash \beta$ for each $\beta \in \Gamma$.

We will need to be precise about what we mean by a logical system in this paper. Formally, a *logic* will here be taken to be a pair $(\mathcal{L}, \vDash)$ where $\mathcal{L}$ is a set containing the set of variables $Prop$ and $\vDash \subseteq 2^{\mathcal{L}} \times \mathcal{L}$. Thus, $(\mathcal{L}_{DDL}, \vDash_{DDL})$ is a logical system, which we denote by S_{DDL}.

To see why S_{DDL} is paradoxical, we ask the reader to verify that the following validity holds:

$$\vDash_{DDL} \ \neg B\neg\alpha \wedge B\beta \to [*\alpha]B\beta$$

and, furthermore, that we have the following validity:

$$\vDash_{DDL} \ [*\alpha]B\alpha$$

The former validy is called *Preservation* by Lindström and Rabinowicz, and the latter validity is called *Success*. The validity of *Preservation* is a direct consequence of the fact that the *Vacuity* postulate is built into the semantics. From *Preservation*, in turn, we derive the paradoxes: let α, β be any formulas. Then as a instance of *Preservation* we have

$$\vDash_{DDL} \ \neg B\neg\alpha \wedge B\neg B\alpha \rightarrow [*\alpha]B\neg B\alpha$$

On the other hand, from *Success* follows trivially by classical logic that:

$$\vDash_{DDL} \ \neg B\neg\alpha \wedge B\neg B\alpha \rightarrow [*\alpha]B\alpha$$

But clearly the operator $[*\alpha]$ is normal, so that we have

$$[*\alpha]B\alpha \wedge [*\alpha]B\neg B\alpha \vDash_{DDL} [*\alpha](B\alpha \wedge B\neg B\alpha)$$

By classical logic we can now derive:

$$\vDash_{DDL} \neg B\neg\alpha \wedge B\neg B\alpha \rightarrow [*\alpha](B\alpha \wedge B\neg B\alpha)$$

This is the formula deemed paradoxical by Lindström and Rabinowicz, and it would be hard to deny that it is quite bizarre. To see why, toss a coin, without looking at it when it lands. Presumably, given that the coin is fair, you now have no opinion at all on whether the coin landed heads or tails. Let α stand for the proposition that the coin landed heads. Since you have no opinion on whether the toss came out heads or tails, you do not believe that the coin did *not* land heads. That is, your current belief state satisfies the condition $\neg B\neg\alpha$. But you do not believe that the coin *did* land heads, and we think that you have the required powers of introspection to be aware of this fact. Thus, your current belief state also satisfies the condition $B\neg B\alpha$. But then, according to DDL, the condition $[*\alpha](B\alpha \wedge B\neg B\alpha)$ should also be true. This means that if you were to take a look at the coin and learn that it did in fact land heads, as a result you should believe that the coin landed heads, but at the same time you should *believe that you do not believe it*. Under perfectly ordinary circumstances, revision of beliefs has lead to a curious, or even incoherent, state of belief.

If we simply dropped the *Vacuity* postulate, then the problem would disappear. But for those who are strongly convinced of the validity of Vacuity, the more attractive route would be to try and retain some semantic version of the *Vacuity* postulate, while employing some strategy to avoid the paradoxes. In the following section, we present Segerberg's own strategy for doing so.

3 Segerberg's solution

Segerberg treats the paradoxes of introspective belief change, which he refers to as "Moore problems", in a paper from 2006 [11]. In this paper, he proposes a solution based on Sorensen's notion of a *blindspot* from his 1988 book [13].

In Segerbergs terminology, an agent has a *Moore problem* (of rank 0) if $B(\phi \wedge \neg B\phi)$ or $B(\phi \wedge B\neg\phi)$ is true (in a certain situation and with respect to his beliefs). In the former case, the problem is said to be *acute*, in the latter *grave*. More generally, the agent has a Moore problem of rank n, where n is a nonnegative integer, if, for some formula ϕ, either $B^n(\phi \wedge \neg B\phi)$ or $B^n(\phi \wedge B\neg\phi)$, where B^n abbreviates

$$\underbrace{B \ldots B}_{n \text{ times}}$$

Segerberg is very clear on the desirability of avoiding Moore problems:

> It is probably impossible to compile a complete list of all the ways in which a doxastic agent may be incoherent or exhibit some degree of inconsistency, but certainly an agent with a Moore problem of any rank is not perfect. Doxastically ambitious agents will stay clear of Moore problems as far as possible! ([11], p.96)

Segerbergs solution seems radical on first sight: he proposes to reject the assumption that the star operator correctly formalizes revision. Revision by ϕ is not to be formalizes as $*\phi$ but rather as

$$\mathbf{R}\phi =_{df.} *(\phi \wedge B\phi)$$

As Segerberg points out, the Preservation and Success conditions are not affected by this definition, meaning that the derivations of the Moorean sentences are still valid inferences. Yet, the conclusions are no longer an embarrassment:

> [...] for the fact that , in a certain possible situation, a star change leads to a Moore problem is not embarrassing, however plausible the situation - why would one want to perform a star change anyway? ([11], p. 101).

What *would* be troublesome is if the corresponding sentences could be derived for revision, i.e. if we could derive

$$(\neg B\neg\phi \wedge B\neg B\phi) \rightarrow [\mathbf{R}\phi]B(\phi \wedge \neg B\phi)$$

and

$$(\neg B\neg\phi \wedge BB\neg\phi) \rightarrow [\mathbf{R}\phi]B(\phi \wedge B\neg\phi)$$

But these sentences are not derivable. Hence his new definition of revision avoids the Moore problems of rank 0. However, as Segerberg shows, some new problems crop up in their stead. Suppose ϕ is such that before revision by ϕ, $\neg B\neg(\phi \wedge B\phi)$ is true, and that, before revision, either $B\neg BB\phi$ or $BB\neg B\phi$ or $BBB\neg\phi$ is true. Then it follows, using Preservation and Success, that after revision by ϕ, on the new understanding of revision, at least one of $B(B\phi \wedge \neg BB\phi)$ or $BB(\phi \wedge \neg B\phi)$ or $BB(\phi \wedge B\neg\phi)$ is true. Thus the agent is confronted with a Moore problem of rank 1.

How can this situation be avoided? Segerberg's main idea is that the predicament can be avoided by making the problematic sets of sentences *inconsistent*, "for inconsistent sets describe (what according to the logic) are impossible situations, and it is of no concern that Moore problems arise in impossible situations" ([11], p. 100). In the present case, this strategy translates into finding a plausible underlying logic that makes each of the following sets inconsistent: $\{\neg B\neg(\phi \wedge B\phi), B\neg BB\phi\}$, $\{\neg B\neg(\phi \wedge B\phi), BB\neg B\phi\}$ and $\{\neg B\neg(\phi \wedge B\phi), BBB\neg\phi\}$. Segerberg notes that the weakest normal logic satisfying this condition is the normal extension of K by the following schemata:

(1A) $B\neg BB\phi \rightarrow B\neg(\phi \wedge B\phi)$

(1B) $BB\neg B\phi \rightarrow B\neg(\phi \wedge B\phi)$

(1C) $BBB\neg\phi \rightarrow B\neg(\phi \wedge B\phi)$

Segerberg shows that all three are derivable, for instance, in KD4 which is a favorite with many doxastic logicians ([11], p. 102). He then goes on to generalize this approach to Moore problems at rank n and of rank ω, showing that the problematic situations can be excluded by a reasonable choice

of underlying doxastic logic. Finally, Segerberg connects his approach to Sorensens concept of a blindspot by defining a *blindspot* as a sentence ϕ such that either ϕ is not entertainable or revision by it leads to an inconsistent state and showing that the following principle comes out as valid on his approach: revision by an entertainable proposition leads to a consistent doxastic state if and only if the sentence in question is not a blindspot. Since

$$[\mathbf{R}(\phi \wedge \neg B\phi)]B \perp$$

and

$$[\mathbf{R}(\phi \wedge B\neg\phi)]B \perp$$

are theorems in all logics recommended by Segerberg, in those logics the original Moore sentences $\phi \wedge \neg B\phi$ and $\phi \wedge B\neg\phi$ are blindspots.

This is certainly an impressive treatment of the Moore problems, especially considering the proposal, which we will grant, that the Moore problems arise in impossible situations where what is impossible or not is defined in a principled manner relative to logical frameworks that have an independent standing in the literature. Segerberg can hardly be accused of adhockery in that respect. However, Segerberg's strategy may still be ad hoc in another regard. Consider again Segerberg's new definition of revision by ϕ, i.e. $\mathbf{R}\phi =_{df.} *(\phi \wedge B\phi)$. First of all, it surely is less simple and striking than the old one. But second and more important, Segerberg does not give any independent motivation for his new definition of revision. Certainly, defining revision in this way does the job of providing a framework within which Moore problems can be avoided, but apart from this fact little speaks in favor of the new definition. And, one might ask, why should every revision by ϕ be, as it were, accompanied by a revision by $B\phi$? Suppose ϕ is an object level sentence such as it is raining. Why should updating by it is raining involve updating by I believe that it is raining? Of course, it may often be the case that these two propositions are accommodated in one swoop, but it is less clear that it has to be that way. For certain kinds of introspective agents the new definition of revision may be fine. But what about agents that adopt beliefs routinely without reflecting on those beliefs at the time of adoption? So there is still a sense in which Segerberg's approach is, at least to some extent, ad hoc.

Another way of putting it is that Segerberg gives but a partial solution to the Moore problems, a solution that takes care of those problems for reflective agents (by which we mean agents for which an update by ϕ is

always accompanied by an update by $B\phi$), but that he has little to say about the prospects of dealing with those problems from the perspective of unreflective agents.

In the light of these remarks, it is natural to ask whether there is some other way to treat the paradoxes of full DDL. In the next section, we present *three* different logics for belief revision that can be found in the literature, each of which can be shown to avoid the paradoxes of introspective belief change. We shall begin by introducing each variant formally, and then discuss what we believe is the common structure behind each approach.

4 Three alternative solutions

4.1 First solution: two-dimensional DDL

The first solution we consider is due to Sten Lindström and Wlodek Rabinowicz. The approach suggested by Lindström and Rabinowicz is to adopt a modified, two-dimensional semantics for DDL in which formulas are no longer evaluated at single worlds, but rather at *pairs* of worlds. Here, the idea is that in an evaluation point (u, v), the left component u serves as a *point of reference*, while v functions as a *point of evaluation*. In addition, rather than an accessibility relation B over the universe of a model, a class of accessibility relations is used, on relative to each world in the universe. Each accessibility relation B_v, where $v \in W$, represents the agent's beliefs *about* the point of reference v.

Formally, the definition of a model from the system S_{DDL} is modified as follows:

Definition 2. A *two-dimensional revision model* is a structure

$$\langle W, \{B_u\}_{u \in W}, \{R^*_u\}_{u \in W}, V \rangle$$

defined as follows: for each $u \in W$, B_u is a binary relation over W. For each $u \in W$ $R^*_u : 2^W \to 2^{W \times W}$ is a function from propositions to relations between possible worlds, such that for each $X \subseteq W$, if $vR^*_u(X)w$ then

1. $B_u(w) \subseteq X$
2. if $X \neq \emptyset$ then $B_u(w) \neq \emptyset$
3. if $B_u(v) \cap X \neq \emptyset$ then $B_u(w) = B_u(v) \cap X$

A *pointed* two-dimensional revision model is a triple $(\mathfrak{A}, u, v)$ where $\mathfrak{A}$ is a two-dimensional revision model and $u, v \in W$.

To speak about these models, we use an extension of the language $\mathcal{L}_{DDL}$. Formally, the language $\mathcal{L}_{2D}$ is given by the following definition where, again, $p \in Prop$:

$$\mathcal{L}_{2D} : p \mid \neg\alpha \mid \alpha \vee \alpha \mid B\alpha \mid [*\alpha]\alpha \mid \dagger\, \alpha$$

The truth definition for formulas is given as follows:

- $(\mathfrak{A}, u, v) \vDash p$ *iff* $v \in V(p)$
- standard Boolean clauses
- $(\mathfrak{A}, u, v) \vDash B\alpha$ *iff* $(\mathfrak{A}, u, w) \vDash \alpha$ for each w such that vB_uw
- $(\mathfrak{A}, u, v) \vDash [*\alpha]\beta$ *iff* $(\mathfrak{A}, u, w) \vDash \beta$ for each w such that $vR^*_u(\|\alpha\|_u)w$
- $(\mathfrak{A}, u, v) \vDash \dagger\, \alpha$ *iff* $(\mathfrak{A}, v, v) \vDash \alpha$

The consequence relation $\vDash_{2D}$ is defined from this semantics as before, and we let S_{2D} denote the logical system $(\mathcal{L}_{2D}, \vDash_{2D})$. The new component of the language of this logic is the †-operator, although the meanings of the modal operators present in $\mathcal{L}_{DDL}$ have changed. This operator has the effect of making the current point of evaluation the current point of reference as well. The formula $\dagger\, \alpha$ can informally be interpreted as saying that α is true *about* the present point of evaluation.

How does this avoid the paradoxes of DDL? As noted by Lindström and Rabinowicz, for each formula α the paradoxical formula of S_{DDL} which we recall was:

$$\neg B\neg\alpha \wedge B\neg B\alpha \rightarrow [*\alpha](B\alpha \wedge B\neg B\alpha)$$

is still valid in this semantics. *But*, as we said, the meaning of the connectives has changed. Consider an evaluation point (u, u) in a model $\mathfrak{A}$ (here, the point of reference and the point of evaluation is the same). Suppose that

$$(\mathfrak{A}, u, u) \vDash \neg B\neg\alpha \wedge B\neg B\alpha$$

so that the agent does not disbelieve α at (u, u) and she believes that she does not believe α. According to the validity of the previously deemed paradoxical formula, we must have

$$(\mathfrak{A}, u, u) \vDash [*\alpha](B\alpha \wedge B\neg B\alpha)$$

This means that

$$(\mathfrak{A}, u, v) \vDash B\alpha \wedge B\neg B\alpha$$

But this is not incoherent, since the righthand conjunct here means that at the *present* point of evaluation (v), the agent believes that the condition $\neg B\alpha$ holds for the point of evaluation *prior* to revision by α (u). By contrast, the formula

$$\neg B\neg\alpha \wedge B\neg B\alpha \rightarrow [*\alpha]\dagger\ (B\alpha \wedge B\neg B\alpha)$$

is paradoxical, since the beliefs described as resulting after revision by α are now beliefs about the point of evaluation after revision, not the one prior to it. But this formula is *not* valid in S_{2D}. Thus, the system S_{2D} is free of this paradoxical feature of S_{DDL}.

4.2 Second solution: temporally indexed beliefs

The second solution, present in Giacomo Bonanno's "simple" modal logic for belief revision, consists in letting a model represent an ω-ordered discrete time-line and use belief operators supplied with indexes representing points in the succession of time. Each move forward in time corresponds to an act of revision by some new piece of information. The expression $B_n\alpha$, where $n \in \omega$, says that the agent believes α *at* the nth point in the succession of time.

Formally, the language $\mathcal{L}_{Temp}$ for Bonanno's system is defined as follows, where n is any natural number:

$$\mathcal{L}_{Temp} : p \mid \neg\alpha \mid \alpha \vee \alpha \mid B_n\alpha \mid I_n\alpha$$

The new operator I_n is interpreted so that $I_n\alpha$ means, informally, that α is the last piece of information received at the nth point in time, or that α it is the input of the revision that results in the belief state at time n.

Semantics for $\mathcal{L}_{Temp}$ is given by the following definitions:

Definition 3. A *temporal belief model* is a structure

$$\langle W, \{B_n\}_{n\in\omega}, \{I_n\}_{n\in\omega}, V\rangle$$

such that:

1. $B_n(u) \subseteq I_n(u)$

2. if $I_n(u) \neq \emptyset$ then $B_n(u) \neq \emptyset$

3. if $B_n(u) \cap I_{n+1}(u) \neq \emptyset$ then $B_{n+1}(u) = B_n(u) \cap I_{n+1}(u)$

A *pointed* temporal belief model is a pair $(\mathfrak{A}, u)$ where $\mathfrak{A}$ is a temporal belief model and $u \in W$.

Truth definitions of formulas in a pointed temporal belief model are:

- $(\mathfrak{A}, u) \vDash p$ *iff* $u \in V(p)$
- standard clauses for Boolean connectives
- $(\mathfrak{A}, u) \vDash B_n\alpha$ *iff* $(\mathfrak{A}, v) \vDash \alpha$ for each v such that uB_nv
- $(\mathfrak{A}, u) \vDash I_n\alpha$ *iff* $I_n(u) = \|\alpha\|$

Here, as before, $\|\alpha\| = \{v \in W : (\mathfrak{A}, v) \vDash \alpha\}$. From this semantics we define the consequence relation $\vDash_{Temp}$ as before, and we define S_{Temp} to be the logical system $(\mathcal{L}_{Temp}, \vDash_{Temp})$. To get a grasp of how the language works, consider the syntactic form of the *Success* postulate in this system; this is captured by the validity

$$\vDash_{Temp} \ I_n\alpha \rightarrow B_n\alpha$$

This says that if the belief state at time n is the result of the revision of a prior belief state by α, then α is believed at time n.

The way that the paradoxes of DDL are avoided in this system is simple. We do have a form of the *Preservation* formula valid in S_{Temp}, namely:

$$\vDash_{Temp} \ \neg B_n\neg\alpha \wedge B_n\beta \rightarrow (I_{n+1}\alpha \rightarrow B_{n+1}\beta)$$

That is, if α is consistent with the agent's beliefs at time n, and the next revision at time $n+1$ has the input formula α, then everything the agent believes at time n she believes at time $n+1$ also. But we cannot derive any paradoxes from this formula, since belief operators come with temporal indexes. To see this, let's try to derive a paradox in the same manner as before. Consider any formula α. As an instance of the previous validity we get

$$\vDash_{Temp} \ \neg B_n\neg\alpha \wedge B_n\neg B_n\alpha \rightarrow (I_{n+1}\alpha \rightarrow B_{n+1}\neg B_n\alpha)$$

From this, using the S_{Temp}-version of *Success* above together with classical logic, we can derive

$$\vDash_{Temp} \ \neg B_n\neg\alpha \wedge B_n\neg B_n\alpha \rightarrow (I_{n+1}\alpha \rightarrow B_{n+1}\alpha \wedge B_{n+1}\neg B_n\alpha)$$

The informal content of this formula looks a lot like that of the paradoxical formula we derived in S_{DDL}. But of course, it is not paradoxical. It says that if α is consistent with the agents beliefs at time n, and the agent is aware that she does not believe α at time n, then after revision by α at time $n+1$, she believes α and *believes that she did not believe it* at time n. By contrast, the formula

$$\models_{Temp} \ \neg B_n \neg\alpha \wedge B_n \neg B_n \alpha \rightarrow (I_{n+1}\alpha \rightarrow B_{n+1}\alpha \wedge B_{n+1}\neg B_{n+1}\alpha)$$

is paradoxical but not valid.

4.3 Third solution: the DEL method

The third alternative way of getting out of the paradoxes of DDL we consider in this paper is found in Johan van Benthem's dynamic logic for belief revision. The system is built on a method used in Dynamic Epistemic Logic (DEL), a framework for studying dynamics of multi-agent epistemic scenarios. The relevant aspect of DEL here is not the multi-agent feature, but rather the way in which dynamics is modelled semantically and reasoned about syntactically.

The method can be described in this way: to model changes of some type of *states*, one should first develop a *static base language* for reasoning about the states and provide it with a semantics, i.e. define *models* for it. Then, changes of states are modelled as operations on models for the static base language, which is extended with dynamic operators to reason about these operations. If the static base logic is rich enough in expressive strength, then it is often possible to translate any dynamic formula into a semantically equivalent formula of the static base logic via socalled *reduction axioms*.

For brevity we will present the static and the dynamic part of van Benthem's system all in one swoop. For a gentler presentation of the system we refer to [14]. For an introduction to DEL, see [15].

We begin by defining the models for the static part of the logic:

Definition 4. A *conditional belief model* is a structure

$$\langle W, \{\sigma_u\}_{u\in W}, V\rangle$$

defined as follows. For each $u \in W$, $\sigma_u : 2^W \rightarrow 2^W$ is called a *selection function* and satisfies the following properties:

1. $\sigma_u(X) \subseteq X$
2. $X \neq \emptyset$ implies $\sigma_u(X) \neq \emptyset$
3. if $\sigma_u(X) \cap Y \neq \emptyset$ then $\sigma_u(X \cap Y) = \sigma_u(X) \cap Y$

V is a valuation function as before, and pointed conditional belief models are defined as before.

The central component of these models is the set of selection functions, which can be thought of as encoding the *conditional beliefs* of the agent. The intuitive explanation is that, for each proposition $X \subseteq W$, the set $\sigma_u(X)$ consists of the "most plausible" worlds from the agent's point of view at the world u. *Actual beliefs* of the agent are defined as beliefs conditional on the trivial proposition. That is, the set of possible worlds compatible with the agent's actual beliefs at the world u is the set $\sigma_u(W)$. The semantics presented in [14] is a bit different from the presentation here and uses orders of plausibility rather than selection functions, but this is irrelevant to the current issue.

To model dynamics of the models, we will use an operation that van Benthem calls *lexicographic upgrade*. Or, rather, we use a version of this operation, adapted to the semantics here based on selection functions which is slightly more general than van Benthem's semantics. Consider a proposition $X \subseteq W$ in a model $\mathfrak{A}$; we want a way to *revise* the selection function u at a world u by X. This is provided by the following definition:

Definition 5. $\sigma_u^{\Uparrow X}(Y) = \begin{cases} \sigma_u(Y) \cap X & \text{if } \sigma_u(Y) \cap X \neq \emptyset \\ \sigma_u(X) & \text{if } \sigma_u(Y) \cap X = \emptyset \end{cases}$

Given a conditional belief model $\mathfrak{A} = \langle W, \{\sigma_u\}_{u \in W}, V \rangle$ and $X \subseteq W$, we define the revised model $\mathfrak{A} \Uparrow X$ by

$$\mathfrak{A} \Uparrow X =_{df.} \langle W, \{\sigma_u^{\Uparrow X}\}_{u \in W}, V \rangle$$

We leave it to the reader to check that this is always a well defined conditional belief model. Notice that we have

$$\sigma_u(W) \cap X \neq \emptyset \quad \Longrightarrow \quad \sigma_u^{\Uparrow X}(W) = \sigma_u(W) \cap X$$

With the definition of actual beliefs as beliefs conditional on the trivial proposition, this property can be seen as a semantic formulation of the *Vacuity* postulate.

Turning to the syntactic side of the system, we define the language $\mathcal{L}_{DEL}$:

$$\mathcal{L}_{DEL} : p \mid \neg\alpha \mid \alpha \vee \alpha \mid B(\alpha \mid \alpha) \mid A\alpha \mid [\Uparrow \alpha]\alpha$$

Here, $B(\alpha \mid \beta)$ says that α is believed conditionally on β, and $[\Uparrow \alpha]\beta$ says that the condition β will hold after revision by α. The operator A is the *global necessity* operator (see [2]). $A\alpha$ means that α holds in all possible worlds of a model; it can be thought of as expressing *logical* necessity. A *static* formula of $\mathcal{L}_{DEL}$ is a formula without any occurrences of the dynamic operators. We define an operator for actual beliefs by $B\alpha =_{df.} B(\alpha \mid p \vee \neg p)$, where p is a propositional variable.

Truth definitions for formulas in a pointed model are:

- $(\mathfrak{A}, u) \vDash p$ *iff* $u \in V(p)$
- standard clauses for Boolean connectives
- $(\mathfrak{A}, u) \vDash B(\alpha \mid \beta)$ *iff* $\sigma_u(\|\beta\|) \subseteq \|\alpha\|$
- $(\mathfrak{A}, u) \vDash A\alpha$ *iff* $(\mathfrak{A}, v) \vDash \alpha$ for each $v \in W$
- $(\mathfrak{A}, u) \vDash [\Uparrow \alpha]\beta$ *iff* $(\mathfrak{A} \Uparrow \|\alpha\|, u) \vDash \beta$

The consequence relation $\vDash_{DEL}$ and the system S_{DEL} are now defined as before.

It is instructive to look at the reduction axioms for S_{DEL}. These are as follows (we follow van Benthem's axiomatization almost without any modification):

$\Uparrow$**1:** $[\Uparrow\gamma]\, q \leftrightarrow q$, q *a propositional atom*

$\Uparrow$**2:** $[\Uparrow \gamma]\neg\alpha \leftrightarrow \neg[\Uparrow \gamma]\alpha$

$\Uparrow$**3:** $[\Uparrow \gamma](\alpha \vee \beta) \leftrightarrow ([\Uparrow \gamma]\alpha \vee [\Uparrow \gamma]\beta)$

$\Uparrow$**4:** $[\Uparrow \gamma]A\alpha \leftrightarrow A[\Uparrow \gamma]\alpha$

$\Uparrow$**5:** $[\Uparrow \gamma]B(\alpha \mid \beta) \leftrightarrow$
$\leftrightarrow (E(\gamma \wedge [\Uparrow \gamma]\beta) \wedge B([\Uparrow \gamma]\beta \rightarrow [\Uparrow \gamma]\alpha \mid \gamma) \vee$
$\vee (\neg E(\gamma \wedge [\Uparrow \gamma]\beta) \wedge B([\Uparrow \gamma]\alpha \mid [\Uparrow \gamma]\beta))$

The reader can check that these axioms are sound in the semantics for S_{DEL}. These axioms can be thought of as providing recursive *definitions* of the truth conditions of dynamic formulas in terms of static formulas. Together with a suitable set of complete axioms for the static fragment of S_{DEL} and a rule for substitution of equivalents, they provide a complete axiomatization of S_{DEL}. To prove this result, one exploits the soundness of the reduction axioms to prove the following proposition as a lemma. The proof is excluded here.

Proposition 1. *There exists a function $\rho : \mathcal{L}_{DEL} \to \mathcal{L}_{DEL}$ such that for each formula $\alpha \in \mathcal{L}_{DEL}$, the formula $\rho(\alpha)$ is a static formula and, furthermore, for each pointed conditional belief model $(\mathfrak{A}, u)$,*

$$(\mathfrak{A}, u) \vDash \alpha \Longleftrightarrow (\mathfrak{A}, u) \vDash \rho(\alpha)$$

To get a feel for the system, let us look at some validities. Here, p, q are two propositional variables and $\bot$ is any tautological contradiction. First, revision by p leads the agent to believe p:

$$\vDash_{DEL} [\Uparrow p]Bp$$

Second, revision by a consistent sentence results in a consistent belief state:

$$\vDash_{DEL} \neg A \neg p \to [\Uparrow p] \neg B \bot$$

What about *Preservation*? We do indeed have a form of the *Preservation* principle valid in this system:

$$(i) \quad \vDash_{DEL} \neg B \neg p \wedge Bq \to [\Uparrow p]Bq$$

Now, if validity in S_{DEL} were closed under substitutions for propositional variables (as is the case in most logics), then obviously we could derive a paradox in the same manner as in S_{DDL}. However, this is not the case, and it is in fact this feature of S_{DEL} that makes it non-paradoxical. In particular, the following substition instance of (i):

$$(ii) \quad \neg B \neg \alpha \wedge B \neg B \alpha \to [\Uparrow \alpha] B \neg B \alpha$$

is invalid. This is exactly the formula that would be required to derive a paradox in S_{DEL}. By contrast, the following formula is valid:

$$(iii) \quad \neg B \neg \alpha \wedge B \neg B \alpha \to B(\neg B \alpha \mid \alpha)$$

Now, what does this formula say? it says that, if $\neg B\neg\alpha$ and $B\neg B\alpha$ are true at some world in a model, than from the point of view of the agent in that world, $\neg B\alpha$ will be *true in the most plausible worlds where* α *is true*. Now, the most plausible worlds where α is true, *prior* to revision by α, are exactly those worlds that are compatible with the agent's actual beliefs *after* revision. But since the truth values of formulas involving beliefs will change at *every* world in a model through the act of revision by α, it does *not* follow from this that $\neg B\alpha$ will be true at all worlds that are compatible with the agent's beliefs after the revision. This is why (ii) fails to be valid.

5 Comparison of the solutions

5.1 What the three approaches have in common

The three solutions we have just presented are, we think, essentially one and the same. All three of them are based on making a distinction between two different perspectives, the state of affairs *prior* to revision and the one *after* revision. This is perhaps clearest in Lindström and Rabinowicz's system; it is embodied quite explicitly in the distinction between the point of *reference* (typically being the state *prior* to revision) and the point of *evaluation* (typically the state *after* revision).

But we see the same distinction very clearly in Bonanno's temporal system of belief revision, although in a different form. Here, it turns up through the temporally indexed belief operators. In particular, in the formula

$$\neg B_n\neg\alpha \wedge B_n\neg B_n\alpha \rightarrow (I_{n+1}\alpha \rightarrow B_{n+1}\alpha \wedge B_{n+1}\neg B_n\alpha)$$

which is provable in S_{Temp}, the state prior to revision corresponds to the time-point represented by the number n, and the state after revision corresponds to $n+1$.

It is perhaps a bit less obvious how van Benthem's system S_{DEL} fits into this picture, but we think it does. We postpone the task of explaining this to section 5.3, where we will be better prepared to do so. The fact that the same solution to the paradoxes can be found in three seemingly rather different frameworks for belief revision counts, we think, as evidence in favor of this approach as a particularly natural way to resolve the paradoxes. Think of it in analogy with the case of various definitions of computable functions, for example in terms of recursive functions or in terms of Turing machines.

The wellknown fact that these definitions turn out to be equivalent speaks strongly in favor of the idea that they all capture the pre-formal notion of computability in a natural way. The present situation, where three different formalisms can be seen to resolve the paradoxes of DDL in the same way, is similar.

To strengthen these claims, we shall establish a formal correspondence between the three logical systems S_{2D}, S_{Temp} and S_{DEL}. More specifically, we shall show that the system S_{2D} can in a precise sense be *interpreted* in S_{Temp}, and in turn, S_{DEL} can be interpreted in S_{2D}. From this will follow that S_{DEL} can be interpreted in S_{Temp} also. These interpretation results will help to clarify the deeper connection that we think exists between the different systems, particularly with respect to how they handle the paradoxes of DDL. In order to formally prove these claims, we need to make precise what it means that a logical system can be interpreted in another. This is captured by the following definition.

Definition 6. Given logical systems $S_1 = (\mathcal{L}_1, \vDash_1)$ and $S_2 = (\mathcal{L}_2, \vDash_2)$, an *interpretation* of S_1 in S_2 is any function $F : \mathcal{L}_1 \to \mathcal{L}_2$ such that $F(p) = p$ for any $p \in Prop$. The interpretation F is said to be a *sound* interpretation of S_1 if, for all sets of formulas $\Gamma \cup \{\alpha\} \subseteq \mathcal{L}_1$, we have

$$\Gamma \vDash_1 \alpha \Longrightarrow F(\Gamma) \vDash_2 F(\alpha)$$

So a sound interpretation of a logical system S_1 in S_2 is a translation that maps sentences of S_1 to sentences of S_2 in a way that preserves logically valid consequences. Just like when we interpret a logical system in a semantics, we might consider the question of whether an interpretation is *complete* in addition to being sound. We could say that an interpretation F of S_1 in S_2 is *sound and complete* if, for all $\Gamma \cup \{\alpha\} \subseteq \mathcal{L}_1$, we have

$$\Gamma \vDash_1 \alpha \Longleftrightarrow F(\Gamma) \vDash_2 F(\alpha)$$

The issue of completeness will not concern us in this paper. Rather, we will focus on soundness. Completeness is a welcome property of any interpretation of a logical system, but soundness is absolutely crucial. If an interpretation is not sound, it is doubtful whether it can be called a proper interpretation at all. Also, as we shall see in the next section, the soundness propery of the interpretations we provide is enough to make the correspondence quite enlightening.

5.2 Interpreting S_{2D} in S_{Temp}

Our first result is that, in the sense of Definition 6, there exists a sound interpretation F of S_{2D} in S_{Temp}. First, by induction over the complexity of formulas, we define the class of functions

$$\tau_{n,m} : \mathcal{L}_{2D} \to \mathcal{L}_{Temp}$$

where $n, m \in \omega$ as follows:

1. $\tau_{n,m}(p) = p$
2. $\tau_{n,m}(\neg\alpha) = \neg\tau_{n,m}(\alpha)$
3. $\tau_{n,m}(\alpha \vee \beta) = \tau_{n,m}(\alpha) \vee \tau_{n,m}(\beta)$
4. $\tau_{n,m}(B\alpha) = B_m\tau_{n,n}(\alpha)$
5. $\tau_{n,m}([*\alpha]\beta) = I_{m+1}\tau_{n,n}(\alpha) \to \tau_{n,m+1}(\beta)$
6. $\tau_{n,m}(\dagger\ \alpha) = \tau_{m,m}(\alpha)$

We then set $F =_{df.} \tau_{0,0}$. For this mapping F we have the following result, proved in Appendix A.1:

Theorem 1. *The translation F constitutes a sound interpretation of the system S_{2D} in the system S_{Temp}.*

To see how this interpretation relates the two systems to each other, let us consider the interpretation of the formula

$$\neg B\neg p \wedge B\neg Bp \to [*p](Bp \wedge B\neg Bp)$$

given by F. This formula is an instance of the paradoxical schema we derived in S_{DDL}. As we mentioned earlier, the formula is valid in S_{2D} also, and therefore by soundness its interpretation under F is valid in S_{Temp}. Now, Lindström and Rabinowicz claim that this formula is *not* paradoxical under the interpretation given to it in two-dimensional semantics. Then, its interpretation under F had better not be paradoxical either!

Indeed it is not. For the formula

$$(1) \quad F(\neg B\neg p \wedge B\neg Bp \to [*p](Bp \wedge B\neg Bp))$$

is identical to

$$(2) \quad (\neg B_0 \neg p \wedge \neg B_0 \neg B_0 p) \rightarrow (I_1 p \rightarrow B_1 p \wedge B_1 \neg B_0 p)$$

which is perfectly fine. We can derive this as follows: first, recalling that $F = \tau_{0,0}$ and using translation clauses for Boolean connectives, atomic formulas and B, the formula (1) becomes

$$\neg B_0 p \wedge B_0 \neg B_0 p \rightarrow \tau_{0,0}([*p](Bp \wedge B \neg Bp))$$

Carrying out the translation further, we get

$$\neg B_0 p \wedge B_0 \neg B_0 p \rightarrow (I_1 p \rightarrow \tau_{0,1}(Bp) \wedge \tau_{0,1}(B \neg Bp))$$

Applying the function $\tau_{0,1}$ to its arguments here, we get

$$\neg B_0 p \wedge B_0 \neg B_0 p \rightarrow (I_1 p \rightarrow B_1 p \wedge B_1 \tau_{0,0}(\neg Bp))$$

But $\tau_{0,0}(\neg Bp) = \neg B_0 p$, so now we arrive at (2) as desired.

By contrast, let's look at the translation of the formula

$$\neg B \neg p \wedge B \neg Bp \rightarrow [*p] \dagger (Bp \wedge B \neg Bp)$$

which *is* paradoxical. Applying the translation F to this formula, instead of (2) we will get the formula

$$(3) \quad (\neg B_0 \neg p \wedge \neg B_0 \neg B_0 p) \rightarrow (I_1 p \rightarrow B_1 p \wedge B_1 \neg B_1 p)$$

which is indeed paradoxical, and not valid in S_{Temp}. To see what happens here, we can carry out the translation step by step and check that we eventually arrive at the formula

$$\neg B_0 p \wedge B_0 \neg B_0 p \rightarrow (I_1 p \rightarrow \tau_{0,1}(\dagger(Bp \wedge B \neg Bp))$$

Applying the translation clause for $\dagger$, this becomes

$$\neg B_0 p \wedge B_0 \neg B_0 p \rightarrow (I_1 p \rightarrow \tau_{1,1}(Bp \wedge B \neg Bp)$$

But

$$\tau_{1,1}(Bp \wedge B \neg Bp) = B_1 p \wedge B_1 \tau_{1,1}(\neg Bp) = B_1 p \wedge B_1 \neg B_1 p$$

and so we arrive at (3).

5.3 Interpreting S_{DEL} in S_{2D}

We now show how to interpret S_{DEL} in S_{2D}. The central observation here is that, since we know that there is a translation ρ that sends every formula α to an equivalent *static* formula $\rho(\alpha)$, it suffices to interpret the static formulas of S_{DEL} in order to get a full interpretation of S_{DEL} in S_{2D}.

Formally, we define the mapping τ as follows:

1. $\tau(p) = p$
2. $\tau(\neg\alpha) = \neg\tau(\alpha)$
3. $\tau(\alpha \vee \beta) = \tau(\alpha) \vee \tau(\beta)$
4. $\tau(A\alpha) = [*\neg\tau(\alpha)]B \bot$
5. $\tau(B(\alpha \mid \beta)) = [*\tau(\beta)]B\tau(\alpha)$

Clearly, every *static* formula of $\mathcal{L}_{DEL}$ receives an interpretation by this mapping. Letting ρ be any translation function as specified in Proposition 1, we define an interpretation $F : \mathcal{L}_{DEL} \to \mathcal{L}_{2D}$ by setting

$$F(\alpha) = \tau(\rho(\alpha))$$

for each $\alpha \in \mathcal{L}_{DEL}$. As before, we have the following soundness result for this interpretation:

Theorem 2. *The translation F constitutes a sound interpretation of the system S_{DEL} in the system S_{2D}.*

The proof of this result is in Appendix A.2. Furthermore, the composition of two sound interpretations (whenever it is well defined) is obviously a sound interpretation. So by the existence of a sound interpretation of S_{DEL} in S_{2D} and a sound interpretation of S_{2D} in S_{Temp}, we get:

Corollary 1. *There exists a sound interpretation of S_{DEL} in S_{Temp}.*

An interesting aspect of the translation F presented in this section is that, clearly, for any $\mathcal{L}_{DEL}$-formula α, the corresponding $\mathcal{L}_{2D}$-formula $F(\alpha)$ will never contain any occurrence of the operator †. Our analysis of this state of affairs is this: consider a formula

$$(A) \quad B(\alpha \mid \beta)$$

contrasted with

$$(B) \quad [\Uparrow \beta]B\alpha$$

What is the difference in meaning between these two formulas? We think it can be understood in terms of Lindström and Rabinowicz's distinction between point of *evaluation* and point of *reference*. Both (A) and (B) can be thought of as expressing that the formula α is believed after revision by β, but the formula α has different meaning in the two cases. In (A), the point of reference is held fixed, while in (B), the formula α is evaluated against a different point of reference than β. However, since the interpretation F takes a detour through the static fragment of the system S_{DEL}, in which no formulas of the form (B) occur, it makes sense that the operator $\dagger$ does not occur in the interpretation of any formulas: it has exactly the effect of changing the point of reference.

Thus, by extracting this insight from our interpretation of S_{DEL} in S_{2D}, we have managed to show how S_{2D} also fits into the picture we described earlier. The distinction between a perspective corresponding to the states of affairs before and after revision, respectively, is mirrored in S_{DEL} by the distinction between expressions of the form (A) and (B). Expressions of the first kind describe our revised beliefs about the state prior to revision, and expressions of the second kind describe our revised beliefs about the state of affairs after revision. Really, we do not have three different solutions; we have three different logical systems, each of which solves the problem with full DDL in one and the same way.

6 Discussion

We have argued that the systems S_{2D}, S_{Temp} and S_{DEL} all solve the problems of full DDL by distinguishing between two perspectives, expressed most explicitly in Lindström and Rabinowicz's two-dimensional approach. Given this, it is striking to find that Segerberg himself has suggested a two-dimensional approach to resolve another well-known paradox, namely Fitch's paradox (in a paper from 1994 with Rabinowicz [9]). Given the obvious similarities between Fitch's paradox and the paradoxes of full DDL, and given that Segerberg argued for a two-dimensional approach to the former, one would have expected him to embrace a two-dimensional approach to the latter as well. Thus it is surprising that Segerberg instead bases his solution

on Sorensen's notion of a blindspot, which is essentially unrelated to the two-dimensional approach.

In fact, it is not only surprising but, we think, it is questionable from a methodological point of view. Given the affinities between these paradoxes it would be desirable to treat them in a uniform fashion. Thus, for Segerberg, who is associated with two-dimensional semantics and the blindspot approach, the following uniform approaches suggests themselves:

(1) Treating both paradoxes as involving blindspots

(2) Treating both paradoxes in a two-dimensional semantics

By contrast, the following would seem less attractive from a systematic perspective:

(3) Treating Fitchs paradox in a two-dimensional semantics and Moores paradox as involving blindspots

(4) Treating Fitchs paradox as involving blindspots and Moores paradox in a two-dimensional semantics

And yet, as we saw, Segerbergs published responses to the paradoxes correspond to option (3), a suboptimal strategy from a systematic perspective. Finally, the result of the present article suggests that option (2) is, in a sense, considerably more plausible than meets the eye. More precisely, (2) is but a specific variant of a more general approach:

(2′) Treating both paradoxes as arising from failure to distinguish between different perspectives

As we have argued, two-dimensional DDL, Bonannos temporal system and van Benthem's DEL-style system are all instances of (2′). They all resolve the paradoxes by distinguishing between two different perspectives, in the two-dimensional case the point of reference and the point of evaluation, in Bonanno's case the time before and after revision, and in van Benthem's logic between conditional beliefs and beliefs after revision. Thus, the main competitor to the blindspot approach, as things must look from Segerberg's point of view, is more widely adopted, and thus has a stronger standing in the research community, than the apparent diversity could lead one to believe. Perhaps even more important is the fact that the main competitor -

the perspectival strategy - is a very natural way of dealing with the problems, or else researchers with widely different starting points would not have converged on it. Furthermore, to reconnect with our discussion of Segerberg's own solution, the perspectival strategy is perfectly compatible with the traditional view that we often revise simply by α rather than by $\alpha \wedge B\alpha$, and are quite rational in doing so. Not only does this accord better with our pre-theoretical conceptions of things (at least those of the present authors), but it means that this strategy works for reflective agents and unreflective agents alike. Unlike the perspectival strategy, Segerberg's solution is dependent on the assumption that the agent in question is reflective. Thus, unless an independent motivation is provided for not taking unreflective agents into consideration, the perspectival strategy stands out as the more general solution.

A Proofs of main results

A.1 Proof of Theorem 1

The proof is based on constructing models for S_{2D} out of models for S_{Temp}, in the following manner:

Definition 7. Given a temporal belief model $\mathfrak{A} = \langle W, \{B_n\}_{n\in\omega}, \{I_n\}_{n\in\omega}, V\rangle$, we define the two-dimensional revision model

$$\mathfrak{A}_{2D} = \langle W^*, \{B_u\}_{u\in W}, \{R^*_u\}_{u\in W}, V^*\rangle$$

as follows. We set

$$W^* = \{(u, n) : u \in W \ \& \ n \in \omega\}$$

For all $(u, n), (v, m), (w, k) \in W^*$, we set $(u, n)B_{(v,m)}(w, k)$ iff uB_nw and $k = m$. We set $(u, n)R^*_{(v,m)}(X)(w, k)$ iff

- $u = w$,
- $k = n + 1$ and
- $Z = I_k(u)$, where $Z = \{t \in W : (t, m) \in X\}$

Finally, we set $(u, n) \in V^*(p)$ iff $u \in V(p)$.

The construction is sound by the following proposition:

Proposition 2. *$\mathfrak{A}_{2D}$ is a two-dimensional revision model, for any temporal belief model $\mathfrak{A}$.*

Proof. We need to check that, for each $X \subseteq W^*$, if $(u,m)R^*_{(v,n)}(X)(w,k)$ then

1. $B_{(v,n)}(w,k) \subseteq X$
2. if $X \neq \emptyset$ then $B_{(v,n)}(w,k) \neq \emptyset$
3. if $B_{(v,n)}(u,m) \cap X \neq \emptyset$ then $B_{(v,n)}(w,k) = B_{(v,n)}(u,m) \cap X$

So suppose $(u,m)R^*_{(v,n)}(X)(w,k)$. Then $u = w$, $k = m+1$ and

$$I_{m+1}(u) = \{t \in W : (t,n) \in X\}$$

Now, since

$$B_{m+1}(u) \subseteq I_{m+1}(u)$$

item (1) follows easily by definition of the relation $B_{(v,n)}$: for if $(u,m+1)B_{v,n}(w',k')$, then $k' = n$ and $uB_{m+1}w'$, so $w' \in I_{m+1}(u)$, so $(w',n) = (w',k') \in X$.

For (2), note that $X \neq \emptyset$ implies $I_{m+1}(u) \neq \emptyset$, so $B_{m+1}(u) \neq \emptyset$. Pick w' such that $uB_{m+1}w'$. Then $(u,m+1)B_{v,n}(w',n)$ so $B_{(v,n)}(u,m+1) \neq \emptyset$.

Lastly, for (3), suppose $B_{(v,n)}(u,m) \cap X \neq \emptyset$. Let $(w',k') \in B_{(v,n)}(u,m) \cap X \neq \emptyset$; then $k' = n$ and uB_mw'. Since $(w',n) \in X$, $w' \in I_{m+1}(u)$. So

$$B_m(u) \cap I_{m+1}(u) \neq \emptyset$$

and hence

$$B_{m+1}(u) = B_m(u) \cap I_{m+1}(u)$$

This means that

$$B_{(v,n)}(u,m+1) = B_{(v,n)(u,m)} \cap X$$

To see this, suppose $(u,m+1)B_{(v,n)}(s,i)$. Then $i = n$, and $uB_{m+1}s$. But then uB_ms and $s \in I_{m+1}(u)$. So $(u,m)B_{(v,n)}(s,n)$ and $(s,n) \in X$.

Conversely, suppose $(u,m+1)B_{(v,n)}(s,i)$ and $(s,i) \in X$. By definition of $B_{v,n}$, $i = n$. So $(s,n) \in X$ and therefore $s \in I_{m+1}(u)$. Furthermore, $uB_{m+1}s$. So $s \in B_m(u) \cap I_{m+1}(u)$, hence $s \in B_{m+1}(u)$. By definition this means that $(u,m+1)B_{v,n}(s,n)$, i.e. $(u,m+1)B_{(v,n)}(s,i)$ as required. □

We now define a mapping G from pointed temporal belief models to pointed two-dimensional revision models by setting

$$G(\mathfrak{A}, u) =_{df.} (\mathfrak{A}_{2D}, (u, 0), (u, 0))$$

for each pointed temporal revision model $(\mathfrak{A}, u)$. We then have the following result, which gives the key to the soundness result for F:

Lemma 1. *For any pointed temporal model $(\mathfrak{A}, u)$ and any $\mathcal{L}_{2D}$-formula α, we have*

$$(\mathfrak{A}, u) \vDash F(\alpha) \Longleftrightarrow G(\mathfrak{A}, u) \vDash \alpha$$

Proof. We show, for any formula α, that for each world u in the universe of $\mathfrak{A}$, we have both

$$(1) \quad (\mathfrak{A}, u) \vDash \tau_{n,m}(\alpha) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D}, (v, n), (u, m)) \vDash \alpha$$

and

$$(2) \quad (\mathfrak{A}, u) \nvDash \tau_{n,m}(\alpha) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D}, (v, n), (u, m)) \nvDash \alpha$$

for all $v \in W$. From (1) and (2) together it follows that

$$(\mathfrak{A}, u) \vDash \tau_{0,0}(\alpha) \Longleftrightarrow (\mathfrak{A}_{2D}, (u, 0), (u, 0)) \vDash \alpha$$

i.e.

$$(\mathfrak{A}, u) \vDash F(\alpha) \Longleftrightarrow G(\mathfrak{A}, u) \vDash \alpha$$

as desired.

The proof goes by induction on the length of α. For propositional variables, both clauses are immediate, and the steps for Boolean connectives are easy.

Step for B: Suppose $(\mathfrak{A}, u) \vDash \tau_{n,m}(B\alpha)$, i.e. $(\mathfrak{A}, u) \vDash B_m\tau_{n,n}(\alpha)$. Let $v \in W$ and let (w, k) be such that $(u, m)B_{v,n}(w, k)$. Then by definition uB_mw and $k = n$, so we must have $(\mathfrak{A}, w) \vDash \tau_{n,n}(\alpha)$ and by clause (1) of the IH we get $(\mathfrak{A}_{2D}, (v, n), (w, n)) \vDash \alpha$. So we must have $(\mathfrak{A}_{2D}, (v, n), (u, m)) \vDash B\alpha$. This shows that

$$(1) \quad (\mathfrak{A}, u) \vDash \tau_{n,m}(B\alpha) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D}, (v, n), (u, m)) \vDash B\alpha$$

Suppose that $(\mathfrak{A}, u) \nvDash \tau_{n,m}(B\alpha)$, i.e. $(\mathfrak{A}, u) \nvDash B_m\tau_{n,n}(\alpha)$. Then there exists $w \in W$ such that uB_mw and $(\mathfrak{A}, v) \nvDash \tau_{n,n}\alpha$. Let $v \in W$; then we have

$(u,m)B_{(v,n)}(w,n)$ and by clause (2) of IH we have $(\mathfrak{A}_{2D},(v,n),(w,n)) \not\models \alpha$, hence $(\mathfrak{A}_{2D},(v,n),(u,m)) \not\models B\alpha$. We have shown that

$$(2) \quad (\mathfrak{A},u) \not\models \tau_{n,m}(B\alpha) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D},(v,n),(u,m)) \not\models B\alpha$$

as required.

Step for $*$: Suppose $(\mathfrak{A},u) \models \tau_{n,m}([*\alpha]\beta)$, i.e.

$$(\mathfrak{A},u) \models I_{m+1}\tau_{n,n}(\alpha) \to \tau_{n,m+1}(\beta)$$

We note that by the IH we have, for each $v \in W$,

$$(\ddagger) \quad \|\tau_{n,n}(\alpha)\|_{\mathfrak{A}} = \{t \in W : (t,n) \in \|\alpha\|_{(v,n)}\}$$

Suppose for $v \in W$ that $(u,m)R^*_{(v,n)}(\|\alpha\|_{(v,n)})(w,k)$. Then $k = m+1$. Furthermore, by definition and by $(\ddagger)$ we get

$$I_{m+1}(u) = \{t \in W : (t,n) \in \|\alpha\|_{(v,n)}\} = \|\tau_{n,n}(\alpha)\|_{\mathfrak{A}}$$

So $(\mathfrak{A},u) \models I_{m+1}(\tau_{n,n}(\alpha))$. Thus, we get $(\mathfrak{A},u) \models \tau_{n,m+1}(\beta)$. By clause (1) of the IH, this gives $(\mathfrak{A}_{2D},(v,n),(w,m+1)) \models \beta$, i.e. $(\mathfrak{A}_{2D},(v,n),(w,k)) \models \beta$. So $(\mathfrak{A}_{2D},(v,n),(u,m)) \models [*\alpha]\beta$. We have thus shown

$$(1) \quad (\mathfrak{A},u) \models \tau_{n,m}([*\alpha]\beta) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D},(v,n),(u,m)) \models [*\alpha]\beta$$

Suppose $(\mathfrak{A},u) \not\models \tau_{n,m}([*\alpha]\beta)$, i.e. $(\mathfrak{A},u) \models I_{m+1}\tau_{n,n}(\alpha)$ but $(\mathfrak{A},u) \not\models \tau_{n,m+1}(\beta)$. Pick $v \in W$. Using $(\ddagger)$ we obtain

$$I_{m+1}(u) = \|\tau_{n,n}(\alpha)\|_{\mathfrak{A}} = \{t \in W : (t,n) \in \|\alpha\|_{(v,n)}\}$$

From this we can conclude that $(u,m)R^*_{(v,n)}(\|\alpha\|_{(v,n)})(u,m+1)$. Furthermore, by clause (2) of the IH we have $(\mathfrak{A}_{2D},(v,n),(u,m+1) \not\models \beta)$, so $(\mathfrak{A}_{2D},(v,n),(u,m) \not\models [*\alpha]\beta)$. We have thus shown

$$(2) \quad (\mathfrak{A},u) \not\models \tau_{n,m}([*\alpha]\beta) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D},(v,n),(u,m)) \not\models [*\alpha]\beta$$

as required.

Step for $\dagger$: Given that the IH holds for α, suppose first that we have $(\mathfrak{A},u) \models \tau_{n,m}(\dagger\,\alpha)$, i.e. $(\mathfrak{A},u) \models \tau_{m,m}(\alpha)$. Then we have by clause (1) of IH: for all $v \in W$, $(\mathfrak{A}_{2D},(v,m),(u,m)) \models \alpha$. In particular, $(\mathfrak{A}_{2D},(u,m),(u,m)) \models$

α. This means that, for all $v \in W$, $(\mathfrak{A}_{2D}, (v,n), (u,m)) \vDash \dagger\ \alpha$. We have established:

$$(1) \quad (\mathfrak{A}, u) \vDash \tau_{n,m}(\dagger\ \alpha) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D}, (v,n), (u,m)) \vDash \dagger\ \alpha$$

On the other hand, suppose $(\mathfrak{A}, u) \nvDash \tau_{n,m} \dagger\ \alpha$, i.e. $(\mathfrak{A}, u) \nvDash \tau_{m,m}(\alpha)$. Then we have by clause (2) of IH: for all $v \in W$, $(\mathfrak{A}_{2D}, (v,m), (u,m)) \nvDash \alpha$. In particular, $(\mathfrak{A}_{2D}, (u,m), (u,m)) \nvDash \alpha$. This means that, for all $v \in W$, $(\mathfrak{A}_{2D}, (v,n), (u,m)) \nvDash \dagger\ \alpha$. We have established:

$$(2) \quad (\mathfrak{A}, u) \nvDash \tau_{n,m}(\dagger\ \alpha) \Longrightarrow \forall v \in W : (\mathfrak{A}_{2D}, (v,n), (u,m)) \nvDash \dagger\ \alpha$$

This ends the proof. □

We now prove Theorem 1 as follows: suppose $F(\Gamma) \nvDash_{Temp} F(\alpha)$. Then there is a pointed temporal belief model $(\mathfrak{A}, u)$ such that $(\mathfrak{A}, u) \vDash F(\Gamma)$ but $(\mathfrak{A}, u) \nvDash F(\alpha)$. By the previous theorem, $G(\mathfrak{A}, u) \vDash \Gamma$ but $G(\mathfrak{A}, u) \nvDash \alpha$. Hence $\Gamma \nvDash_{2D} \alpha$. This ends the proof of the theorem.

A.2 Proof of Theorem 2

We use the same strategy as in the previous section:

Definition 8. Given a two-dimensional model $\mathfrak{A}$ and a world v in the universe of $\mathfrak{A}$, we define the conditional belief model

$$\mathfrak{A}_{DEL}[v] = \langle W^*, \{\sigma_u\}_{u \in W^*}, V^* \rangle$$

as follows: we set $W^* = W$ and $V^* = V$. For each $u \in W$ and $X \subseteq W$, we set

$$\sigma_u(X) = \{w \in W : \exists p \in W[uR^*_v(X)p \text{ and } pB_v w]\}$$

It is easily checked that $\mathfrak{A}_{DEL}[v]$ is a conditional belief model. We define the mapping G from pointed two-dimensional revision models to pointed conditional belief models by setting $G(\mathfrak{A}, v, u) = (\mathfrak{A}_{DEL}[v], u)$ for a pointed two-dimensional revision model $(\mathfrak{A}, v, u)$. We have the following result:

Lemma 2. *For any pointed two-dimensional model $(\mathfrak{A}, u, v)$ and any <u>static</u> $\mathcal{L}_{DEL}$-formula α we have*

$$(\mathfrak{A}, u, v) \vDash \tau(\alpha) \Longleftrightarrow G(\mathfrak{A}, u, v) \vDash \alpha$$

Proof. By induction over the length of static formulas we show that, for all $v \in W$ we have

$$(\mathfrak{A}, u, v) \vDash \tau(\alpha) \Longleftrightarrow (\mathfrak{A}_{DEL}[u], v) \vDash \alpha$$

The steps for atomic formulas and Boolean connectives are trivial.

Step for A**:** suppose $(\mathfrak{A}, u, v) \vDash \tau(A\alpha)$, i.e. $(\mathfrak{A}, u, v) \vDash [*\neg\tau(\alpha)]B\bot$. By seriality of $R^*_u(\|\neg\tau(\alpha)\|^{\mathfrak{A}}_u)$ there must be some w such that $vR^*_u(\|\neg\tau(\alpha)\|^{\mathfrak{A}}_u)w$. Furthermore, clearly we have $B_u(w) = \emptyset$, and this means that $\|\neg\tau(\alpha)\|^{\mathfrak{A}}_u = \emptyset$. Hence $\|\tau(\alpha)\|^{\mathfrak{A}}_u = W = W^*$. By the IH, $\|\alpha\|_{\mathfrak{A}_{DEL}[u]} = W^*$, and so we have $(\mathfrak{A}_{DEL}[u], v) \vDash A\alpha$ as required.

Conversely, suppose $(\mathfrak{A}, u, v) \nvDash \tau(A\alpha)$, i.e. $(\mathfrak{A}, u, v) \nvDash [*\neg\tau(\alpha)]B\bot$. Then there is some w such that $vR^*_u(\|\neg\tau(\alpha)\|^{\mathfrak{A}}_u)w$ and $B_u(w) \neq \emptyset$. Hence there is some s such that wB_us. By the definition of a two-dimensional model, $s \in \|\neg\tau(\alpha)\|$ i.e. $(\mathfrak{A}, u, s) \vDash \neg\tau(\alpha)$. Hence $(\mathfrak{A}, u, s) \nvDash \tau(\alpha)$, and by the IH $(\mathfrak{A}_{DEL}[u], s) \nvDash \alpha$. Hence $(\mathfrak{A}_{DEL}[u], v) \nvDash A\alpha$ as required.

Step for B**:** suppose $(\mathfrak{A}, u, v) \vDash \tau(B(\alpha \mid \beta))$, i.e.

$$(\mathfrak{A}, u, v) \vDash [*\tau(\beta)]B\tau(\alpha)$$

Suppose $w \in \sigma_v(\|\beta\|_{\mathfrak{A}_{DEL}[u]})$. By the IH this means that $w \in \sigma_v(\|\tau(\beta)\|^{\mathfrak{A}}_u)$, so there is some s such that $vR^*_u(\|\alpha\|^{\mathfrak{A}}_u)s$ and sB_uw. Since $(\mathfrak{A}, u, v) \vDash [*\tau(\beta)]B\tau(\alpha)$ we have $(\mathfrak{A}, u, s) \vDash B\tau(\alpha)$ so $(\mathfrak{A}, u, w) \vDash \tau(\alpha)$. By IH we get $(\mathfrak{A}_{DEL}[u], w) \vDash \alpha$. We have thus shown that $(\mathfrak{A}, u, v) \vDash B(\alpha \mid \beta)$ as required.

Conversely, suppose that $(\mathfrak{A}, u, v) \nvDash \tau(B(\alpha \mid \beta))$, i.e.

$$(\mathfrak{A}, u, v) \nvDash [*\tau(\beta)]B\tau(\alpha)$$

Then there is some s such that $vR^*_u(\|\tau(\beta)\|^{\mathfrak{A}}_u)s$ and $(\mathfrak{A}, u, s) \nvDash B\tau(\alpha)$. This means that for some w we have sB_uw and $(\mathfrak{A}, u, w) \nvDash \tau(\alpha)$. By the IH we have $vR^*_u(\|\beta\|_{\mathfrak{A}_{DEL}[u]})s$, and thus we have $w \in \sigma_v(\|\beta\|_{\mathfrak{A}_{DEL}[u]})$. Furthermore, by the IH again, we have $(\mathfrak{A}_{DEL}[u], w) \nvDash \alpha$. Thus $(\mathfrak{A}_{DEL}[u], v) \nvDash B(\alpha \mid \beta)$ as required. □

Using the fundamental property of the translation ρ used in the construction of F, this lemma immediately entails:

Corollary 2. *For any pointed two-dimensional model* $(\mathfrak{A}, u, v)$ *and any* $\mathcal{L}_{DEL}$*-formula* α *we have*

$$(\mathfrak{A}, u, v) \vDash F(\alpha) \Longleftrightarrow G(\mathfrak{A}, u, v) \vDash \alpha$$

From this result, we can prove Theorem 2 just like we proved Theorem 1.

References

[1] Alchourrón, C.E., Gärdenfors,P., Makinson, D.: On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. J. Symbolic Logic **50**, 510–530 (1985)

[2] Blackburn,P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press (2001)

[3] Bonanno, G.: A simple modal logic for belief revision, Synthese **147**, 193–228 (2005)

[4] Gärdenfors, P.: Knowledge in Flux: Modeling the dynamics of epistemic states. Cambridge, Massachusetts: MIT Press (1988)

[5] Hansson, S-O.: A Textbook of Belief Dynamics: Theory Change and Database Updating. Kluwer Academic Publishers (1999)

[6] Harel,D., Kozen, D., Tiuryn, J.: Dynamic Logic. The MIT Press (2000)

[7] Hintikka, J.: Knowledge and Belief: an Introduction to the Logic of the Two Notions. Cornell University Press (1962)

[8] Lindström, S., Rabinowicz, W.: DDL Unlimited: Dynamic Doxastic Logic for Introspective Agents. Erkenntnis **50**, 353–385 (1999)

[9] Rabinowicz, W., Segerberg, K.: Actual truth, possible knowledge. Topoi **13**, 101–115 (1994)

[10] Segerberg, K.: Irrevocable belief revision in dynamic doxastic logic. Notre Dame J. Formal Logic **39**, 287–306 (1998)

[11] Segerberg, K.: Moore problems in full dynamic doxastic logic. In Malinowski, J., Pietruszczak, A. (eds.), Poznan Studies in the Philosophy of the Sciences and the Humanities, Essays in Logic and Ontology, pp.95-110. Rodopi (2006)

[12] Segerberg, K., Leitgeb, H.: Dynamic doxastic logic - why, how and where to? Synthese **155**, 167–190 (2007)

[13] Sorensen, R.A.: Blindspots. Oxford: Clarendon Press (1988)

[14] van Benthem, J.: Dynamic logic for belief revision. J. Appl. Non-Classical Logics **17**, 129–155 (2007)

[15] van Ditmarsch, H.P., van der Hoek, P., Kooi, B.P.: Dynamic Epistemic Logic. Dordrecht: Springer (2005)