

---

# Group Belief Dynamics under Iterated Revision: Fixed Points and Cycles of Joint Upgrades

---

**Alexandru Baltag**  
Computing Laboratory  
Oxford University  
Oxford, UK

**Sonja Smets\***  
Dept. of Artificial Intelligence & Dept. of Philosophy  
University of Groningen  
Groningen, The Netherlands

## 1 Introduction

What happens if in the Muddy Children story [22] we drop the assumption that the public announcements (made by the father and by the children) are commonly known to be always true, and instead we simply assume that they *are true and commonly believed to be true*? More generally, what happens *in the long term* with a group's beliefs, knowledge and "epistemic states" (fully describable in fact by conditional beliefs), when receiving (or exchanging) a *sequence of public announcements of truthful but uncertain information*? Do the agents' beliefs (or knowledge, or conditional beliefs, or other doxastic attitudes such as "strong beliefs") reach a *fixed point*? Or do they exhibit instead a *cyclic behavior*, oscillating forever?

These questions are of obvious importance for Belief Revision theory, Learning theory and Social Choice theory, and may have some relevance to Game Theory as well. In fact, some of these questions came to our attention due to a recent talk by J. van Benthem (Chennai, January 2009), in which he was refining his previous "dynamic" analysis [14] of backward induction solution in perfect information games. This extended abstract provides some partial answers to some of the questions above, as well as a convenient setting for investigating further the ones that are still open.

## 2 Beliefs and Joint Upgrades in Plausibility Models

In this section, we review some basic notions and results from [5]. We use finite "plausibility" frames, in the sense of our papers [5, 7, 6, 8, 10, 9]. These kind of semantic structures are the *natural multi-agent generalizations of structures that are standard*, in one form or another, in Belief Revision: Halpern's "preferential models" [24], Spohn's ordinal-ranked models [28], Board's "belief-revision structures" [18], Grove's "sphere" models [23]. Unlike the set-

tings in [10, 9], we restrict here to the *finite* case, for at least three reasons. First, since in the finite case all the above-mentioned semantic approaches are equivalent, this assumption gives a certain "semantic *robustness*" to the results in this paper: they are independent on which of the above semantic approaches is preferred. Secondly, it seems *realistic* to assume that an agent starts with a finite belief base, consisting of finitely many sentences that are believed on their own merits (even if, by logical omniscience, the agent also believes all their infinitely many consequences); but, since all the languages considered here have the finite model property, any such finite base can be represented in a finite plausibility model. Thirdly, our questions in the Introduction make sense primarily for finite structures: *only for finite models it is reasonable at all to expect that iterated announcements might reach a fixed point after finitely many iterations*.

For a given set  $\mathcal{A}$  of labels called "agents", a (*finite, multi-agent*) *plausibility frame* is a structure  $\mathbf{S} = (S, \geq_a)_{a \in \mathcal{A}}$ , consisting of a *finite* set  $S$  of "states", together with a family of *locally connected preorders*  $\geq_a \subseteq S \times S$ , labeled by agents. Here, a "locally connected preorder"  $\geq \subseteq S \times S$  is a reflexive and transitive relation such that: if  $s \geq t$  and  $s \geq w$  then either  $t \geq w$  or  $w \geq t$ ; and if  $t \geq s$  and  $w \geq s$  then either  $t \geq w$  or  $w \geq t$ . See [5] for a justification and motivation for these conditions.<sup>1</sup> We use the notation  $s \sim_a t$  for the *comparability relation* with respect to  $\geq_a$  (i.e.  $s \sim_a t$  iff either  $s \geq_a t$  or  $t \geq_a s$ ),  $s >_a t$  for the corresponding *strict order relation* (i.e.  $s >_a t$  iff  $s \geq_a t$  but  $t \not\geq_a s$ ), and  $s \cong_a t$  for the corresponding *indifference relation* (i.e.  $s \cong_a t$  iff both  $s \geq_a t$  and  $t \geq_a s$ ). In any plausibility frame, the comparability relations  $\sim_a$  are *equivalence relations*, and so they induce *partitions*. We denote by  $s(a) := \{t \in S : s \sim_a t\}$  the  $\sim_a$ -*partition cell* of  $s$ .

**(Pointed) Plausibility Models** A (*finite, multi-agent*) *pointed plausibility model* (or "model", for short) is a structure  $\mathbf{S} = (S, \geq_a, \|\cdot\|, s_0)_{a \in \mathcal{A}}$ , consisting of a plau-

---

\*Research Associate of IEG, Research Group on the Philosophy of Information, Oxford University UK

---

<sup>1</sup>In the infinite case, one has to add a well-foundedness condition, obtaining "locally well-preordered" relations.

sibility frame  $(S, \geq_a)_{a \in \mathcal{A}}$  together with a *valuation map*  $\|\cdot\| : \Phi \rightarrow \mathcal{P}(S)$ , mapping every element  $p$  of some given set  $\Phi$  of “atomic sentences” into a set of states  $\|p\| \subseteq S$ , and together with a *designated state*  $s_0 \in S$ , called the “actual state”.

**(Common) Knowledge and (Conditional) Belief** Given a plausibility model  $\mathbf{S}$ , sets  $P, Q \subseteq S$  of states, an agent  $a \in \mathcal{A}$  and some group  $G \subseteq \mathcal{A}$ , we define:  $best_a P := \text{Min}_{\geq_a} P := \{s \in P : s' \geq_a s \text{ for all } s' \in P\}$ ,  $K_a P := \{s \in S : s(a) \subseteq P\}$ ,  $B_a P := \{s \in S : best_a s(a) \subseteq P\}$ ,  $B_a^Q P := \{s \in S : best_a (s(a) \cap Q) \subseteq P\}$ ,  $Ek_G P := \bigcap_{a \in G} K_a P$ ,  $Eb_G P := \bigcap_{a \in G} B_a P$ ,  $Ck_G P := \bigcap_{n \in \mathbb{N}} Ek_G^n P$  (where  $Ek_G^0 P := P$  and  $Ek_G^{n+1} := Ek_G(Ek_G^n P)$ ),  $Eb P := Eb_{\mathcal{A}} P$ , and  $Ck P := Ck_{\mathcal{A}} P$ .

**Interpretation.** The elements of  $S$  represent the *possible states*, or “possible worlds”, of a system: *possible descriptions of the real world*. The *correct* description of the real world is given by the “actual state”  $s_0$ . The atomic sentences  $p \in \Phi$  represent “*ontic*” (*non-doxastic*) *facts*, that might hold or not in a given state. The valuation tells us which facts hold at which worlds. For each agent  $a$ , the equivalence relation  $\sim_a$  represents the agent  $a$ ’s *epistemic indistinguishability relation*, inducing  $a$ ’s *information partition*;  $s(a)$  is the state  $s$ ’s *information cell* with respect to  $a$ ’s partition: if  $s$  were the real state, then agent  $a$  would consider all the states  $t \in s(a)$  as “epistemically possible”.  $K_a P$  is the proposition “agent  $a$  knows  $P$ ”: observe that this is indeed the same as Aumann’s partition-based definition of knowledge. The plausibility relation  $\geq_a$  is agent  $a$ ’s *plausibility order* between her “epistemically possible” states: we read  $s \geq_a t$  as “agent  $a$  considers  $t$  at least as plausible as  $s$  (though the two are *epistemically indistinguishable*)”. This is meant to capture the agent’s (*conditional*) *beliefs about the state* of the system. Note that  $s \geq_a t$  implies  $s \sim_a t$ , so that the agent only compares the plausibility of states that are epistemically indistinguishable: so *we are not concerned here with counterfactual beliefs* (going against the agent’s knowledge), *but only with conditional beliefs* (if given new evidence that must be compatible with prior knowledge). So  $B_a^Q P$  is read “agent  $a$  believes  $P$  conditional on  $Q$ ” and means that, if  $a$  would receive some further (certain) information  $Q$  (to be added to what she already knows) then she would believe that  $P$  was the case. So conditional beliefs  $B_a^Q$  give descriptions of the agent’s *plan* (or *commitments*) about what would she believe (about the current state) if she would learn some new information  $Q$ . To quote J. van Benthem in [13], conditional beliefs are “*static pre-encodings*” of the agent’s *potential belief changes* in the face of new information. The above definition says that  $B_a^Q P$  holds iff  $P$  holds in all the “best” (i.e. the most plausible)  $Q$ -states (that are consistent with  $a$ ’s knowledge). In particular, a *simple (non-conditional) belief*  $B_a P$  holds iff  $P$  holds in all the best

states that are epistemically possible for  $a$ .

**“Strong Belief”** Another important doxastic attitude, called *strong belief*, is given by:

$$Sb_a P = \{s \in P : s(a) \cap P \neq \emptyset \text{ and } w \succ_a t \text{ for all } t \in s(a) \cap P \text{ and all } w \in s(a) \setminus P\}.$$

So  $P$  is strong belief at a state  $s$  iff  $P$  is *epistemically possible* and moreover *all epistemically possible  $P$ -states at  $s$  are more plausible than all epistemically possible non- $P$  states*. This notion corresponds to the “strong belief” of Battigalli and Siniscalchi [11], and to the “robust belief” of Stalnaker [29]. As for belief and knowledge, we define “everybody in group  $G$  strongly believes” by  $ESb_G P := \bigcap_{a \in G} Sb_a P$ .

**Doxastic Propositions** As long as the model  $\mathbf{S}$  is kept fixed, we can identify (as we did above) a “proposition” with any set  $P \subseteq S$  of states in  $\mathbf{S}$ . But, since later we will proceed to study systematic *changes* of models, we need a notion of proposition that can be interpreted in *any* model. A *doxastic proposition* is a map  $\mathbf{P}$  assigning to each plausibility model  $\mathbf{S}$  some set  $\mathbf{P}_{\mathbf{S}} \subseteq S$  of states in  $S$ . We write  $s \models_{\mathbf{S}} \mathbf{P}$ , and say that the proposition  $\mathbf{P}$  is *true at state  $s \in S$  in the model  $\mathbf{S}$* , iff  $s \in \mathbf{P}_{\mathbf{S}}$ . We say that a doxastic proposition  $\mathbf{P}$  is *true at (the pointed model)  $\mathbf{S}$* , and write  $\mathbf{S} \models \mathbf{P}$  if it is true at the “actual state”  $s_0$  in the model  $\mathbf{S}$ , i.e. if  $s_0 \models_{\mathbf{S}} \mathbf{P}$ . In particular, we have the “always true”  $\top$  and “always false”  $\perp$  propositions  $\perp_{\mathbf{S}} := \emptyset$ ,  $\top_{\mathbf{S}} := S$ . Also, for each atomic sentence  $p$ , there exists a corresponding doxastic proposition  $\mathbf{p}$ , given by  $\mathbf{p}_{\mathbf{S}} = \|p\|_{\mathbf{S}}$ . All the operations on sets can be similarly “*lifted*” *pointwise* to propositions: negation  $(\neg \mathbf{P})_{\mathbf{S}} := S \setminus \mathbf{P}_{\mathbf{S}}$ , conjunction  $(\mathbf{P} \wedge \mathbf{R})_{\mathbf{S}} := \mathbf{P}_{\mathbf{S}} \cap \mathbf{R}_{\mathbf{S}}$  etc, the “*best*” operator  $(best_a \mathbf{P})_{\mathbf{S}} := best_a \mathbf{P}_{\mathbf{S}}$ , all the *modalities*  $(K_a \mathbf{P})_{\mathbf{S}} := K_a \mathbf{P}_{\mathbf{S}}$ ,  $(B_a \mathbf{P})_{\mathbf{S}} := B_a \mathbf{P}_{\mathbf{S}}$ ,  $(B_a^Q \mathbf{P})_{\mathbf{S}} := B_a^{Q_{\mathbf{S}}} \mathbf{P}_{\mathbf{S}}$  etc.

**Characterization of strong belief** It is easy to see that we have the following equivalence:

$$\mathbf{S} \models Sb_a \mathbf{P} \text{ iff:}$$

$$\mathbf{S} \models B_a \mathbf{P} \text{ and } \mathbf{S} \models B_a^Q \mathbf{P}$$

$$\text{for every } \mathbf{Q} \text{ such that } \mathbf{S} \models \neg K_a (\mathbf{Q} \rightarrow \neg \mathbf{P}).$$

In other words: *something is strong belief iff it is believed and if this belief can only be defeated by evidence (truthful or not) that is known to contradict it*. An example is the “presumption of innocence” in a trial: requiring the members of the jury to hold the accused as “innocent until proven guilty” means asking them to start the trial with a “strong belief” in innocence.

**Example 1** Consider a pollster (Charles) with the following beliefs about how a voter (Alice) will vote:



In the representation, the arrows represent the plausibility relations  $\geq_a$  and  $\geq_c$  for Alice and Charles, but we skip the loops and the arrows that can be obtained by transitivity. Since Alice *knows* how she votes, there is no uncertainty for her, hence no plausibility arrows (except for the loops). There are three possible worlds  $s$  (in which Alice votes Republican),  $w$  (in which she votes Democrat) and  $t$  (in which she doesn't vote). We assume there are no other options: i.e. there are no other candidates and it is impossible to vote for both candidates. The atomic sentences are  $r$  (for “voting Republican”) and  $d$  (for “voting Democrat”). The valuation is given in the diagram:  $\|r\| = \{s\}$ ,  $\|d\| = \{w\}$ . We assume the real world is  $s$ , so Alice will vote Republican ( $r$ )! But Charles believes that she will vote Democrat ( $d$ ); and in case this turns out wrong, he'd rather believe that she won't vote ( $\neg d \wedge \neg r$ ) than accepting that she may vote Republican: so world  $t$  is more plausible than  $s$ .

**Modal Languages** One can consider a number of modal languages for the above models. See e.g. [5] for details on a number of logics, including one that can *define strong beliefs* (in terms of a “safe belief” operator). For each language  $L$ , the semantics is given by an interpretation map, which assigns to each sentence  $\varphi$  some doxastic proposition  $\|\varphi\|$ . In this paper, we only consider the language of *doxastic-epistemic logic* (having only operators for belief  $B_a\varphi$ , knowledge  $K_a\varphi$  and common knowledge  $Ck\varphi$ , with the obvious compositional semantic clauses) and its extension with *conditional belief* operators  $B_a^\varphi\psi$ .

**G-Bisimulation** For a group  $G \subseteq \mathcal{A}$  of agents, we say the pointed models  $\mathbf{S} = (S, \geq_a, \|\cdot\|, s_0)_{a \in \mathcal{A}}$  and  $\mathbf{S}' = (S', \geq'_a, \|\cdot\|', s'_0)_{a \in \mathcal{A}}$  are *G-bisimilar*, and write  $\mathbf{S} \simeq_G \mathbf{S}'$ , if the pointed Kripke models  $(S, \geq_a, \|\cdot\|, s_0)_{a \in G}$  and  $(S', \geq'_a, \|\cdot\|', s'_0)_{a \in G}$  (having as accessibility relations only the  $G$ -labeled relations) are bisimilar in the usual sense from Modal Logic [17]. When  $G = \mathcal{A}$ , we simply write  $\mathbf{S} \simeq \mathbf{S}'$ , and say  $\mathbf{S}$  and  $\mathbf{S}'$  are *bisimilar*. Bisimilar models *differ only formally*: they encode precisely the same doxastic-epistemic information, and they satisfy the same modal sentences.

### Types of Public Announcements: “Joint Upgrades”

We move on now to *dynamics*: what happens when some proposition  $\mathbf{P}$  is *publicly announced*? According to Dynamic Epistemic Logic, this induces a *change of model*: a “model transformer”. However, the specific change depends on *the agents' attitudes to the plausibility* of the announcement: *how certain is the new information*? Three main possibilities have been discussed in the literature: (1) the announcement  $\mathbf{P}$  is *certainly true*: it is *common knowledge* that the speaker tells the truth; (2) the announcement is *strongly believed* to be true by everybody: it is *common knowledge that everybody strongly believes* that the speaker tells the truth; (3) the announcement is *(simply) believed*: it is *common knowledge that everybody believes (in the sim-*

*ple, “weak” sense)* that the speaker tells the truth. These three alternatives correspond to three forms of “learning” a public announcement, forms discussed in [13, 15] in a Dynamic Epistemic Logic context: “update”<sup>2</sup>  $\! \uparrow \mathbf{P}$ , “*radical upgrade*”  $\! \uparrow \mathbf{P}$  and “*conservative upgrade*”  $\! \uparrow \mathbf{P}$ . Under various names, they have been previously proposed in the literature on Belief Revision, e.g. by Rott [27], and in the literature on dynamic semantics for natural language by Veltman [30].

We will use “*joint upgrades*” as a general term for all these three model transformers, and denote them in general by  $\! \uparrow \mathbf{P}$ , where  $\! \uparrow \in \{!, \uparrow, \uparrow\}$ . Formally, each of our joint upgrades is a (possibly partial) function taking as inputs pointed models  $\mathbf{S} = (S, \geq_a, \|\cdot\|, s_0)$  and returning new (“upgraded”) pointed models  $\! \uparrow \mathbf{P}(\mathbf{S}) = (S', \geq'_a, \|\cdot\|', s'_0)$ , with  $S' \subseteq S$ . Since upgrades are purely doxastic, they *won't affect the real world or the “ontic facts” of each world*: i.e. they all satisfy  $s'_0 = s_0$  and  $\|p\|' = \|p\| \cap S'$ , for atomic  $p$ . So, in order to completely describe a given upgrade, we only have to specify (a) *its possible inputs*  $\mathbf{S}$ , (b) *the new set of states*  $S'$ ; (c) *the new relations*  $\geq'_a$ .

**(1) Learning Certain information: Joint “Update”.** The update  $\! \uparrow \mathbf{P}$  is an operation on pointed models which is *executable* (on a pointed model  $\mathbf{S}$ ) *iff*  $\mathbf{P}$  is *true* (at  $\mathbf{S}$ ) and which *deletes all the non- $\mathbf{P}$ -worlds from the pointed model, leaving everything else the same*. Formally, an update  $\! \uparrow \mathbf{P}$  is an upgrade such that: (a) it takes as inputs only pointed models  $\mathbf{S}$ , such that  $\mathbf{S} \models \mathbf{P}$ ; (b) the new set of states is  $S' = \mathbf{P}_\mathbf{S}$ ; (c)  $s \geq'_a t$  iff  $s \geq_a t$  and  $s, t \in S'$ .

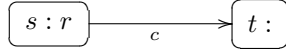
**(2) Learning from a Strongly Trusted Source: (Joint) “Radical” Upgrade.** The “*radical upgrade*” (or “*lexicographic upgrade*”)  $\! \uparrow \mathbf{P}$ , as an operation on pointed plausibility models, can be described as “*promoting*” *all the  $\mathbf{P}$ -worlds within each information cell so that they become “better” (more plausible) than all  $\neg \mathbf{P}$ -worlds in the same information cell, while keeping everything else the same*: the valuation, the actual world and the relative ordering between worlds within either of the two zones ( $\mathbf{P}$  and  $\neg \mathbf{P}$ ) stay the same. Formally, a radical upgrade  $\! \uparrow \mathbf{P}$  is (a) a *total upgrade* (taking as input *any* model  $\mathbf{S}$ ), such that (b)  $S' = S$ , and (c):  $s \geq'_a t$  iff either  $s \notin \mathbf{P}_\mathbf{S}$  and  $t \in s(a) \cap \mathbf{P}_\mathbf{S}$ , or  $s \geq_a t$ .

**(3) “Barely believing” what you hear: (Joint) “Conservative” Upgrade.** The so-called “*conservative upgrade*”  $\! \uparrow \mathbf{P}$  (called “*minimal conditional revision*” by Boutilier [19]) performs in a sense the minimal possible revision of a model that is forced by believing the new information  $\mathbf{P}$ .

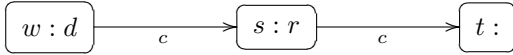
<sup>2</sup>Note that in Belief Revision, the term “belief update” is used for a totally different operation (the Katzuno-Mendelzon update[25]), while what we call “update” is known as “conditioning”. We choose to follow here the terminology used in Dynamic Epistemic Logic, but we want to warn the reader against any possible confusions with the KM update.

As an operation on pointed models, it can be described as “promoting” only the “best” (most plausible)  $\mathbf{P}$ -worlds, so that they become the most plausible in their information cell, while keeping everything else the same. Formally,  $\uparrow \mathbf{P}$  is (a) a total upgrade, such that (b)  $S' = S$ , and (c):  $s \geq'_a t$  iff either  $t \in \text{best}_a(s(a) \cap \mathbf{P}_S)$  or  $s \geq_a t$ .

**Examples:** As an example of update, consider the case in which an absolutely infallible authority (a truth-telling “Oracle”) publicly announces that Alice will definitely not vote Democrat: this is the update  $\uparrow(\neg d)$ , whose result is the updated model

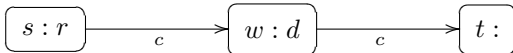


In the same initial situation (from Example 1), consider the case in which there is no Oracle, but instead a trusted, but fallible source publicly announces that Alice won’t vote Democrat. The agents have a *strong trust* in the veracity of any information coming from this source, but they *cannot exclude* the possibility that *it might still be a lie*: so this is not an update! Instead, it is a *radical upgrade*  $\uparrow(\neg d)$ , whose result is:



Now, Charles believes that Alice will not vote at all; but even if he later would learn this was not the case, he’d still keep his newly acquired belief that Alice doesn’t vote Democrat (so in this eventuality he’d conclude, correctly, that she votes Republican).

Contrast this situation with the case in which the agents “barely trust” what they hear; e.g. they just hear a rumor that Alice will not vote Democrat. They *believe* the rumor, but *not strongly*: in case they later are forced to revise their beliefs, they’d immediately give up the belief in the inherent veracity of the rumor. In this case, we interpret the learning event as a *conservative upgrade*  $\uparrow(\neg d)$  of the original model in Example 1, upgrade whose result is:



Now, Charles also believes Alice will not vote, but in case he would later learn this was not the case, he’d dismiss the rumor and revert to his older belief that Alice votes Democrat.

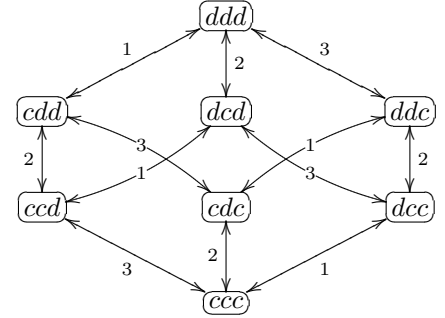
**Truthfulness** A joint upgrade  $\uparrow \mathbf{P}$  is *truthful* in a pointed model  $\mathbf{S}$  if  $\mathbf{P}$  is true at  $\mathbf{S}$ .

### 3 Iterated Upgrades

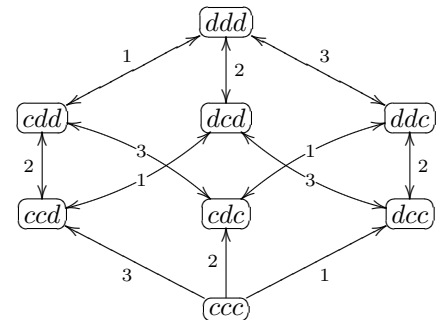
We are concerned in this paper with the problem of *long-term learning via iterated belief revision*. So we need to look at *sequences (“streams”) of upgrades*. We are in particular interested in learning true information, and so in iterating *correct, or at least truthful, upgrades*. In fact, our

main concern is with *whether the iterated belief revision process induced by truthful upgrades converges to a fixed point or not*.

**Motivation** Let us take a fresh look at the classical Muddy Children example [22]: there are  $n$  children, and  $k$  of them have mud on their faces; each can see the others, but not himself. We assume that, in the beginning, the children consider as equi-plausible all states that they cannot distinguish: so at the outset, for every child, being dirty is as plausible as being clean. For three children such that only child 1 and child 2 are dirty, the model is

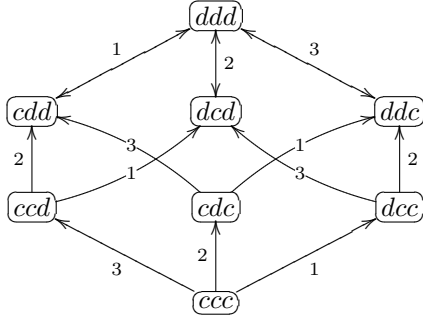


where the real state is  $ddc$  (and as before we skip all the loops in the picture). The father announces publicly: “At least one of you is dirty”. The traditional assumption is that the announcements provide “hard information”: it is common knowledge that nobody lies. But now let us relax this assumption, to obtain a “soft” version of the story. We still assume that *in fact nobody lies*, and that it is *common knowledge that everybody believes* (strongly or not, it won’t make any difference) *that nobody lies*: children trust each other and they trust the father. But nevertheless they *don’t necessarily know* that nobody lies. So we interpret father’s announcement as either a *radical upgrade*  $\uparrow(d_1 \vee d_2 \vee d_3)$  or a *conservative one*  $\uparrow(\neg d_1 \vee \neg d_2 \vee \neg d_3)$ . Both change the model to:



Then the father asks, repeatedly: “Do you believe you are dirty?” Children answer (truthfully and sincerely), and as before they all *trust* each other’s answers to be true, but they *don’t know* that they are true. Given this uncertainty, we interpret their answers again as either *radical upgrades*  $\uparrow(\neg B_1 d_1 \wedge \neg B_2 d_2 \wedge \neg B_3 d_3)$  or *conservative upgrades*  $\uparrow(\neg B_1 \neg d_1 \wedge \neg B_2 \neg d_2 \wedge \neg B_3 \neg d_3)$  (instead of updates, as traditionally done). So *no states are deleted, the uncertainty is*

never reduced, but the plausibility relations change:



Now (in the real world  $ddc$ ), children 1 and 2 believe they are dirty! In general, one can show that the classical scenario still applies in this sense: after  $k$  upgrades (including the father's announcement and  $k - 1$  answers), all dirty children will believe they are dirty. If now they answer sincerely once again, the process reaches a fixed point: no matter how many times the question is asked again, the model will stay the same from now on, since any further answering is redundant.

**Question:** Is the above phenomenon general? We know that iterated updates (on a finite model) always converge to a fixed point (see e.g. [12] and Proposition 3 below), so the classical Muddy Children story is a symptom of a general phenomenon. Is the same true for the above "soft" version?

**Redundancy, Informativity, Fixed Points** A joint upgrade  $\dagger\mathbf{P}$  is *redundant on a pointed model  $\mathbf{S}$  with respect to a group of agents  $G$*  if  $\dagger\mathbf{P}(\mathbf{S}) \simeq_G \mathbf{S}$ . Essentially, this means that, as far as the group  $G$  is concerned,  $\dagger\mathbf{P}$  doesn't change anything (when applied to model  $\mathbf{S}$ ): all the group  $G$ 's mutual beliefs, conditional beliefs, strong beliefs, mutual knowledge, common knowledge etc *stay the same* after the upgrade. An upgrade  $\dagger\mathbf{P}$  is *informative (on  $\mathbf{S}$ ) to group  $G$*  if it is not redundant with respect to  $G$ . An upgrade  $\dagger\mathbf{P}$  is *redundant with respect to (or informative to) an agent  $a$*  if it is redundant with respect to (or informative to) the singleton group  $\{a\}$ . A model  $\mathbf{S}$  is a *fixed point* of  $\dagger\mathbf{P}$  if  $\mathbf{S} \simeq \dagger\mathbf{P}(\mathbf{S})$ , i.e. if  $\dagger\mathbf{P}$  is redundant on  $\mathbf{S}$  with respect to the group of all agents  $\mathcal{A}$ .

**Proposition 1** (Logical Characterizations of Redundancy and Fixed Points)

1.  $!\mathbf{P}$  is redundant with respect to a group  $G$  iff  $\mathbf{P}$  is common knowledge in the group  $G$ ; i.e.  $\mathbf{S} \simeq_G !\mathbf{P}(\mathbf{S})$  iff  $\mathbf{S} \models Ck_G \mathbf{P}$ . Special case:  $!\mathbf{P}$  is redundant with respect to an agent  $a$  iff  $a$  knows  $\mathbf{P}$ . Another special case:  $\mathbf{S}$  is a fixed point of  $!\mathbf{P}$  iff  $\mathbf{S} \models Ck \mathbf{P}$ .
2.  $\uparrow\mathbf{P}$  is redundant with respect to a group  $G$  iff it is common knowledge in the group  $G$  that  $\mathbf{P}$  is strongly believed (by all  $G$ -agents); i.e.  $\mathbf{S} \simeq_G \uparrow\mathbf{P}(\mathbf{S})$  iff  $\mathbf{S} \models Ck_G(ESb_G \mathbf{P})$ . Special case:  $\uparrow\mathbf{P}$  is redundant with respect to an agent  $a$  iff  $a$  strongly believes

$\mathbf{P}$ . Another special case:  $\mathbf{S}$  is a fixed point of  $\uparrow\mathbf{P}$  iff  $\mathbf{S} \models Ck(ESb \mathbf{P})$ .

3.  $\uparrow\mathbf{P}$  is redundant with respect to a group  $G$  iff it is common knowledge in the group  $G$  that  $\mathbf{P}$  is believed (by all  $G$ -agents); i.e.  $\mathbf{S} \simeq_G \uparrow\mathbf{P}(\mathbf{S})$  iff  $\mathbf{S} \models Ck_G(Eb_G \mathbf{P})$ . Special case:  $\uparrow\mathbf{P}$  is redundant with respect to an agent  $a$  iff  $a$  believes  $\mathbf{P}$ . Another special case:  $\mathbf{S}$  is a fixed point of  $\uparrow\mathbf{P}$  iff  $\mathbf{S} \models Ck(Eb \mathbf{P})$ .

Redundancy and informativity are especially important if we want to capture the "sincerity" of an announcement made by a speaker. Intuitively, an announcement is sincere when it doesn't go against the speaker's prior epistemic state: *accepting the announcement should not change the speaker's own state*. So an announcement  $\dagger\mathbf{P}$  (made by an agent  $a$ ) is said to be "sincere" for (the speaker)  $a$  on a pointed model  $\mathbf{S}$  iff the upgrade  $\dagger\mathbf{P}$  is redundant with respect to  $a$ . In particular, an update  $!\mathbf{P}$  is sincere for  $a$  iff  $a$  (already) knows  $\mathbf{P}$ ; a radical upgrade  $\uparrow\mathbf{P}$  is sincere for  $a$  iff  $a$  (already) strongly believes  $\mathbf{P}$ ; and  $\dagger\mathbf{P}$  is sincere for  $a$  iff  $a$  (already) believes  $\mathbf{P}$ .

**Upgrade Streams** An *upgrade stream*  $\dagger\vec{\mathbf{P}} = (\dagger\mathbf{P}_n)_{n \in \mathbb{N}}$  is an infinite sequence of joint upgrades  $\dagger\mathbf{P}_n$  of the same type  $\dagger \in \{!, \uparrow, \dagger\}$ . An upgrade stream is *definable in a logic  $L$*  if all  $\mathbf{P}_n$  are of the form  $\mathbf{P}_n = \|\varphi_n\|$  for some  $\varphi_n \in L$ .

Any upgrade stream  $\dagger\vec{\mathbf{P}}$  induces a function mapping every pointed model  $\mathbf{S}$  into an infinite sequence  $\dagger\vec{\mathbf{P}}(\mathbf{S}) = (\mathbf{S}_n)_{n \in \mathbb{N}}$  of pointed models, defined inductively by:

$$\mathbf{S}_0 = \mathbf{S}, \quad \text{and} \quad \mathbf{S}_{n+1} = \dagger\mathbf{P}_n(\mathbf{S}_n).$$

The upgrade stream  $\dagger\vec{\mathbf{P}}$  is *truthful* if every  $\dagger\mathbf{P}_n$  is truthful with respect to  $\mathbf{S}_n$  (i.e.  $\mathbf{S}_n \models \mathbf{P}_n$ ). The stream  $\dagger\vec{\mathbf{P}}$  is *sincere* if all its upgrades are sincere for at least one agent at the moment of speaking: i.e. for every  $n$  there exists  $a_n \in \mathcal{A}$  such that  $\dagger\mathbf{P}_n$  is sincere for  $a_n$  on  $\mathbf{S}_n$ . Sincere upgrade streams model (sincere) communication processes within a group, while truthful streams model learning processes (by the group).

A *repeated truthful upgrade* is a truthful upgrade stream of the form  $(\dagger\mathbf{P}_n)_{n \in \mathbb{N}}$ , where  $\mathbf{P}_n \in \{\mathbf{P}, \neg\mathbf{P}\}$  for some proposition  $\mathbf{P}$ . In other words, it consists in repeatedly learning the answer to the "same" question  $\mathbf{P}$ ? (such as the Father's repeated question in the Muddy Children story).

We say that a stream  $\dagger\vec{\mathbf{P}}$  *stabilizes a (pointed) model  $\mathbf{S}$*  if there exists some  $n \in \mathbb{N}$  such that  $\mathbf{S}_n \simeq \mathbf{S}_m$  for all  $m > n$ . Obviously, a *repeated upgrade* stabilizes  $\mathbf{S}$  if it reaches a fixed point of  $\dagger\mathbf{P}$  or of  $\dagger(\neg\mathbf{P})$ .

We say that  $\dagger\vec{\mathbf{P}}$  *stabilizes all (simple, non-conditional) beliefs on the model  $\mathbf{S}$*  if the process of belief-changing induced by  $\dagger\vec{\mathbf{P}}$  on  $\mathbf{S}$  reaches a fixed point; i.e. if there exists some  $n \in \mathbb{N}$  such that  $\mathbf{S}_n \models B_a \mathbf{P}$  iff  $\mathbf{S}_m \models B_a \mathbf{P}$ , for all  $a \in \mathcal{A}$ , all  $m > n$  and all doxastic propositions

$\mathbf{P}$ . Equivalently, iff there exists some  $n \in N$  such that  $(\text{best}_{a s_0}(a))_{\mathbf{S}_n} = (\text{best}_{a s_0}(a))_{\mathbf{S}_m}$  for all  $a \in \mathcal{A}$  and all  $m > n$ .

Similarly, we say that  $\dagger \vec{\mathbf{P}}$  *stabilizes all conditional beliefs on the model  $\mathbf{S}$*  if the process of conditional-belief-changing induced by  $\dagger \vec{\mathbf{P}}$  on  $\mathbf{S}$  reaches a fixed point; i.e. if there exists  $n \in N$  such that  $\mathbf{S}_n \models B_a^{\mathbf{R}} \mathbf{P}$  iff  $\mathbf{S}_m \models B_a^{\mathbf{R}} \mathbf{P}$ , for all  $a \in \mathcal{A}$ , all  $m > n$  and all doxastic propositions  $\mathbf{P}, \mathbf{R}$ . Equivalently, iff there exists  $n \in N$  such that  $(\text{best}_{a s_0}(a) \cap \mathbf{R})_{\mathbf{S}_n} = (\text{best}_{a s_0}(a) \cap \mathbf{R})_{\mathbf{S}_m}$  for all  $a \in \mathcal{A}$ , all  $m > n$  and all  $\mathbf{R}$ .

A similar definition can be formulated for *stabilization of strong beliefs*. Finally,  $\dagger \vec{\mathbf{P}}$  *stabilizes all knowledge on the model  $\mathbf{S}$*  if the knowledge-changing process induced by  $\dagger \vec{\mathbf{P}}$  on  $\mathbf{S}$  reaches a fixed point; i.e. if there exists  $n \in N$  such that  $\mathbf{S}_n \models K_a \mathbf{P}$  iff  $\mathbf{S}_m \models K_a \mathbf{P}$  for all  $a \in \mathcal{A}$ , all  $m > n$  and all propositions  $\mathbf{P}$ . Equivalently, iff there exists  $n \in N$  such that  $(s_0(a))_{\mathbf{S}_n} = (s_0(a))_{\mathbf{S}_m}$  for all  $a \in \mathcal{A}$  and all  $m > n$ .

**Lemma 2** *The following are equivalent:*

- An upgrade stream  $\dagger \vec{\mathbf{P}}$  stabilizes a pointed model  $\mathbf{S}$ .
- $\dagger \vec{\mathbf{P}}$  stabilizes all conditional beliefs on  $\mathbf{S}$ .

Also, the above conditions imply that  $\dagger \vec{\mathbf{P}}$  stabilizes all knowledge, all (simple) beliefs and all strong beliefs.

The following result is just an adapted version of a result in [12]:

**Proposition 3** Every update stream stabilizes every model on which it is executable<sup>3</sup>.

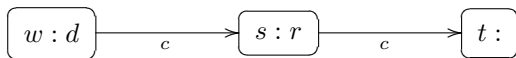
**Corollary 4** Every upgrade stream stabilizes all knowledge.

The analogue of Proposition 3 is *not true* for arbitrary upgrade streams, *not even for truthful upgrade streams*. Indeed, *it is not even true for repeated truthful upgrades*:

**Counterexample:** In the situation from Example 1, suppose that the *strongly trusted (but fallible) source* publicly announces the following true statement  $\mathbf{P}$ : “If Charles would truthfully learn that Alice won’t vote Republican, then his resulting belief about whether or not she votes Democrat would be wrong”.  $\mathbf{P}$  can be rendered as

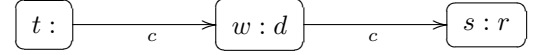
$$\neg r \Rightarrow (B^{-r} d \Leftrightarrow \neg d).$$

This is a truthful radical upgrade  $\uparrow \mathbf{P}$  (since the proposition  $\mathbf{P}$  is true in the actual world  $s$ , as well as in  $t$ ), and yields the model

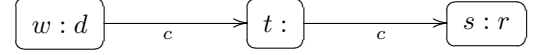


<sup>3</sup>An update stream  $(!\mathbf{P}_n)_{n \in N}$  is *executable* on  $\mathbf{S}$  if each  $!\mathbf{P}_n$  is executable at its turn, i.e. if  $\mathbf{S}_n \models \mathbf{P}_n$  for all  $n$ .

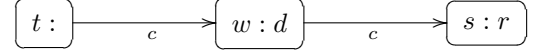
The same sentence is again true in (the real world)  $s$  and in  $w$ , so  $\uparrow \mathbf{P}$  is again truthful, resulting in:



Another truthful upgrade with the same proposition produces



then another truthful upgrade  $\uparrow \mathbf{P}$  gets us back to the previous model:



Clearly from now on the last two models *will keep reappearing, in an endless cycle*: hence in this example, *conditional beliefs never stabilize!* The same applies to strong beliefs. But note that *the simple beliefs are the same* in these last two models, since  $s$  is the most plausible world in both: so *the set of simple (non-conditional) beliefs stays the same from now on*. This is not an accident, but a symptom of a more general converge phenomenon:

**Theorem 5** Every truthful radical upgrade stream  $(\uparrow \mathbf{P}_n)_{n \in N}$  stabilizes all (simple, non-conditional) beliefs (even if it doesn’t stabilize the model).

This is the *main theorem* of this paper, and it has a number of *important consequences*. Its *proof* (included in the *Appendix*) is non-trivial, and needs a number of other preliminary Lemmas.

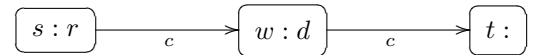
**Corollary 6** Every repeated truthful radical upgrade definable in doxastic-epistemic logic (i.e. in the language of simple belief and knowledge operators, without any conditional beliefs) stabilizes every model (with respect to which it is correct), and thus stabilizes all conditional beliefs.

Nevertheless, the (analogues of the) last two results are *not true for conservative upgrades*:

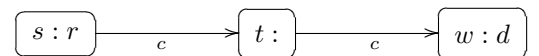
**Counterexample:** Suppose that, in the situation described in Example 1, the agents hears a *rumor*  $\mathbf{Q}$  saying that: “Either Alice will vote Republican or else Charles’ beliefs about whether or not she votes Democrat are wrong”.

$$r \vee (d \wedge \neg B d) \vee (\neg d \wedge B d)$$

This is a *truthful conservative upgrade*  $\uparrow \mathbf{Q}$ , since  $\mathbf{Q}$  is true in  $s$  (as well as in  $t$ ). Its result is:



Again, the sentence  $\mathbf{Q}$  is true in the actual state  $s$  (as well as in  $w$ ), so  $\uparrow \mathbf{Q}$  can again applied, producing:



So these two models (supporting *opposite beliefs!*) *will keep reappearing, in an endless cycle*.

## 4 Conclusions

This paper is just *a first investigation of the long-term behavior of doxastic structures under iterated revision with higher-level doxastic information*. In the context of very simple forms of new information, this type of problem has been studied in Learning Theory. In the context of updating only with purely *propositional* information, the problem has been tackled in the Belief Revision literature. But as far as we are aware, there are essentially no results on convergence of iterated revision with higher-level doxastic information in the literature to date. So our investigation here is a first step in tackling this question.

As is natural in science when a new problem is first considered, we start with the *simplest* possible setting, by considering here only (the belief revision induced by) the simplest type of communication: *truthful public announcements (though with various degrees of inherent plausibility)*. The simplicity of this type of information update (its “primitive and 1980’s-like” nature, to use the unnecessarily abusive label affixed to our paper by one unsympathetic TARK referee) is *not* due to a lack of knowledge or appreciation for wider settings allowing for richer and more sophisticated types of information update. Nor it is due to a limitation of current logical formalisms: against the impression of some economists such as the mentioned referee (impression explainable only by ignorance of the logical literature), the state of the art in Economics with respect to modeling communication does not seem to be “above and beyond” the current state in Logic. In particular, in the Dynamic Epistemic Logic tradition there exist sophisticated models for a wide range of types of communication, starting with the work in [3, 2] on “epistemic actions” that can be partially opaque or misleading to at least some of the participants, continuing with the work in [26, 16] on probabilistic epistemic updates and the work in [1, 20, 21] on multi-agent belief revision, and including our own recent contributions [5, 7, 6, 8, 10, 9] on the belief revision induced by various forms of deceiving communication, e.g. lying, misreporting, impersonation, secrecy, secret interception of messages (wiretapping) etc.

But aiming for maximum generality is *not* the point of this paper. On the contrary, if one restates our results for the most general class of communication updates, then they become *trivial*: it is *straightforward* to generate infinite belief-revision cycles with no fixed points *if one allows for cheating, lying and misreporting!* Indeed, if the truthfulness requirement is given up, then an easy example is the infinite sequence of purely propositional upgrades  $\uparrow p, \uparrow \neg p, \uparrow p, \uparrow \neg p, \dots$ . But the main point of our paper is precisely to investigate the iterated belief revision induced by “*ideal*”, *non-deceiving, public communication by a commonly trusted (though not infallible) agent*. We show that *even in such ideal circumstances*, iterated up-

dates with *higher-level* information pose new challenges and lead to a *highly non-trivial long-term dynamics*. So our “naive” simplifying assumptions about communication do *not* make our results weaker, or less widely applicable, or of more limited value. On the contrary, these results are *stronger and more general* than their analogues for wider classes of information updates; they show that infinite cycles and divergence of belief revision *cannot be avoided even in the most ideal case*, but that they are a *necessary product of iterated revision with higher-level doxastic information from a potentially fallible (even if highly trusted, and even if in fact always truthful) source*.

To summarize our results, we provide *answers to the Question* in Section 3, regarding *how typical is the “soft” version of the Muddy Children Story*, as far as convergence is concerned. The answer depends on the children’s degree of trust in the announcements. If we consider them as radical upgrades, the Muddy Children outcome is symptomatic for a general phenomenon: *repeated truthful radical upgrades with a (same) proposition definable in doxastic-epistemic logic lead to a fixed point*. However, this phenomenon is not as general as it might be: unlike the case of repeated updates, *repeated radical upgrades with truthful information do not in general lead to a true fixed point (although they stabilize the simple beliefs)*. Moreover, if we consider the announcements as conservative upgrades, then the Muddy Children outcome is atypical: *repeated conservative upgrades (even with truthful sentences in doxastic-epistemic logic) do not in general reach a fixed point (nor do they stabilize the simple beliefs)*.

The contributions in this paper are *not* tied to the specific way we model knowledge, belief, updates etc. First of all, as explained in Section 2, our plausibility-model setting is nothing but *the most natural and obvious multi-agent generalization* of structures that are *standard in Belief Revision Theory*. In their turn, these structures are simply the *qualitative core* of structures (such as *conditional lexicographic probability spaces*) that are commonly used in the Economics literature to deal with belief revision. Our qualitative version is simpler and more general, as it subsumes conditional lexicographic probability spaces, as well as conditional probability spaces in the style of Popper, Renyi and de Finetti, but in the same time it is rich and expressive enough to define all the relevant qualitative notions. In particular, our notions of “belief”, “conditional belief” and “strong belief” are essentially the same as the ones of Battigalli and Siniscalchi [11]. Second, the type of information upgrades we consider in this paper are the simplest and most natural qualitative analogues of the conditionalization rules that are most commonly used in probabilistic settings for belief update, and in their applications to Economics: in particular, our “update” !P is the qualitative (multi-agent and public) analogue of Bayesian conditionalization, while our “radical upgrade” is the qualitative (multi-agent and

public) analogue of Jeffrey conditionalization (or, more precisely, its special case of conditionalizing only with information given by a binary partition<sup>4</sup>). Thirdly, results analogue to ours *can in fact be proved directly* for conditional lexicographic probability spaces and other probabilistic belief-revision structures used in Game Theory, if one considers iterated revision by Jeffrey conditionalization with higher-level information (about the agents' conditional probabilities). As far as we know, these results are also new, and they form the topic of another forthcoming paper. The reasons we choose to stick here with a qualitative setting are *logical simplicity, generality* and *transparency*. Our setting makes obvious the fact that our results *do not depend on the specific quantitative-probabilistic details* of modeling belief and belief-revision, but they are *general, robust, qualitative results* telling us something important about the long-term belief dynamics under revision with higher-level information.

A non-logician (and in particular an economist or a philosopher of language) in the TARK audience of our talk may feel justified to follow one of our referees in asking himself or herself the question “What have I learned that I didn't really already know, at the conceptual level?”. Our answer is that this paper reveals the *surprisingly non-trivial long-term belief dynamics* induced even by *very simple, “ideal” forms of learning*, such as *truthful public communication by a non-deceiving, commonly trusted agent*. We show that, *even in our simple, general qualitative setting*, the convergence or the cyclic behavior of doxastic attitudes under iterated learning are *highly dependent on subtle details*, such as the *specific doxastic attitude whose convergence is investigated* (belief, strong belief, knowledge or conditional belief), the *listener's degree of trust in the speaker* (e.g. her *strong belief* in the speaker's veracity versus her *simple belief* in it), the *type of new information* that is received and the *language in which it is expressible* (e.g. simple propositional information, higher-level doxastic information expressible only in terms of simple belief and knowledge, or even more complex information that can refer to conditional beliefs). And we show that, *despite this complexity, certain doxastic attitudes (simple belief, knowledge) do converge in general enough conditions*.<sup>5</sup>

## Acknowledgments

The authors give special thanks to J. van Benthem for providing the seminal ideas and challenges that gave rise to this paper. We thank D. Mackinson for his insightful commentary on the second's author's LSE presentation of preliminary work leading to this paper. We thank to the anonymous

TARK referees for their comments. The research of the first author was partially supported by the Netherlands Organization for Scientific Research, grant number B 62-635, which is herewith gratefully acknowledged. The second author acknowledges the support by the University of Groningen via a Rosalind Franklin research position.

## References

- [1] G. Aucher. A combined system for update logic and belief revision. Master's thesis, ILLC, University of Amsterdam, Amsterdam, the Netherlands, 2003.
- [2] A. Baltag and L.S. Moss. Logics for epistemic programs. *Synthese*, 139:165–224, 2004. Knowledge, Rationality & Action 1–60.
- [3] A. Baltag, L.S. Moss, and S. Solecki. The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa, editor, *Proc. of TARK 98*, pages 43–56, 1998.
- [4] A. Baltag and S. Smets. Learning by questions and answers: from belief-revision cycles to doxastic fixed points. In *Proc. of WOLLIC'09*. Manuscript, to appear, 2009.
- [5] A. Baltag and S. Smets. Conditional doxastic models: a qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, 165:5–21, 2006.
- [6] A. Baltag and S. Smets. Dynamic belief revision over multi-agent plausibility models. In W. van der Hoek and M. Wooldridge, editors, *Proceedings of LOFT'06*, pages 11–24. University of Liverpool, Liverpool, 2006.
- [7] A. Baltag and S. Smets. The logic of conditional doxastic actions: a theory of dynamic multi-agent belief revision. In *Proceedings of ESSLLI Workshop on Rationality and Knowledge*. 2006.
- [8] A. Baltag and S. Smets. Probabilistic dynamic belief revision. In J. van Benthem, S. Ju, and F. Veltman, editors, *Proceedings of LORI'07*, pages 21–39. College Publications, London, 2007.
- [9] A. Baltag and S. Smets. The logic of conditional doxastic actions. In K. Apt and R. van Rooij, editors, *Texts in Logic and Games*, volume 4, pages 9–32. Amsterdam University Press, 2008.
- [10] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Texts in Logic and Games*, volume 3, pages 9–58. Amsterdam University Press, 2008.

<sup>4</sup>We consider a *more general* qualitative analogue of Jeffrey's rule in our forthcoming paper [4], where we also prove in that general setting results analogue to the ones in this paper.

<sup>5</sup>Indeed, our convergence results are generalized to arbitrary Jeffrey-type “upgrades” in our forthcoming paper [4].

- [11] P. Battigalli and M. Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 105:356–391, 2002.
- [12] J.F.A.K. van Benthem. One is a lonely number. In P. Koepke Z. Chatzidakis and W. Pohlers, editors, *Logic Colloquium 2002*, pages 96–129. ASL and A.K. Peters, Wellesley MA, 2006.
- [13] J.F.A.K. van Benthem. Dynamic logic of belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.
- [14] J.F.A.K. van Benthem. Rational dynamics. *International Game Theory Review*, 9(1):13–45, 2007.
- [15] J.F.A.K. van Benthem. Logical dynamics of information and interaction. In *Manuscript*. 2009. To appear.
- [16] J.F.A.K. van Benthem, J. van Eijck, and B.P. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.
- [17] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Number 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, 2001.
- [18] O. Board. Dynamic interactive epistemology. *Games and Economic Behaviour*, 49:49–80, 2002.
- [19] C. Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25(3):262–305, 1996.
- [20] H.P. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147:229–275, 2005.
- [21] H.P. van Ditmarsch and W. Labuschagne. My beliefs about your beliefs: a case study in theory of mind and epistemic logic. *Synthese*, 155:191–209, 2007.
- [22] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge MA, 1995.
- [23] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.
- [24] J.Y. Halpern. *Reasoning about Uncertainty*. MIT Press, Cambridge MA, 2003.
- [25] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Cambridge Tracts in Theoretical Computer Science*, pages 183–203, 1992.
- [26] B.P. Kooi. Probabilistic dynamic epistemic logic. *Journal of Logic, Language and Information*, 12:381–408, 2003.
- [27] H. Rott. Conditionals and theory change: revisions, expansions, and additions. *Synthese*, 81:91–113, 1989.
- [28] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134, 1988.
- [29] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.
- [30] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

## Appendix: Proofs

**Proof of Lemma 2:** (1)  $\Rightarrow$  (2) is obvious: if the model stabilizes, then so do the conditional beliefs.

For (2)  $\Rightarrow$  (1): suppose all conditional beliefs stabilized at stage  $n$ . Then, for  $s, t \in S$ , we have the sequence of equivalencies:  $s <_a t$  in  $\mathbf{S}_n$  iff  $\mathbf{S}_n \models B_a^{\{s,t\}} \neg\{t\}$  iff  $\mathbf{S}_m \models B_a^{\{s,t\}} \neg\{t\}$  iff  $s <_a t$  in  $\mathbf{S}_m$ . (Here, to be pedantic, we should actually replace  $\{s\}, \{t\}, \{s, t\}$  by any doxastic propositions  $\mathbf{P}, \mathbf{Q}, \mathbf{R}$  such that  $\mathbf{P}_S = \{s\}, \mathbf{Q}_S = \{t\}$  and  $\mathbf{R}_S = \{s, t\}$ .)

The rest of the claims follow from the fact that knowledge, simple belief and strong belief can be completely characterized in terms of conditional beliefs:  $K_a \mathbf{P} = B_a^- \mathbf{P}$ ;  $B_a \mathbf{P} = B_a^+ \mathbf{P}$ ; and (by the Characterization given in the paper)  $\mathbf{S} \models S b_a \mathbf{P}$  iff:  $\mathbf{S} \models B_a \mathbf{P}$  and  $\mathbf{S} \models B_a^{\mathbf{Q}} \mathbf{P}$  for every  $\mathbf{Q}$  such that  $\mathbf{S} \models \neg K_a(\mathbf{Q} \rightarrow \neg \mathbf{P})$ .

**Proof of Proposition 3:** An update changes only the set of possible states, leaving the same order, same valuation etc. on the remaining states. But the sequence  $(S_n)_{n \in \mathbb{N}}$  of sets of states of the models generated by any executable update is an *infinite descending chain*  $S_0 \supseteq S_1 \supseteq \dots \supseteq S_n \supseteq \dots$  of finite sets, and thus it *must stabilize*: there exists some stage  $n$  such that  $S_n = S_m$  for all  $m > n$ . Since everything else stays the same, the model itself stabilizes at that stage.

**Proof of Corollary 4:** For update streams, this follows from Proposition 3. The other types of upgrades never eliminate any states and never change the information partitions, so knowledge stays the same.

**For the Proof of Theorem 5,** we assume the following setting: let  $(\uparrow \mathbf{P}_n)_{n \in \mathbb{N}}$  be a radical upgrade stream, let  $\mathbf{S} = (S, \leq_a, \|\cdot\|, s_0)$  be a pointed model, and let  $(\mathbf{S}_n)_{n \in \mathbb{N}}$

be the sequence of pointed models  $\mathbf{S}_n = (S_n, \leq_a^n, \|\cdot\|, s_0)$  defined by applying in turn each of the upgrades:

$$\mathbf{S}_0 = \mathbf{S}, \quad \text{and} \quad \mathbf{S}_{n+1} = \uparrow \mathbf{P}_n(\mathbf{S}_n).$$

In the following we assume that  $(\uparrow \mathbf{P}_n)_{n \in N}$  is *truthful* with respect to  $\mathbf{S}$ , i.e.  $\mathbf{S}_n \models \mathbf{P}_n$  for all  $n \in N$ .

For  $P \subseteq S$ , we denote by  $best_a^n P := \text{Min}_{\leq_a^n} P$ .

In order to prove Theorem 5, we need a *number of preliminary lemmas*.

**Lemma 7** For all  $n \in N$ , all  $s \in s_0(a) \cap \bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}$  and all  $t \in s_0(a) \setminus \left(\bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}\right)$ , we have:  $s <_a^n t$ .

*Proof.* This is by induction on  $n$ : for  $n = 0$ , it is trivially true (since  $\bigcap_{i < 0} (\mathbf{P}_i)_{\mathbf{S}_i} = \emptyset$ ). For the induction step: we assume it true for  $n$ . After applying  $\uparrow \mathbf{P}_n$  to  $\mathbf{S}_n$ , we have (by the definition of an upgrade) that:  $s <_a^{n+1} w$  for all  $s \in s_0(a) \cap (\mathbf{P}_n)_{\mathbf{S}_n}$  and all  $w \in s_0(a) \setminus (\mathbf{P}_n)_{\mathbf{S}_n}$  (since all  $\mathbf{P}_n$ -worlds get “promoted”); and also that

$$s <_a^{n+1} t \text{ for all } s \in s_0(a) \cap (\mathbf{P}_n)_{\mathbf{S}_n} \cap \left(\bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}\right) \text{ and}$$

$$\text{for all } t \in s_0(a) \cap (\mathbf{P}_n)_{\mathbf{S}_n} \setminus \left(\bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}\right)$$

(because of the induction assumption and of the fact that inside the  $s_0(a) \cap (\mathbf{P}_n)_{\mathbf{S}_n}$ -zone the old order  $\leq_a^n$  is preserved by applying  $\uparrow \mathbf{P}_n$ ). Putting these together, and using the transitivity of  $<_a^{n+1}$ , we get the desired conclusion:

$$s <_a^{n+1} t \text{ for all } s \in s_0(a) \cap \left(\bigcap_{i < n+1} (\mathbf{P}_i)_{\mathbf{S}_i}\right) \text{ and}$$

$$\text{for all } t \in s_0(a) \setminus \left(\bigcap_{i < n+1} (\mathbf{P}_i)_{\mathbf{S}_i}\right).$$

**Lemma 8** For all  $n \in N$ , we have:

$$best_a^n(s_0(a)) \subseteq \bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}.$$

*Proof.* Suppose towards a contradiction that, for some  $n$ , there is some state  $t \in best_a^n(s_0(a))$  such that  $t \notin \bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}$ . Then  $t \in s_0(a) \setminus \left(\bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}\right)$ . By the “truthfulness” assumption, the real world  $s_0$  has the property that  $s_0 \in (\mathbf{P}_i)_{\mathbf{S}_i}$  for all  $i$ , and hence  $s_0 \in s_0(a) \cap \left(\bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}\right)$ . By Lemma 7,  $s_0 <_a^n t$ , which contradicts the assumption that  $t \in best_a^n(s_0(a))$ .

**Lemma 9** For all  $n \in N$ , we have:

$$best_a^n s_0(a) = best_a^n(s_0(a) \cap \bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i}).$$

*Proof.* This follows immediately from the previous Lemma, together with the definition of  $best_a^n$  and our assumption that  $\geq_a$  is locally connected (so that it is in fact *connected on the cell*  $s_0(a)$ ).

**Lemma 10** There exists some natural number  $n_0$  such that

$$s_0(a) \cap \bigcap_{i < n_0} (\mathbf{P}_i)_{\mathbf{S}_i} = s_0(a) \cap \bigcap_{i < m} (\mathbf{P}_i)_{\mathbf{S}_i}, \text{ for all } m \geq n_0.$$

*Proof.* This is because the sequence

$s_0(a) \supseteq s_0(a) \cap (\mathbf{P}_0)_{\mathbf{S}_0} \supseteq \dots \supseteq s_0(a) \cap \bigcap_{i < n} (\mathbf{P}_i)_{\mathbf{S}_i} \supseteq \dots$  is an infinite descending sequence of finite sets, so it must stabilize at some stage  $n_0$ .

Now we can finish the proof of our main result.

**Proof of Theorem 5:** By the definition of radical upgrades, we know that, for every  $m$ , the order inside the  $s_0(a) \cap (\mathbf{P}_m)_{\mathbf{S}_m}$  zone is left the same by  $\uparrow \mathbf{P}_m$ .

Let now  $n_0$  be as in Lemma 10. So we must have

$$s_0(a) \cap \bigcap_{i < n_0} (\mathbf{P}_i)_{\mathbf{S}_i} \subseteq s_0(a) \cap (\mathbf{P}_m)_{\mathbf{S}_m} \text{ for all } m \geq n_0.$$

Hence, the order inside  $s_0(a) \cap \bigcap_{i < n_0} (\mathbf{P}_i)_{\mathbf{S}_i}$  is left the same by all future upgrades  $\uparrow \mathbf{P}_m$ , with  $m \geq n_0$ . As a consequence: for all  $m \geq n_0$ , we have

$$best_a^{n_0} \left( s_0(a) \cap \bigcap_{i < n_0} (\mathbf{P}_i)_{\mathbf{S}_i} \right) = best_a^m \left( s_0(a) \cap \bigcap_{i < n_0} (\mathbf{P}_i)_{\mathbf{S}_i} \right)$$

This, together with the previous two Lemmas, gives us that:  $best_a^{n_0}(s_0(a)) = best_a^{n_0}(s_0(a) \cap \bigcap_{i < n_0} (\mathbf{P}_i)_{\mathbf{S}_i}) = best_a^m(s_0(a) \cap \bigcap_{i < n_0} (\mathbf{P}_i)_{\mathbf{S}_i}) = best_a^m(s_0(a) \cap \bigcap_{i < m} (\mathbf{P}_i)_{\mathbf{S}_i}) = best_a^m(s_0(a))$ , for all  $m \geq n_0$ . So the sequence of most plausible states in  $s_0(a)$  stabilizes at  $n_0$ .

**Proof of Corollary 6:** Suppose we have a repeated truthful radical upgrade  $\uparrow \mathbf{P}, \dots$ , such that  $\mathbf{P} = \|\varphi\|$  is definable by a sentence  $\varphi$  in doxastic-epistemic logic. By Corollary 4 and Theorem 5, we know that there is some  $n_0$  such that  $(s_0(a))_{\mathbf{S}_{n_0}} = (s_0(a))_{\mathbf{S}_m}$  and  $best_a^{n_0}(s_0(a)) = best_a^m(s_0(a))$ , for all  $m \geq n_0$ . We also know that the valuation is stable. Using these (and the full introspection of knowledge and beliefs), we can check (by induction on  $\psi$ ) that, for every doxastic-epistemic sentence  $\psi$ , we have  $\|\psi\|_{\mathbf{S}_{n_0}} = \|\psi\|_{\mathbf{S}_m}$  for all  $m \geq n_0$ . So, in particular, *the interpretation of the sentence  $\varphi$  stabilizes at stage  $n_0$* , and hence the set  $\mathbf{P}_{\mathbf{S}_n}$  stabilizes at stage  $n = n_0$ , and so does the set  $(\neg \mathbf{P})_{\mathbf{S}_n} = S \setminus \mathbf{P}_{\mathbf{S}_n}$  (and so in particular *only one of  $\mathbf{P}$  or  $\neg \mathbf{P}$  is true at  $\mathbf{S}_m$  for all  $m \geq n_0$* , so only one of the upgrades  $\uparrow \mathbf{P}$  or  $\uparrow (\neg \mathbf{P})$  is truthful for all  $m \geq n_0$ ). Hence (by the definition of the model transformations induced by  $\uparrow \mathbf{P}$  and by  $\uparrow (\neg \mathbf{P})$ ), applying (whichever of the two is) the *truthful* upgrade  $\uparrow \mathbf{P}$  (or  $\uparrow (\neg \mathbf{P})$ ) will not produce any more changes after this stage.