Chapter 39
A Qualitative Theory of Dynamic Interactive Belief Revision

Alexandru Baltag and Sonja Smets

Introduction

This paper contributes to the recent and on-going work in the logical community (Aucher 2003; Baltag and Sadrzadeh 2006; Baltag and Smets 2006a,b,c; van Benthem 2007; van Ditmarsch 2005) on dealing with mechanisms for belief revision and update within the Dynamic-Epistemic Logic (DEL) paradigm. DEL originates in the work of Gerbrandy and Groeneveld (1997) and Gerbrandy (1999), anticipated by Plaza in (1989), and further developed by numerous authors Baltag et al. (1998), Gerbrandy (1999), van Ditmarsch (2000, 2002), Baltag (2002), Kooi (2003), Baltag and Moss (2004), and van Benthem et al. (2006a,b) etc. In its standard incarnation, as presented e.g., in the recent textbook by van Ditmarsch et al. (2007), the DEL approach is particularly well fit to deal with *complex multi-agent learning actions* by which groups of interactive agents update their beliefs (including *higher-level beliefs* about the others' beliefs), *as long as the newly received information is consistent with the agents' prior beliefs*. On the other hand, the classical AGM theory and its more recent extensions have been very successful in dealing with the problem of *revising one-agent, first-level (factual) beliefs when they are contradicted by new information*. So it is natural to look for a way to combine these approaches.

A. Baltag (✉) • S. Smets
ILLC, University of Amsterdam, The Netherlands
e-mail: thealexandrubaltag@gmail.com; S.J.L.Smets@uva.nl

© Springer International Publishing Switzerland 2016

H. Arló-Costa et al. (eds.), *Readings in Formal Epistemology*, Springer Graduate Texts in Philosophy 1, DOI 10.1007/978-3-319-20451-2_39

We develop here a notion of *doxastic actions*,[1] general enough to cover most examples of multi-agent communication actions encountered in the literature, but also flexible enough to deal with *(both static and dynamic) belief revision*, and in particular to *implement various "belief-revision policies" in a unified setting*. Our approach can be seen as a natural extension of the work in Baltag and Moss (2004) and Baltag et al. (1998) on "epistemic actions", incorporating ideas from the AGM theory along the lines pioneered in Aucher (2003) and van Ditmarsch (2005), but using a *qualitative* approach based on *conditional beliefs*, in the line of Stalnaker (1968), Bonanno (2005), Board (2002), and van Benthem (2007).

Our paper assumes the general distinction, made in Baltag and Smets (2006a); van Benthem (2007); van Ditmarsch (2005), between *"dynamic" and "static" belief revision*. It is usually acknowledged that the classical AGM theory in Alchourrón et al. (1985) and Gärdenfors (and embodied in our setting by the *conditional belief* operators $B_a^P Q$) is indeed "static", in the sense that it captures *the agent's changing beliefs about an unchanging world*. But in fact, when we take into account all the higher-level beliefs, the "world" (that these higher-level beliefs are about) includes all agent's (real) beliefs.[2] Thus, such a world is *always changed by our changes of beliefs*! So we can better understand a belief conditional on P as capturing the agent's beliefs *after revising with P* about the state of the world *before the revision*: the statement $B_a^P Q$ says that, *if agent a would learn P, then she would come to believe that Q was the case (before the learning)*. In contrast, "dynamic" belief revision uses dynamic modalities to capture the agent's revised beliefs about the world *as it is after revision*: $[!\,P]B_a Q$ says that *after learning P, agent a would come to believe that Q is the case (in the world after the learning)*. The standard alternative (Katsuno and Mendelzon 1992) to the AGM theory calls this *belief update*, but like the AGM approach, it only deals with "first-level" (factual) beliefs from a non-modal perspective, neglecting any higher-order "beliefs about beliefs". As a result, *it completely misses the changes induced* (in our own or the other agents' epistemic-doxastic states) *by the learning actions themselves* (e.g., the learning of a Moore sentence, see the third section on ""Dynamic" Belief Revision"). This is reflected in the acceptance in Katsuno and Mendelzon (1992) of the AGM "Success Axiom": in dynamic notation, this is the axiom $[!\,P]B_a P$ (which cannot accommodate Moore sentences). Instead, Katsuno and Mendelzon (1992) exclusively concentrate on the possible changes of (ontic) facts that may have occurred during our learning (but *not due to our learning*). In contrast, our approach to belief update (following the DEL tradition) may be thought of as "dual" to the one in Katsuno and Mendelzon

[1] Or "doxastic events", in the terminology of van Benthem (2007).

[2] To verify that a higher-level belief about another belief is "true" we need to check the content of that higher-level belief (i.e., the existence of the second, lower-level belief) against the "real world". So the real world has to include the agent's beliefs.

(1992): we completely neglect here the ontic changes,[3] considering only the changes induced by *"purely doxastic" actions* (learning by observation, communication, etc.).

Our formalism for "static" revision can best be understood as a modal-logic implementation of the well-known view of belief revision in terms of *conditional reasoning* (Stalnaker 1968, 2006). In Baltag and Smets (2006a,c), we introduced two equivalent semantic settings for conditional beliefs in a multi-agent epistemic context (*conditional doxastic models* and *epistemic plausibility models*), taking the first setting as the basic one. Here, we adopt the second setting, which is closer to the standard semantic structures used in the literature on modeling belief revision (Board 2002; Friedmann and Halpern 1994; Grove 1988; Spohn 1988; Stalnaker 2006; van Benthem 2007, 2004). We use this setting to define notions of *knowledge* K_aP, *belief* B_aP and *conditional belief* B_a^QP. Our concept of "knowledge" is the standard S5-notion, partition-based and fully introspective, that is commonly used in Computer Science and Economics, and is sometimes known as "Aumann knowledge", as a reference to Aumann (1999). The conditional belief operator is a way to "internalize", in a sense, the "static" (AGM) belief revision within a modal framework: saying that, at state s, agent a believes P conditional on Q is a way of saying that Q belongs to a's revised "theory" (capturing her revised beliefs) after revision with P (of a's current theory/beliefs) at state s. Our conditional formulation of "static" belief revision is close to the one in Stalnaker (1968), Ryan and Schobbens (1997), Board (2002), Bonanno (2005), and Rott (1989). As in Board (2002), the preference relation is assumed to be well-preordered; as a result, the logic CDL of conditional beliefs is equivalent to the strongest system in Board (2002).

We also consider other modalities, capturing other "doxastic attitudes" than just knowledge and conditional belief. The most important such notion expresses a form of "weak (non-introspective) knowledge" $\Box_aP$, first introduced by Stalnaker in his modal formalization (Stalnaker 1968, 2006) of Lehrer's *defeasibility analysis of knowledge* (Lehrer 1990; Lehrer and Paxson 1969). We call this notion *safe belief*, to distinguish it from our (Aumann-type) concept of knowledge. Safe belief can be understood as belief that is *persistent under revision with any true information*. We use this notion to give a new solution to the so-called "Paradox of the Perfect Believer". We also solve the open problem posed in Board (2002), by providing a *complete axiomatization of the "static" logic* $K\Box$ *of conditional belief, knowledge and safe belief*. In a forthcoming paper, we apply the concept of safe belief to Game Theory, improving on Aumann's epistemic analysis of backwards induction in games of perfect information.

Moving thus on to *dynamic belief revision*, the first thing to note is that (unlike the case of "static" revision), *the doxastic features of the actual "triggering event"* that induced the belief change *are essential* for understanding this change (as a "dynamic

[3] But our approach can be easily modified to incorporate ontic changes, along the lines of van Benthem et al. (2006b).

revision", i.e., in terms of the revised beliefs about the state of the world after revision). For instance, our beliefs about *the current situation after* hearing a *public* announcement (say, of some *factual* information, denoted by an atomic sentence p) are different from our beliefs after receiving a *fully private* announcement with the same content p. Indeed, in the public case, we come to believe that p is now *common knowledge* (or at least *common belief*). While, in the private case, we come to believe that the content of the announcement forms now our *secret knowledge*. So the agent's *beliefs about the learning actions* in which she is currently engaged affect the way she updates her previous beliefs.

This distinction is irrelevant for "static" revision, since e.g., in both cases above (public as well as private announcement) we learn the same thing about the situation that existed *before the learning*: our beliefs about that past situation will change in the same way in both cases. More generally, our beliefs about the "triggering action" are irrelevant, as far as our "static" revision is concerned. This explains a fact observed in van Benthem (2007), namely that by and large, the standard literature on belief revision (or belief update) *does not usually make explicit the doxastic events that "trigger" the belief change* (dealing instead only with types of abstract operations on beliefs, such as update, revision and contraction etc). The reason for this lies in the "static" character of AGM revision, as well as its restriction (shared with the "updates" of Katsuno and Mendelzon 1992) to one-agent, first-level, factual beliefs.

A "truly dynamic" logic of belief revision has to be able to capture the *doxastic-epistemic features (e.g., publicity, complete privacy etc.) of specific "learning events"*. We need to be able to model the agents' "dynamic beliefs", i.e., their *beliefs about the learning action itself*: the *appearance* of this action (while it is happening) to each of the agents. In Baltag and Moss (2004), it was argued that a natural way to do this is to use *the same type of formalism that was used to model "static" beliefs*: *epistemic actions should be modeled in essentially the same way as epistemic states*; and this common setting was taken there to be given by *epistemic Kripke models*.

A similar move is made here in the context of our richer doxastic-plausibility structures, by introducing *plausibility pre-orders on actions* and developing a notion of "action plausibility models", that extends the "epistemic action models" from Baltag and Moss (2004), along similar lines to (but without the quantitative features of) the work in Aucher (2003) and van Ditmarsch (2005).

Extending to (pre)ordered models the corresponding notion from Baltag and Moss (2004), we introduce an operation of *product update* of such models, based on the *anti-lexicographic order* on the product of the state model with the action model. The simplest and most natural way to define a connected pre-order on a Cartesian product from connected pre-orders on each of the components is to use either the *lexicographic* or the *anti-lexicographic* order. Our choice is the second, which we regard as the *natural generalization of the AGM theory*, giving *priority to incoming information* (i.e., to "actions" in our sense). This can also be thought of as a generalization of the so-called *"maximal-Spohn" revision*. We call this type of update rule the *"Action-Priority" Update*. The intuition is that the beliefs encoded in the action model express *the "incoming" changes of belief*, while the state model

only captures that *past beliefs*. One could say that the new "beliefs about actions" are *acting* on the prior "beliefs about states", producing the updated (posterior) beliefs. This is embedded in the Motto that we give in the paragraph on "Action Models" in the third section, the Motto is: "*beliefs about changes encode (and induce) changes of beliefs*".

By abstracting away from the quantitative details of the plausibility maps when considering the associated *dynamic logic*, our approach to dynamic belief revision is in the spirit of the one in van Benthem (2007): instead of using "graded belief" operators as in e.g., Aucher (2003) and van Ditmarsch (2005), or probabilistic modal logic as in Kooi (2003), both our account and the one in van Benthem (2007) concentrate on the simple, qualitative language of *conditional beliefs, knowledge and action modalities* (to which we add here the *safe belief* operator). As a consequence, we obtain *simple, elegant, general logical laws of dynamic belief revision*, as natural generalizations of the ones in van Benthem (2007). These "reduction laws" give a *complete axiomatization of the logic of doxastic actions*, "reducing" it to the "static" logic $K\Box$. Compared both to our older axiomatization in Baltag and Smets (2006c) and to the system in Aucher (2003), one can easily see that the introduction of the safe belief operator leads to a major simplification of the reduction laws.

Our qualitative logical setting (in this paper and in Baltag and Smets 2006a,b,c), as well as the closely related setting in van Benthem (2007), are conceptually very different from the more "quantitative" approaches to dynamic belief revision taken in (Aucher 2003; van Ditmarsch 2005; van Ditmarsch and Labuschagne 2007), approaches based on "degrees of belief" given by ordinal plausibility functions. This is not just a matter of interpretation, but it makes a difference for the choice of dynamic revision operators. Indeed, the update mechanisms proposed in Spohn (1988), Aucher (2003), and van Ditmarsch (2005) are essentially quantitative, using various binary functions in transfinite ordinal arithmetic, in order to compute the degree of belief of the output-states in terms of the degrees of the input-states and the degrees of the actions. This leads to an increase in complexity, both in the computation of updates and in the corresponding logical systems. Moreover, there seems to be no canonical choice for the arithmetical formula for updates, various authors proposing various formulas. No clear intuitive justification is provided to any of these formulas, and we see no transparent reason to prefer one to the others. In contrast, classical (AGM) belief revision theory is a qualitative theory, based on natural, intuitive postulates, of great generality and simplicity.

Our approach retains this qualitative flavor of the AGM theory, and aims to build a theory of "dynamic" belief revision of equal simplicity and naturality as the classical "static" account. Moreover (unlike the AGM theory), it aims to provide a *"canonical"* choice for a dynamic revision operator, given by our "Action Priority" update. This notion is a *purely qualitative one*,[4] based on a *simple, natural relational*

[4]One could argue that our plausibility pre-order relation is equivalent to a quantitative notion (of ordinal degrees of plausibility, such as in Spohn (1988)), but unlike in Aucher (2003) and van

definition. From a *formal point of view*, one might see our choice of the anti-lexicographic order as *just one of the many possible options* for developing a belief-revision-friendly notion of update. As already mentioned, it is a generalization of the "maximal-Spohn" revision, already explored in van Ditmarsch (2005) and Aucher (2003), among many other possible formulas for combining the "degrees of belief" of actions and states. But here we justify our option, arguing that our *qualitative interpretation of the plausibility order makes this the only reasonable choice*.

It may seem that by making this choice, we have confined ourselves to *only one of the bewildering multitude of "belief revision policies"* proposed in the literature by Spohn (1988), Rott (1989), Segerberg (1998), Aucher (2003), van Ditmarsch (2005), van Benthem (2004), and van Benthem (2007). But, as argued below, *this apparent limitation is not so limiting after all*, but can instead be regarded as an *advantage*: the power of the "action model" approach is reflected in the fact that *many different belief revision policies* can be recovered as *instances of the same type of update operation*. In this sense, our approach can be seen as a *change of perspective*: the diversity of possible revision policies is replaced by the diversity of possible action models; the differences are now viewed as *differences in input, rather than having different "programs"*. For a computer scientist, this resembles "Currying" in lambda-calculus: if every "operation" is encoded as an input-term, then *one operation* (functional application) *can simulate all operations*.[5] In a sense, this is nothing but the idea of Turing's universal machine, which underlies universal computation.

The title of our paper is a paraphrase of Oliver Board's "Dynamic Interactive Epistemology" (Board 2002), itself a paraphrase of the title ("Interactive Epistemology") of a famous paper by Aumann (1999). We interpret the word "interactive" as referring to the *multiplicity of agents* and the *possibility of communication*. Observe that "interactive" does not necessarily imply "dynamic": indeed, Board and Stalnaker consider Aumann's notion to be "static" (since it doesn't accommodate any non-trivial belief revision). But even Board's logic, as well as Stalnaker's (2006), are "static" in our sense: they cannot directly capture the effect of learning *actions* (but can only express "static" conditional beliefs). In contrast, our DEL-based approach has all the "dynamic" features and advantages of DEL: in addition to "simulating" a range of individual belief-revision policies, it can deal with an even wider range of *complex types of multi-agent learning and communication actions*. We thus think it is realistic to expect that, *within its own natural limits*,[6] our Action-Priority Update Rule could play the role of a *"universal machine" for qualitative dynamic interactive belief-revision*.

Ditmarsch (2005) the way belief update is defined in our account does not make any use of the ordinal "arithmetic" of these degrees.

[5]Note that, as in untyped lambda-calculus, the input-term encoding the operation (i.e., our "action model") and the "static" input-term to be operated upon (i.e., the "state model") are essentially *of the same type*: epistemic plausibility models for the same language (and for the same set of agents).

[6]E.g., our update cannot deal with "forgetful" agents, since "perfect recall" is in-built. But finding out what exactly are the "natural limits" of our approach is for now an open problem.

"Static" Belief Revision

Using the terminology in van Benthem (2007) and Baltag and Smets (2006a,b,c, 2007a), "static" belief revision is about *pre-encoding potential belief revisions as conditional beliefs*. A conditional belief statement $B_a^P Q$ can be thought of as expressing a "doxastic predisposition" or a "plan of doxastic action": the agent is determined to believe that Q was the case, if he learnt that P was the case. The semantics for conditional beliefs is usually given in terms of plausibility models (or equivalent notions, e.g., "spheres", "onions", ordinal functions etc.) As we shall see, both *(Aumann, S5-like) knowledge* and *simple (unconditional) belief* can be defined in terms of conditional belief, which itself could be defined in terms of a *unary belief-revision operator*: $*_a P$ captures *all the revised beliefs* of agent a after revising (her current beliefs) with P.

In addition, we introduce a *safe belief* operator $\Box_a P$, meant to express a weak notion of "defeasible knowledge" (obeying the laws of the modal logic $S4.3$). This concept was defined in Stalnaker (2006) and Board (2002) using a higher-order semantics (quantifying over conditional beliefs). But this is in fact equivalent to a first-order definition, as the Kripke modality for the (converse) plausibility relation. This observation greatly simplifies the task of completely axiomatizing the logic of safe belief and conditional beliefs: indeed, our proof system $K\Box$ below is a solution to the open problem posed in Board (2002).

Plausibility Models: The Single Agent Case

To warm up, we consider first the case of only *one agent*, a case which fits well with the standard models for belief revision.

A *single-agent plausibility frame* is a structure $(S, \leq)$, consisting of a set S of "states" and a "well-preorder" $\leq$, i.e., a reflexive, transitive binary relation on S such that *every non-empty subset has minimal elements*. Using the notation

$$\mathrm{Min}_{\leq} P := \{s \in P : s \leq s' \text{ for all } s' \in P\}$$

for the set of $\leq$-minimal elements of P, the last condition says that: For every set $P \subseteq S$, if $P \neq \emptyset$ then $\mathrm{Min}_{\leq} P \neq \emptyset$.

The usual reading of $s \leq t$ is that "state s is *at least as plausible* as state t". We keep this reading for now, though we will later get back to it and clarify its meaning. The "minimal states" in $\mathrm{Min}_{\leq} P$ are thus the "most plausible states" satisfying proposition P. As usual, we write $s < t$ iff $s \leq t$ but $t \not\leq s$, for the *"strict" plausibility relation* (s is *more plausible* than t). Similarly, we write $s \cong t$ iff both $s \leq t$ and $t \leq s$, for the *"equi-plausibility"* (or *indifference*) relation (s and t are *equally plausible*).

S-propositions and models. Given an epistemic plausibility frame S, an *S-proposition* is any subset $P \subseteq S$. Intuitively, we say that *a state s satisfies the proposition P* if $s \in P$. Observe that a plausibility frame is just a special case of a

relational frame (or *Kripke frame*). So, as it is standard for Kripke frames in general, we can define a *plausibility model* to be a structure $\mathbf{S} = (S, \leq, \|\cdot\|)$, consisting of a plausibility frame $(S, \leq)$ together with a valuation map $\|\cdot\| : \Phi \to \mathcal{P}(S)$, mapping every element of a given set Φ of "atomic sentences" into S-propositions.

Interpretation. The elements of S will represent the *possible states* (or "possible worlds") of a system. 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. Finally, the plausibility relations $\leq$ capture the agent's *(conditional) beliefs about the state* of the system; if e.g., the agent was given the information that the state of the system is either s or t, she would believe that the system was in the *most plausible* of the two. So, if $s < t$, the agent would believe the real state was s; if $t < s$, she would believe it was t; otherwise (if $s \cong t$), the agent would be indifferent between the two alternatives: she will not be able to decide to believe any one alternative rather than the other.

Propositional operators, Kripke modalities. For every model $\mathbf{S}$, we have the usual Boolean operations with S-propositions

$$P \wedge Q := P \cap Q, \quad P \vee Q := P \cup Q,$$

$$\neg P := S \setminus P, \quad P \to Q := \neg P \vee Q,$$

as well as Boolean constants $\top_S := S$ and $\bot_S := \emptyset$. Obviously, one may also introduce *infinitary* conjunctions and disjunctions. In addition, any binary relation $R \subseteq S \times S$ on S gives rise to a *Kripke modality* $[R] : \mathcal{P}(S) \to \mathcal{P}(S)$, defined by

$$[R]Q := \{s \in S : \forall t\, (sRt \Rightarrow t \in Q)\}.$$

Accessibility relations for belief, conditional belief and knowledge. To talk about beliefs, we introduce a *doxastic accessibility relation* $\to$, given by

$$s \to t \ \text{ iff } \ t \in \mathrm{Min}_{\leq} S\,.$$

We read this as saying that: when the actual state is s, the agent believes that *any* of the states t with $s \to t$ *may be* the actual state. This matches the above interpretation of the preorder: the states believed to be possible are the minimal (i.e., "most plausible") ones.

In order to talk about *conditional beliefs*, we can similarly define a *conditional doxastic accessibility relation* for each S-proposition $P \subseteq S$:

$$s \to^P t \ \text{ iff } \ t \in \mathrm{Min}_{\leq} P\,.$$

We read this as saying that: when the actual state is s, if the agent is given the information (that) P (is true at the actual state), then she believes that *any* of the states t with $s \to t$ *may be* the actual state.

Finally, to talk about knowledge, we introduce a relation of *epistemic possibility* (or "indistinguishability") $\sim$. Essentially, this is just the universal relation:

$$s \sim t \text{ iff } s, t \in S.$$

So, in single-agent models, *all* the states in S are assumed to be "epistemically possible": the only thing *known* with absolute certainty about the current state is that it belongs to S. This is natural, in the context of a single agent: the states known to be impossible are *irrelevant* from the point of doxastic-epistemic logic, so they can simply be excluded from our model S. (As seen below, this cannot be done in the case of multiple agents!)

Knowledge and (conditional) belief. We define *knowledge* and *(conditional) belief* as the Kripke modalities for the epistemic and (conditional) doxastic accessibility relations:

$$KP := [\sim]P,$$

$$BP := [\rightarrow]P,$$

$$B^Q P := [\rightarrow^Q]P.$$

We read KP as saying that the (implicit) agent *knows* P. This is "knowledge" in the strong Leibnizian sense of "truth in all possible worlds". We similarly read BP as "P is believed" and $B^Q P$ as "P is believed given (or conditional on) Q". As for *conditional belief* statements $s \in B^Q P$, we interpret them in the following way: if the actual state is s, then after coming to believe that Q is the case (at this actual state), the agent will believe that P *was* the case (at the same actual state, *before* his change of belief). In other words, conditional beliefs B^Q give descriptions of the agent's *plan* (or *commitments*) about what he will believe about the current state after receiving new (believable) information. To quote Johan van Benthem in (2007), conditional beliefs are *"static pre-encodings" of the agent's potential belief changes* in the face of new information.

Discussion on interpretation. Observe that our interpretation of the plausibility relations is *qualitative*, in terms of *conditional beliefs* rather than "degrees of belief": there is no scale of beliefs here, allowing for "intermediary" stages between believing and not believing. Instead, all these beliefs are equally "firm" (*though conditional*): given the condition, something is either believed or not. To repeat, writing $s < t$ is for us just a way to say that: if *given* the information that the state of the system is either s or t, the agent would believe it to be s. So plausibility relations are special cases of conditional belief. This interpretation is based on the following (easily verifiable) equivalence:

$$s < t \text{ iff } s \in B^{\{s,t\}}\{s\} \text{ iff } t \in B^{\{s,t\}}\{s\}.$$

There is nothing quantitative here, no need for us to refer in any way to the "strength" of this agent's belief: though she might have beliefs of unequal strengths, we are not interested here in modeling this quantitative aspect. Instead, we give the agent some information about a state of a virtual system (that it is either s or t) and we ask her a *yes-or-no question* ("Do you believe that virtual state to be s ?"); we write $s < t$ iff the agent's answer is "yes". This is a firm answer, so it expresses a firm belief. "Firm" does not imply "un-revisable" though: if later we reveal to the agent that the state in question was in fact t, she should be able to accept this new information; after all, the agent should be introspective enough to realize that her belief, however firm, was just a belief.

One possible objection against this qualitative interpretation is that our postulate that $\leq$ is a well-preorder (and so in particular a connected pre-order) introduces a hidden "quantitative" feature; indeed, any such preorder can be equivalently described using a plausibility map as in e.g., Spohn (1988), assigning ordinals to states. Our answer is that, first, the specific ordinals will not play any role in our definition of a dynamic belief update; and second, all our postulates can be given a justification in purely qualitative terms, using conditional beliefs. The transitivity condition for $\leq$ is just a *consistency* requirement imposed on a rational agent's conditional beliefs. And the existence of minimal elements in any non-empty subset is simply the natural extension of the above setting to *general* conditional beliefs, not only conditions involving two states: more specifically, for any possible condition $P \subseteq S$ about a system S, the S-proposition $\mathrm{Min}_{\leq} P$ is simply a way to encode everything that the agent would believe about the current state of the system, if she was given the information that the state satisfied condition P.

Note on other models in the literature. Our models are the same as Board's "belief revision structures" (Board 2002), i.e., nothing but "Spohn models" as in Spohn (1988), but with a purely relational description. Spohn models are usually described in terms of a map assigning ordinals to states. But giving such a map is equivalent to introducing a well pre-order $\leq$ on states, and it is easy to see that all the relevant information is captured by this order.

Our conditions on the preorder $\leq$ can also be seen as a *semantic analogue* of Grove's conditions for the (relational version of his) models in Grove (1988). The standard formulation of Grove models is in terms of a "system of spheres" (weakening Lewis' similar notion), but it is equivalent (as proved in Grove 1988) to a relational formulation. Grove's postulates are still *syntax-dependent*, e.g., existence of minimal elements is required only for subsets that are *definable* in his language: this is the so-called "smoothness" condition, which is weaker than our "well-preordered" condition. We prefer a purely semantic condition, independent of the choice of a language, both for reasons of elegance and simplicity and because we want to be able to consider more than one language for the same structure.[7]

[7] Imposing syntactic-dependent conditions in the very definition of a class of structures makes the definition meaningful only for one language; or else, the meaning of what, say, a plausibility model is won't be *robust*: it will change whenever one wants to extend the logic, by adding a few more

So, following Board (2002) and Stalnaker (2006) and others, we adopt the natural semantic analogue of Grove's condition, simply requiring that *every* subset has minimal elements: this will allow our conditional operators to be well-defined on sentences of *any* extension of our logical language.

Note that the minimality condition implies, by itself, that the relation $\leq$ is both *reflexive* (i.e., $s \leq s$ for all $s \in S$) and *connected*[8] (i.e., either $s \leq t$ or $t \leq s$, for all $s, t \in S$). In fact, a "well-preorder" is the same as a *connected, transitive, well-founded*[9] *relation*, which is the setting proposed in Board (2002) for a logic of conditional beliefs equivalent to our logic CDL below. Note also that, when the set S is *finite*, a well-preorder is nothing but a *connected preorder*. This shows that our notion of frame subsumes, not only Grove's setting, but also some of the other settings proposed for conditionalization.

Multi-agent Plausibility Models

In the multi-agent case, *we cannot exclude from the model the states that are known to be impossible* by some agent a: they may still be considered possible by a second agent b. Moreover, they might still be relevant for a's beliefs/knowledge about what b believes or knows. So, in order to define an agent's knowledge, we cannot simply quantify over *all* states, as we did above: instead, we need to consider, as usually done in the Kripke-model semantics of knowledge, only the "possible" states, i.e., the ones that are *indistinguishable* from the real state, as far as a given agent is concerned. It is thus natural, in the multi-agent context, to explicitly specify the agents' epistemic indistinguishability relations $\sim_a$ (labeled with the agents' names) as part of the basic structure, in addition to the plausibility relations $\leq_a$. Taking this natural step, we obtain *epistemic plausibility frames* $(S, \sim_a, \leq_a)$. As in the case of a single agent, specifying epistemic relations turns out to be *superfluous*: the relations $\sim_a$ can be recovered from the relations $\leq_a$. Hence, we will simplify the above structures, obtaining the equivalent setting of *multi-agent plausibility frames* $(S, \leq_a)$.

Before going on to define these notions, observe that it doesn't make sense anymore to require the plausibility relations $\leq_a$ to be connected (and even less sense to require them to be well-preordered): if two states s, t are *distinguishable* by an agent a, i.e., $s \not\sim_a t$, then a will never consider both of them as epistemically possible in the same time. If she was given the information that the real state is either s or t, agent a will immediately *know* which of the two: if the real state was s, she would be able to distinguish this state from t, and would thus know the state

operators. This is very undesirable, since then one cannot compare the expressivity of different logics on the same class of models.

[8] In the Economics literature, connectedness is called "completeness", see e.g., Board (2002).

[9] I.e., there exists no infinite descending chain $s_0 > s_1 > \cdots$.

was s; similarly, if the real state was t, she would know it to be t. Her beliefs will play no role in this, and it would be meaningless to ask her which of the two states is more plausible to her. So only the states in the same $\sim_a$-equivalence class could, and should, be $\leq_a$-comparable; i.e., $s \leq_a t$ *implies* $s \sim_a t$, and the restriction of $\leq_a$ to each $\sim_a$-equivalence class is connected. Extending the same argument to arbitrary conditional beliefs, we can see that *the restriction of* $\leq_a$ *to each* $\sim_a$*-equivalence class must be well-preordered.*

Epistemic plausibility frames. Let $\mathcal{A}$ be a finite set of labels, called *agents*. An *epistemic plausibility frame* over $\mathcal{A}$ (EPF, for short) is a structure $\mathbf{S} = (S, \sim_a, \leq_a)_{a\in\mathcal{A}}$, consisting of a set S of "states", endowed with a family of equivalence relations $\sim_a$, called *epistemic indistinguishability relations*, and a family of *plausibility relations* $\leq_a$, both labeled by "agents" and assumed to satisfy two conditions: (1) $\leq_a$-comparable states are $\sim_a$-indistinguishable (i.e., $s \leq_a t$ implies $s \sim_a t$); (2) the restriction of each plausibility relation $\leq_a$ to each $\sim_a$-equivalence class is a well-preorder. As before, we use the notation $\mathrm{Min}_{\leq_a} P$ for the set of $\leq_a$-minimal elements of P. We write $s <_a t$ iff $s \leq_a t$ but $t \not\leq_a s$ (the *"strict" plausibility relation*), and write $s \cong_a t$ iff both $s \leq_a t$ and $t \leq_a s$ (the *"equi-plausibility" relation*). The notion of *epistemic plausibility models* (EPM, for short) is defined in the same way as the plausibility models in the previous section.

Epistemic plausibility models. We define a *(multi-agent) epistemic plausibility model* (EPM, for short) as a multi-agent EPF together with a valuation over it (the same way that single-agent plausibility models were defined in the previous section).

It is easy to see that our definition of EPFs includes superfluous information: in an EPF, the knowledge relation $\sim_a$ can be recovered from the plausibility relation $\leq_a$, via the following rule:

$$s \sim_a t \text{ iff either } s \leq_a t \text{ or } t \leq_a s\,.$$

In other words, two states are indistinguishable for a iff they are *comparable* (with respect to $\leq_a$).

So, in fact, one could present epistemic plausibility frames simply as *multi-agent plausibility frames*. To give this alternative presentation, we use, for *any preorder relation* $\leq$, the notation $\sim$ for the associated *comparability relation*

$$\sim \;:= \leq \cup \geq$$

(where $\geq$ is the converse of $\leq$). A *comparability class* is a set of the form $\{t : s \leq t \text{ or } t \leq s\}$, for some state s. A relation $\leq$ is called *locally well-preordered* if it is a preorder such that its restriction to each comparability class is well-preordered. Note that, when the underlying set S is *finite*, a locally well-preordered relation is nothing but a *locally connected preorder*: a preorder whose restrictions to any comparability class are connected. More generally, *a locally well-preordered relation is the same as a locally connected and well-founded preorder.*

Multi-agent plausibility frames. A *multi-agent plausibility frame* (MPF, for short) is a structure $(S, \leq_a)_{a\in\mathcal{A}}$, consisting of a set of states S together with a family of locally well-preordered relations $\leq_a$, one for each agent $a \in \mathcal{A}$. Oliver Board (2002) calls multi-agent plausibility frames "belief revision structures". A *multi-agent plausibility model* (MPM, for short) is an MPF together with a valuation map.

Bijective correspondence between EPFs and MPFs. *Every MPF can be canonically mapped into an EPF*, obtained by defining epistemic indistinguishability via the above rule ($\sim_a := \leq_a \cup \geq_a$). Conversely, every EPF gives rise to an MPF, via the map that "forgets" the indistinguishability structure. It is easy to see that these two maps are the inverse of each other. Consequently, from now on we *identify MPFs and EPFs, and similarly identify MPMs and EPMs*; e.g., we can talk about "knowledge", "(conditional) belief" etc. in an MPM, defined in terms of the associated EPM.

So from now on we identify the two classes of models, via the above canonical bijection, and talk about "plausibility models" in general. One can also see how this approach relates to another widely adopted definition for conditional beliefs; in Board (2002), van Ditmarsch (2005), and van Benthem (2007), this definition involves the assumption of a *"local plausibility" relation at a given state* $s \leq_a^w t$, to be read as: "at state w, agent a considers state s at least as plausible as state t ". Given such a relation, the conditional belief operator is usually defined in terms that are equivalent to putting $s \to_a^P t$ iff $t \in \mathrm{Min}_{\leq_a^s} P$. One could easily restate our above definition in this form, by taking:

$$s \leq_a^w t \quad \text{iff} \quad \text{either } w \not\sim_a t \text{ or } s \leq_a t.$$

The converse problem is studied in Board (2002), where it is shown that, if full introspection is assumed, then one can recover "uniform" plausibility relations $\leq_a$ from the relations $\leq_a^w$.

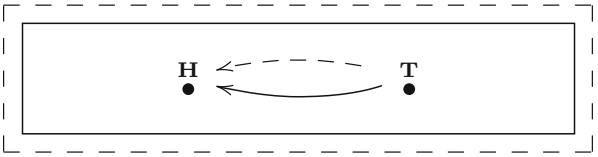
Information cell. The equivalence relation $\sim_a$ induces a partition of the state space S, called *agent a's information partition*. We denote by $s(a)$ the *information cell* of s in a's partition, i.e., the $\sim_a$-equivalence class of s:

$$s(a) := \{t \in S : s \sim_a t\}.$$

The information cell $s(a)$ captures *all the knowledge possessed by the agent* at state s: when the actual state of the system is s, then agent a knows only the state's equivalence class $s(a)$.

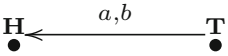
Example 1. Alice and Bob play a game, in which an anonymous referee puts a coin on the table, lying face up but in such a way that the face is covered (so Alice and Bob cannot see it). Based on previous experience, (it is common knowledge that) Alice and Bob believe that the upper face is Heads (since e.g., they noticed that

the referee had a strong preference for Heads). And in fact, they're right: the coin lies Heads up. Neglecting the anonymous referee, the EPM for this example is the following model **S**:

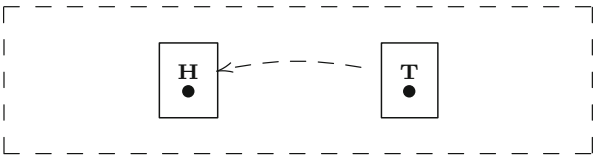


Here, the arrows represent *converse plausibility relations* $\geq$ *between distinct states only* (going from less plausible to more plausible states): since these are always reflexive, we choose to *skip all the loops* for convenience. The squares represent the *information cells* for the two agents. Instead of labels, we use *dashed arrows and squares for Alice*, while using *continuous arrows and squares for Bob*. In this picture, the actual state of the system is the state s on the left (in which **H** is true). Henceforth, in our other examples, we will refer to this particular plausibility model as **S**.

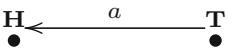
By deleting the squares, we obtain a representation of the corresponding MPM, also denoted by **S** (where we now use labels for agents instead of different types of lines):



Example 2. In front of Alice, the referee shows the face of the coin to Bob, but Alice cannot see the face. The EPM is now the following model **W**:



while the MPM is



Since Bob now knows the state of the coin, his local plausibility relation consists only of loops, and hence we have no arrows for Bob in this diagrammatic representation.

(Conditional) doxastic appearance and (conditional) doxastic accessibility. As in the previous section, we can define a doxastic and epistemic accessibility relations, except that now we have to select, for each state s, the most plausible states in its information cell $s(a)$ (instead of the most plausible in S). For this, it is convenient to introduce some notation and terminology: the *doxastic appearance* of state s to agent a is the set

$$s_a := \mathrm{Min}_{\leq_a} s(a)$$

of the "most plausible" states that are consistent with the agent's knowledge at state s. The doxastic appearance of s captures *the way state s appears to the agent*, or (in the language of Belief Revision) *the agent's current "theory" about the world s*. We can extend this to capture *conditional beliefs* (in full generality), by associating to each S-proposition $P \subseteq S$ and each state $s \in S$ the *conditional doxastic appearance* s_a^P *of state s to agent a, given (information) P*. This can be defined as the S-proposition

$$s_a^P := \mathrm{Min}_{\leq_a} s(a) \cap P$$

given by the set of all $\leq_a$-minimal states of $s(a) \cap P$: these are the "most plausible" states satisfying P that are consistent with the agent's knowledge at state s. The conditional appearance S_a^P gives *the agent's revised theory (after learning P) about the world s*. We can put these in a relational form, by defining *doxastic accessibility relations* $\rightarrow_a$, $\rightarrow_a^P$, as follows:

$$s \rightarrow_a t \text{ iff } t \in s_a\,,$$
$$s \rightarrow_a^P t \text{ iff } t \in s_a^P\,.$$

Knowledge and (conditional) belief. As before, we define the *knowledge* and *(conditional) belief* operators for an agent a as the Kripke modalities for a's epistemic and (conditional) doxastic accessibility relations:

$$K_a P := [\sim_a]P = \{s \in S : s(a) \subseteq P\}\,,$$
$$B_a P := [\rightarrow_a]P = \{s \in S : s_a \subseteq P\}\,,$$
$$B_a^Q P := [\rightarrow_a^Q]P = \{s \in S : s_a^Q \subseteq P\}.$$

We also need a notation for the *dual of the K modality* ("epistemic possibility"):

$$\tilde{K}_a P := \neg K_a \neg P.$$

Doxastic propositions. Until now, our notion of proposition is "local", being specific to a given model: we only have "S-propositions" for each model S. As long as the model is fixed, this notion is enough for interpreting sentences over the

given model. But, since later we will proceed to study systematic *changes* of models (when dealing with *dynamic* belief revision), we need a notion of proposition that is not confined to one model, but makes sense on *all* models:

A *doxastic proposition* is a map $\mathbf{P}$ assigning to each plausibility model $\mathbf{S}$ some S-proposition $\mathbf{P_S} \subseteq S$. We write $s \models_{\mathbf{S}} \mathbf{P}$, and say that the proposition $\mathbf{P}$ is true at $s \in \mathbf{S}$, iff $s \in (\mathbf{P})_{\mathbf{S}}$. We skip the subscript and write $s \models \mathbf{P}$ when the model is understood.

We denote by Prop the family of all doxastic propositions. All the Boolean operations on S-propositions as sets can be *lifted* pointwise to operations on Prop: in particular, we have the "always true" $\top$ and "always false" $\bot$ propositions, given by $(\bot)_{\mathbf{S}} := \emptyset, (\top)_{\mathbf{S}} := S$, negation $(\neg\mathbf{P})_{\mathbf{S}} := S \setminus \mathbf{P_S}$, conjunction $(\mathbf{P} \wedge \mathbf{Q})_{\mathbf{S}} := \mathbf{P_S} \cap \mathbf{Q_S}$, disjunction $(\mathbf{P} \vee \mathbf{Q})_{\mathbf{S}} := \mathbf{P_S} \cup \mathbf{Q_S}$ and all the other standard Boolean operators, including *infinitary* conjunctions and disjunctions. Similarly, we can define pointwise the *epistemic and (conditional) doxastic modalities*: $(K_a\mathbf{P})_{\mathbf{S}} := K_a\mathbf{P_S}$, $(B_a\mathbf{P})_{\mathbf{S}} := B_a\mathbf{P_S}$, $(B_a^{\mathbf{Q}}\mathbf{P})_{\mathbf{S}} := B_a^{\mathbf{Q_S}}\mathbf{P_S}$. It is easy to check that we have: $B_a\mathbf{P} = B_a^{\top}\mathbf{P}$. Finally, the relation of *entailment* $\mathbf{P} \models \mathbf{Q}$ between doxastic propositions is given pointwise by inclusion: $\mathbf{P} \models \mathbf{Q}$ iff $\mathbf{P_S} \subseteq \mathbf{Q_S}$ for all $\mathbf{S}$.

Safe Belief and the Defeasibility Theory of Knowledge

Ever since Plato's *identification of knowledge with "true justified (or justifiable) belief"* was shattered by Gettier's celebrated counterexamples (Gettier 1963), philosophers have been looking for the "missing ingredient" in the Platonic equation. Various authors identify this missing ingredient as "robustness" (Hintikka 1962), "indefeasibility" (Klein 1971; Lehrer 1990; Lehrer and Paxson 1969; Stalnaker 2006) or "stability" (Rott 2004). According to this *defeasibility theory of knowledge* (or "stability theory", as formulated by Rott), a belief counts as "knowledge" if it is *stable under belief revision with any new evidence*: "if a person has knowledge, than that person's justification must be sufficiently strong that it is not capable of being defeated by evidence that he does not possess" (Pappas and Swain 1978).

One of the problems is interpreting what "evidence" means in this context. There are at least two natural interpretations, each giving us a concept of "knowledge". The first, and the most common,[10] interpretation is to take it as meaning "any *true* information". The resulting notion of "knowledge" was formalized by Stalnaker in (2006), and defined there as follows: "an agent knows that φ if and only if φ is true, she believes that φ, and she continues to believe φ if any *true* information is received". This concept differs from the usual notion of knowledge ("Aumann knowledge") in Computer Science and Economics, by the fact that it does not satisfy the laws of the modal system S5 (in fact, negative introspection fails); Stalnaker

[10]This interpretation is the one virtually adopted by all the proponents of the defeasibility theory, from Lehrer to Stalnaker.

shows that the complete modal logic of this modality is the modal system S4.3. As we'll see, this notion ("Stalnaker knowledge") corresponds to what we call "safe belief" $\Box P$. On the other hand, another natural interpretation, considered by at least one author Rott (2004), takes "evidence" to mean "*any* proposition", i.e., to include possible *misinformation*: "real knowledge" should be robust even in the face of false evidence. As shown below, this corresponds to our "knowledge" modality KP, which could be called "absolutely unrevisable belief". This is a partition-based concept of knowledge, identifiable with "Aumann knowledge" and satisfying all the laws of S5. In other words, this last interpretation provides a perfectly decent "defeasibility" defense of S5 and of negative introspection!

In this paper, we adopt the pragmatic point of view of the formal logician: instead of debating which of the two types of "knowledge" is the real one, we simply formalize both notions in a common setting, compare them, axiomatize the logic obtained by combining them and use their joint strength to express interesting properties. Indeed, as shown below, conditional beliefs can be *defined* in terms of knowledge *only* if we combine both the above-mentioned types of "knowledge".

Knowledge as unrevisable belief. Observe that, for all propositions $\mathbf{P}$, we have

$$K_a\mathbf{Q} = \bigwedge_{\mathbf{P}} B_a^{\mathbf{P}}\mathbf{Q}$$

(where the conjunction ranges over *all* doxastic propositions), or equivalently, we have for every state s in every model $\mathbf{S}$:

$$s \models K_a\mathbf{Q} \quad \text{iff} \quad s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for all } \mathbf{P}. \tag{39.1}$$

This gives a characterization of *knowledge as "absolute" belief, invariant under any belief revision*: a given belief is "known" iff it cannot be revised, i.e., it would be still believed in any condition.[11] Observe that this resembles the defeasibility analysis of knowledge, but only if we adopt the *second interpretation* mentioned above (taking "evidence" to include misinformation). Thus, our "knowledge" is more robust than Stalnaker's: it resists any belief revision, not capable of being defeated by *any* evidence (including false evidence). This is a very "strong" notion of knowledge (implying "absolute certainty" and full introspection), which seems to us to fit better with the standard usage of the term in Computer Science literature. Also, unlike the one in Stalnaker (2006), our notion of knowledge *is negatively introspective*.

[11]This of course assumes agents to be "rational" in a sense that excludes "fundamentalist" or "dogmatic" beliefs, i.e., beliefs in unknown propositions but refusing any revision, even when contradicted by facts. But this "rationality" assumption is already built in our plausibility models, which satisfy an epistemically friendly version of the standard AGM postulates of rational belief revision. See Baltag and Smets (2006a) for details.

Another identity[12] that can be easily checked is:

$$K_a\mathbf{Q} = B_a^{\neg\mathbf{Q}}\mathbf{Q} = B_a^{\neg\mathbf{Q}}\bot \tag{39.2}$$

(where $\bot$ is the "always false" proposition). This captures in a different way the "absolute un-revisability" of knowledge: something is "known" if it is believed even if conditionalizing our belief with its negation. In other words, this simply expresses the *impossibility* of accepting its negation as evidence (since such a revision would lead to an inconsistent belief).

Safe belief. To capture "Stalnaker knowledge", we introduce the Kripke modality $\Box_a$ associated to the converse $\geq_a$ of the plausibility relation, going from any state s to all the states that are "at least as plausible" as s. For S-propositions $P \subseteq S$ over any given model $\mathbf{S}$, we put

$$\Box_a P := [\geq_a]P = \{s \in S : t \in P \text{ for all } t \leq_a s\},$$

and this induces pointwise an operator $\Box_a\mathbf{P}$ on doxastic propositions. We read $s \models \Box_a\mathbf{P}$ as saying that: *at state s, agent a's belief in* $\mathbf{P}$ *is safe*; or at *state s, a safely believes that* $\mathbf{P}$. We will explain this reading below, but first observe that: $\Box_a$ is an $S4$-modality (since $\geq_a$ is reflexive and transitive), but not necessarily $S5$; i.e., *safe beliefs are truthful* ($\Box_a\mathbf{P} \models \mathbf{P}$) and *positively introspective* ($\Box_a\mathbf{P} \models \Box_a\Box_a\mathbf{P}$), but not necessarily negatively introspective: in general, $\neg\Box_a\mathbf{P} \not\models \Box_a\neg\Box_a\mathbf{P}$.

Relations between knowledge, safe belief and conditional belief. First, *knowledge entails safe belief*

$$K_a\mathbf{P} \models \Box_a\mathbf{P}\,,$$

and *safe belief entails belief*

$$\Box_a\mathbf{P} \models B_a\mathbf{P}.$$

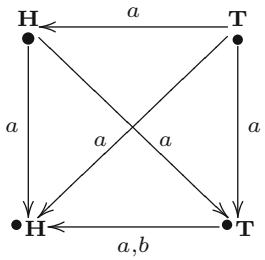
The last observation can be strengthened to characterize safe belief in a similar way to the above characterization (39.1) of knowledge (as belief invariant under any revision): *safe beliefs are precisely the beliefs which are persistent under revision with any true information*. Formally, this says that, for every state s in every model $\mathbf{S}$, we have

$$s \models \Box_a\mathbf{Q} \quad \text{iff:} \quad s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \mathbf{P} \tag{39.3}$$

[12]This identity corresponds to the definition of "necessity" in Stalnaker (1968) in terms of doxastic conditionals.

We can thus see that *safe belief coincides indeed with Stalnaker's notion of "knowledge"*, given by the first interpretation ("evidence as true information") of the defeasibility theory. As mentioned above, we prefer to keep the name "knowledge" for the strong notion (which gives absolute certainty), and call this weaker notion "safe belief": indeed, these are beliefs that are "safe" to hold, in the sense that no future learning of truthful information will force us to revise them.

Example 3 (Dangerous Knowledge). This starts with the situation in Example 1 (when none of the two agents has yet seen the face of the coin). Alice has to get out of the room for a minute, which creates an opportunity for Bob to quickly raise the cover in her absence and take a peek at the coin. He does that, and so he sees that the coin is Heads up. After Alice returns, she obviously doesn't know whether or not Bob took a peek at the coin, but she believes he didn't do it: taking a peek is against the rules of the game, and so she trusts Bob not to do that. The model is now rather complicated, so we only represent the MPM:



Let us call this model $\mathbf{S}'$. The actual state s'_1 is the one in the upper left corner, in which Bob took a peek and saw the coin Heads up, while the state t'_1 in the upper right corner represents the other possibility, in which Bob saw the coin lying Tails up. The two lower states s'_2 and t'_2 represent the case in which Bob *didn't take a peek*. Observe that the above drawing includes the (natural) assumption that Alice keeps her previous belief that the coin lies Heads up (since there is no reason for her to change her mind). Moreover, we assumed that she will keep this belief even if she'd be told that Bob took a peek: this is captured by the a-arrow from t'_1 to s'_1. This seems natural: Bob's taking a peek doesn't change the upper face of the coin, so it shouldn't affect Alice's prior belief about the coin.

In both Examples 1 and 3 above, Alice holds a *true belief* (at the real state) that the coin lies Heads up: the actual state satisfies $B_a\mathbf{H}$. In both cases, this true belief is *not knowledge* (since Alice doesn't know the upper face), but nevertheless in Example 1, this belief is *safe* (although it is *not known by the agent to be safe*): no additional truthful information (about the real state s) can force her to revise this belief. (To see this, observe that any *new* truthful information would reveal to Alice the real state s, thus confirming her belief that Heads is up.) So in the model $\mathbf{S}$ from Example 1, we have $s \models \Box_a\mathbf{H}$ (where s is the actual state). In contrast, in Example 2, Alice's belief (that the coin is Heads up), though true, is *not safe*. There is some piece

of correct information (about the real state s_1') which, if learned by Alice, would make her change this belief: we can represent this piece of correct information as the doxastic proposition $\mathbf{H} \to \mathbf{K_b H}$. It is easy to see that the actual state s_1' of the model $\mathbf{S}'$ satisfies the proposition $B_a^{\mathbf{H} \to \mathbf{K_b H}} \mathbf{T}$ (since $(\mathbf{H} \to K_b \mathbf{H})_{\mathbf{S}'} = \{s_1', t_1', t_2'\}$ and the minimal state in the set $s_1'(a) \cap \{s_1', s_1', t_2'\} = \{s_1', t_1', t_2'\}$ is t_2', which satisfies $\mathbf{T}$.) So, if given this information, Alice would come to wrongly believe that the coin is Tails up! This is an example of a *dangerous truth*: a true information whose learning can lead to wrong beliefs.

Observe that *an agent's belief can be safe without him necessarily knowing this* (in the "strong" sense of knowledge given by K): "safety" (similarly to "truth") is an *external* property of the agent's beliefs, that can be ascertained only by comparing his belief-revision system with reality. Indeed, *the only way* for an agent to *know a belief to be safe* is to actually *know it to be truthful*, i.e., to have actual *knowledge* (not just a belief) of its truth. This is captured by the valid identity

$$K_a \Box_a \mathbf{P} = K_a \mathbf{P}. \tag{39.4}$$

In other words: *knowing that something is safe to believe is the same as just knowing it to be true*. In fact, *all beliefs held by an agent "appear safe" to him*: in order to believe them, he has to believe that they are safe. This is expressed by the valid identity

$$B_a \Box_a \mathbf{P} = B_a \mathbf{P} \tag{39.5}$$

saying that: *believing that something is safe to believe is the same as just believing it*. Contrast this with the situation concerning "knowledge": in our logic (as in most standard doxastic-epistemic logics), we have the identity

$$B_a K_a \mathbf{P} = K_a \mathbf{P}. \tag{39.6}$$

So *believing that something is known is the same as knowing it*!

The Puzzle of the Perfect Believer. The last identity is well-known and has been considered "paradoxical" by many authors. In fact, the so-called "Paradox of the Perfect Believer" in Gochet and Gribomont (2006), Voorbraak (1993), Hoek (1993), Meyer and Hoek (1995), Williamson (2001), and Friedmann and Halpern (1994) is based on it. For a "strong" notion of belief as the one we have here ("belief" = belief with certainty), it seems reasonable to assume the following "axiom":

$$B_a \varphi \to B_a K_a \varphi. \tag{?}$$

Putting this together with (39.6) above, we get a paradoxical conclusion:

$$B_a \varphi \to K_a \varphi. \tag{?!}$$

So this leads to a triviality result: *knowledge and belief collapse to the same thing*, and *all beliefs are always true*! One solution to the "paradox" is to reject (?), as an (intuitive but) *wrong* "axiom". In contrast, various authors Friedmann and Halpern (1994); Hoek (1993); Voorbraak (1993); Williamson (2001) accept (?) and propose other solutions, e.g., giving up the principle of "negative introspection" for knowledge.

Our solution to the paradox, as embodied in the contrasting identities (39.5) and (39.6), combines the advantages of both solutions above: the "axiom" (?) is *correct if we interpret "knowledge" as safe belief* $\Box_a$, since then (?) becomes equivalent to identity (39.5) above; but then *negative introspection fails for this interpretation*! On the other hand, if we interpret "knowledge" as our K_a-modality then negative introspection holds; but then *the above "axiom" (?) fails*, and on the contrary we have the identity (39.6).

So, in our view, *the paradox of the perfect believer arises from the conflation of two different notions of "knowledge":* "Aumann" (partition-based) knowledge and "Stalnaker" knowledge (i.e., safe belief).

(Conditional) beliefs in terms of "knowledge" notions. An important observation is that *one can characterize/define (conditional) beliefs only in terms of our two "knowledge" concepts (K and* $\Box$*)*: For simple beliefs, we have

$$B_a\mathbf{P} \;=\; \tilde{K}_a\Box_a\mathbf{P} \;=\; \Diamond_a\Box_a\mathbf{P}\,,$$

recalling that $\tilde{K}_a\mathbf{P} = \neg K_a\neg\mathbf{P}$ is the Diamond modality for K_a, and $\Diamond_a\mathbf{P} = \neg\Box_a\neg\mathbf{P}$ is the Diamond for $\Box_a$.

The equivalence $B_a\mathbf{P} \;=\; \Diamond_a\Box_a\mathbf{P}$ has recently been observed by Stalnaker in (2006), who took it as the basis of a philosophical analysis of "belief" in terms of "defeasible knowledge" (i.e., safe belief). Unfortunately, this analysis does not apply to conditional belief: one can easily see that *conditional belief cannot be defined in terms of safe belief only*! However, one *can generalize* the identity $B_a\mathbf{P} = \tilde{K}_a\Box_a\mathbf{P}$ above, defining conditional belief in terms of *both our "knowledge" concepts*:

$$B_a^{\mathbf{P}}\mathbf{Q} \;=\; \tilde{K}_a\mathbf{P} \to \tilde{K}_a(\mathbf{P} \wedge \Box_a(\mathbf{P} \to \mathbf{Q}))\,. \tag{39.7}$$

Other Modalities and Doxastic Attitudes

From a modal logic perspective, it is natural to introduce the Kripke modalities $[>_a]$ and $[\cong_a]$ for the other important relations (strict plausibility and equiplausibility): For S-propositions $P \subseteq S$ over a given model $\mathbf{S}$, we put

$$[>_a]P \;:=\; \{s \in S : t \in P \text{ for all } t <_a s\},$$
$$[\cong_a]P \;:=\; \{s \in S : t \in P \text{ for all } t \cong_a s\},$$

and as before these pointwise induce corresponding operators on Prop. The intuitive meaning of these operators is not very clear, but they can be used to define other interesting modalities, capturing various "doxastic attitudes".

Weakly safe belief. We can define a *weakly safe belief* operator $\Box^{\text{weak}}\mathbf{P}$ in terms of the strict order by putting:

$$\Box_a^{\text{weak}}\mathbf{P} \;=\; \mathbf{P} \wedge [>_a]\mathbf{P}\,.$$

Clearly, this gives us the following truth clause:

$$s \models \Box_a^{\text{weak}}\ \mathbf{P} \text{ iff: } \ s \models \mathbf{P} \text{ and } t \models \mathbf{P} \text{ for all } t < s\,.$$

But a more useful characterization is the following:

$$s \models \Box_a^{\text{weak}}\mathbf{Q} \quad \text{iff:} \quad s \models \neg B_a^{\mathbf{P}}\neg\mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \mathbf{P}\,.$$

So "weakly safe beliefs" are *beliefs which (might be lost but) are never reversed (into believing the opposite) when revising with any true information*.

The unary revision operator. Using the strict plausibility modality, we can also define a unary "belief revision" modality $*_a$, which in some sense *internalizes the standard (binary) belief revision operator*, by putting:

$$*_a\mathbf{P} \;=\; \mathbf{P} \wedge [>_a]\neg\mathbf{P}\,.$$

This gives us the following truth clause:

$$s \models *_a\mathbf{P} \text{ iff } s \in s_a^{\mathbf{P}}\,.$$

It is easy to see that $*_a\mathbf{P}$ *selects from any given information cell* $s(a)$ *precisely those states that satisfy agent a's revised theory* s_a^P:

$$*_aP \cap s(a) = s_a^P\,.$$

Recall that $s_a^P \;=\; \text{Min}_{\leq_a} s(a) \cap P$ is the conditional appearance of s to a given P, representing the agent's "revised theory" (after revision with P) about s. This explains our interpretation: the proposition $*_a\mathbf{P}$ is a *complete description of the agent's P-revised "theory" about the current state*.

Another interesting identity is the following:

$$B_a^{\mathbf{P}}\mathbf{Q} \;=\; K_a(*_a\mathbf{P} \to \mathbf{Q})\,. \tag{39.8}$$

In other words: $\mathbf{Q}$ is *a conditional belief (given a condition* $\mathbf{P}$*) iff it is a known consequence of the agent's revised theory (after revision with* $\mathbf{P}$*)*.

Degrees of belief. Spohn's "degrees of belief" in Spohn (1988) were captured by Aucher (2003) and van Ditmarsch (2005) using logical operators $B_a^n\mathbf{P}$. Intuitively, 0-belief $B_a^0\mathbf{P}$ is the same as simple belief $B_a\mathbf{P}$; 1-belief $B_a^1\mathbf{P}$ means that $\mathbf{P}$ is believed conditional on learning that not all the 0-beliefs are true etc. Formally, this can be introduced e.g., by defining by induction a sequence of appearance maps s_a^n for all states s and natural numbers n:

$$s_a^0 = \mathrm{Min}_{\leq_a} s(a) \ , \ s_a^n = \mathrm{Min}_{\leq_a}\Big(s(a) \setminus \bigcup_{i<n} s_a^i\Big)$$

and defining

$$s \models B_a^n\mathbf{P} \ \text{ iff } \ t \models \mathbf{P} \text{ for all } t \in s_a^n.$$

A state s has degree of belief n if we have $s \in s_a^n$. An interesting observation is that the *finite degrees of belief* $B_a^n\mathbf{P}$ *can be defined using the unary revision operator* $*_a\mathbf{P}$ *and the knowledge operator* K_a (and, as a consequence, they can be defined using the plausibility operator $[>_a]\mathbf{P}$ and the knowledge operator). To do this, first put inductively:

$$b_a^0 := *_a\top \ , \ b_a^n := *_a\Big(\bigwedge_{m<n} \neg b_a^m\Big) \ \text{ for all } n \geq 1$$

and then put

$$B_a^n\mathbf{P} \ := \ \bigwedge_{m<n} \neg K_a(b_a^m \to \mathbf{P}) \wedge K_a(b_a^n \to \mathbf{P}) \, .$$

"Strong belief". Another important doxastic attitude can be defined in terms of knowledge and safe belief as:

$$Sb_a\mathbf{P} \ = \ B_a\mathbf{P} \wedge K_a(\mathbf{P} \to \Box_a\mathbf{P}) \, .$$

In terms of the plausibility order, it means that *all the* $\mathbf{P}$*-states in the information cell* $s(a)$ *of* s *are bellow (more plausible than) all the non-*$\mathbf{P}$ *states in* $s(a)$ (and that, moreover, *there are such* $\mathbf{P}$*-states in* $s(a)$). This notion is called "strong belief" by Battigalli and Siniscalchi (2002), while Stalnaker (1996) calls it "robust belief". Another characterization of strong belief is the following

$s \models Sb_a\mathbf{Q}$ iff:

$$s \models B_a\mathbf{Q} \text{ and } s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \neg K_a(\mathbf{P} \to \neg\mathbf{Q}).$$

In other words: *something is strong belief if 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.

The Logic of Conditional Beliefs

The logic CDL ("conditional doxastic logic") introduced in Baltag and Smets (2006a) is a logic of conditional beliefs, equivalent to the strongest logic considered in Board (2002). The *syntax* of CDL (without common knowledge and common belief operators[13]) is:

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a^{\varphi}\varphi$$

while the *semantics* is given by an *interpretation map* associating to each sentence φ of CDL a doxastic proposition $\|\varphi\|$. The definition is by induction, in terms of the obvious compositional clauses (using the doxastic operators $B_a^{\mathbf{P}}\mathbf{Q}$ defined above).

In this logic, *knowledge and simple (unconditional) belief are derived operators*, defined as abbreviations by putting $K_a\varphi := B_a^{\neg\varphi}\varphi$, $B_a\varphi := B_a^{\top}\varphi$ (where $\top := \neg(p \wedge \neg p)$ is some tautological sentence).

Proof system. In addition to the rules and axioms of propositional logic, the *proof system* of CDL includes the following:

Necessitation Rule:	From $\vdash \varphi$ infer $\vdash B_a^{\psi}\varphi$.
Normality:	$\vdash B_a^{\theta}(\varphi \to \psi) \to (B_a^{\theta}\varphi \to B_a^{\theta}\psi)$
Truthfulness of Knowledge:	$\vdash K_a\varphi \to \varphi$
Persistence of Knowledge:	$\vdash K_a\varphi \to B_a^{\theta}\varphi$
Full Introspection:	$\vdash B_a^{\theta}\varphi \to K_a B_a^{\theta}\varphi$,
	$\vdash \neg B_a^{\theta}\varphi \to K_a \neg B_a^{\theta}\varphi$
Success of Belief Revision:	$\vdash B_a^{\varphi}\varphi$
Minimality of Revision:	$\vdash \neg B_a^{\varphi}\neg\psi \to (B_a^{\varphi\wedge\psi}\theta \leftrightarrow B_a^{\varphi}(\psi \to \theta))$

Proposition 4 (Completeness and Decidability). *The above system is complete for MPMs (and so also for EPMs). Moreover, it is decidable and has the finite model property.*

Proof. The proof is essentially the same as in Board (2002). It is easy to see that the proof system above is equivalent to Board's strongest logic in Board (2002) (the one that includes axiom for full introspection), and that our models are equivalent to the "full introspective" version of the semantics in Board (2002). Q.E.D.

[13]The logic in Baltag and Smets (2006a) has these operators, but for simplicity we decided to leave them aside in this presentation.

The Logic of Knowledge and Safe Belief

The problem of finding a complete axiomatization of the logic of "defeasible knowledge" (safe belief) and conditional belief was posed as an *open question* in Board (2002). We answer this question here, by extending the logic CDL above to a complete logic $K\Box$ of *knowledge and safe belief*. Since this logic can *define* conditional belief, it is in fact equivalent to the logic whose axiomatization was required in Board (2002). Solving the question posed there becomes in fact trivial, once we observe that the higher-order definition of "defeasible knowledge" in Stalnaker (2006) and Board (2002) (corresponding to our identity (39.3) above) is in fact equivalent to our simpler, first-order definition of "safe belief" as a Kripke modality.

Syntax and semantics. The *syntax* of the logic $K\Box$ is:

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi \mid K_a\varphi$$

while the *semantics* over plausibility models is given as for CDL, by inductively defining an interpretation map from sentences to doxastic propositions, using the obvious compositional clauses. *Belief and conditional belief are derived operators here, defined as abbreviations*:

$$\begin{aligned} B_a^{\varphi}\psi \ &:= \ \tilde{K}_a\varphi \to \tilde{K}_a(\varphi \wedge \Box_a(\varphi \to \psi)) \,, \\ B_a\varphi \ &:= \ B_a^{\top}\varphi \,, \end{aligned}$$

where $\tilde{K}_a\varphi := \neg K_a\neg\varphi$ is the Diamond modality for K, and $\top = \neg(p \wedge \neg p)$ is some tautological sentence. So *the logic $K\Box$ is more expressive than CDL.*

Proof system. In addition to the rules and axioms of propositional logic, the *proof system* for the logic $K\Box$ includes the following:

- the Necessitation Rules for both K_a and $\Box_a$;
- the $S5$-axioms for K_a;
- the $S4$-axioms for $\Box_a$;
- $K_aP \to \Box_aP$;
- $K_a(P \vee \Box_aQ) \wedge K_a(Q \vee \Box_aP) \to K_aP \vee K_aQ$.

Theorem 5 (Completeness and Decidability). *The logic $K\Box$ is* (weakly) complete *with respect to MPMs (and so also with respect to EPMs). Moreover, it is* decidable *and has* the finite model property.

Proof. A *non-standard frame (model)* is a structure $(S, \geq_a, \sim_a)_a$ (together with a valuation, in the case of models) such that $\sim_a$ are equivalence relations, $\geq_a$ are preorders, $\geq_a \ \subseteq \ \sim_a$ and the restriction of $\geq_a$ to each $\sim_a$-equivalence class is connected. For a logic with two modalities, $\Box_a$ for $\geq_a$ and K_a for the relation $\sim_a$, we can use well-known results in Modal Correspondence Theory to see

that each of these semantic conditions corresponds to one of our modal axioms above. By general classical results on canonicity and modal correspondence,[14] we immediately obtain *completeness for non-standard models*. *Finite model property for these non-standard models* follows from the same general results. But every *finite* strict preorder relation $>$ is well-founded, and an MPM is nothing but a non-standard model whose strict preorders $>_a$ are well-founded. So *completeness for ("standard") MPMs* immediately follows. Then we can use Proposition 4 above to obtain *completeness for EPMs*. Finally, *decidability* follows, in the usual way, from finite model property together with *completeness* (with respect to a *finitary* proof system) and with the *decidability of model-checking on finite models*. (This last property is obvious, given the semantics.) Q.E.D.

"Dynamic" Belief Revision

The revision captured by conditional beliefs is of a *static*, purely *hypothetical*, nature. We *cannot* interpret B_a^φ as referring to the agent's revised beliefs about the situation *after revision*; if we did, then the "Success" axiom

$$\vdash B_a^\varphi \varphi$$

would *fail for higher-level beliefs*. To see this, consider a "Moore sentence"

$$\varphi := p \wedge \neg B_a p\,,$$

saying that some fact p holds but that agent a doesn't believe it. The sentence φ is consistent, so it may very well happen to be true. But agent a's beliefs about the situation after learning that φ was true *cannot* possibly include the sentence φ itself: after learning this sentence, agent *a knows p*, and so he believes p, contrary to what φ asserts. Thus, after learning φ, agent *a knows that* φ *is false now* (after the learning). This directly contradicts the Success axiom: far from believing the sentence after learning it to be true, the agent (knows, and so he correctly) believes that it has become false. There is nothing paradoxical about this: sentences may obviously change their truth values, due to our actions. Since learning the truth of a sentence is itself an action, it is perfectly consistent to have a case in which learning changes the truth value of the very sentence that is being learnt. Indeed, this is always the case with Moore sentences. Though not paradoxical, the existence of Moore sentences shows that the "Success" axiom does not correctly describe a rational agent's (higher-level) beliefs about what is the case after a new truth is being learnt.

[14] See e.g., Blackburn et al. (2001) for the general theory of modal correspondence and canonicity.

The only way to understand the "Success" axiom in the context of higher-level beliefs is to insist on the above-mentioned "static" interpretation of conditional belief operators B_a^{φ}, as expressing the agent's *revised belief* about how the state of the world *was before the revision*.

In contrast, a *belief update* is a dynamic form of belief revision, meant to capture the actual change of beliefs induced by learning: the updated belief is about the state of the world as it is *after the update*. As noticed in Gerbrandy (1999), Baltag et al. (1998), and Baltag and Moss (2004), the original model does not usually include enough states to capture all the epistemic possibilities that arise in this way. While in the previous section the models were kept unchanged during the revision, all the possibilities being already there (so that both the unconditional and the conditional beliefs *referred to the same model*), we now have to allow for belief updates that *change the original model*.

In Baltag and Moss (2004), it was argued that *epistemic events should be modeled in essentially the same way as epistemic states*, and this common setting was taken to be given by *epistemic Kripke models*. Since in this paper we enriched our state models with doxastic plausibility relations to deal with (conditional) beliefs, it is natural to follow Baltag and Moss (2004) into extending the similarity between actions and states to this setting, thus obtaining *(epistemic) action plausibility models*. The idea of such an extension was first developed in Aucher (2003) (for a different notion of plausibility model and a different notion of update product), then generalized in van Ditmarsch (2005), where many types of action plausibility models and notions of update product, that extend the so-called *Baltag-Moss-Solecki (BMS) update product* from Baltag et al. (1998) and Baltag and Moss (2004), are explored. But both these works are based on a *quantitative* interpretation of plausibility ordinals (as "degrees of belief"), and thus they define the various types of products using complex formulas of transfinite ordinal arithmetic, for which no intuitive justification is provided.

In contrast, our notion of update product is a *purely qualitative one*, based on a *simple and intuitive relational definition*: the simplest way to define a total pre-order on a Cartesian product, given total pre-orders on each of the components, is to use either the *lexicographic* or the *anti-lexicographic* order. We choose the second option, as the closest in spirit to the classical AGM theory: it gives *priority to the new, incoming information* (i.e., to "actions" in our sense).[15] We justify this choice by interpreting the action plausibility model as representing the agent's *"incoming" belief*, i.e., the *belief-updating event*, which *"performs" the update, by "acting" on the "prior" beliefs* (as given in the state plausibility model).

[15]This choice can be seen as a generalization of the so-called *"maximal-Spohn" revision*.

Action Models

An *action plausibility model*[16] (APM, for short) is a plausibility frame $(\Sigma, \leq_a)_{a\in\mathcal{A}}$ together with a *precondition map* $\text{pre} : \Sigma \to \mathsf{Prop}$, associating to each element of Σ some doxastic proposition pre_σ. We call the elements of Σ *(basic) doxastic actions* (or "events"), and we call pre_σ the *precondition* of action σ. The basic actions $\sigma \in \Sigma$ are taken to represent *deterministic belief-revising actions* of a particularly simple nature. Intuitively, the precondition defines the *domain of applicability* of action σ: it can be executed on a state s iff s satisfies its precondition. The relations $\leq_a$ give the agents' beliefs about which actions are more plausible than others.

To model *non-determinism*, we introduce the notion of epistemic program. A *doxastic program over a given action model* $\mathbf{\Sigma}$ (or $\mathbf{\Sigma}$*-program*, for short) is simply a *set* $\Gamma \subseteq \Sigma$ of doxastic actions. We can think of doxastic programs as non-deterministic actions: each of the basic actions $\gamma \in \Gamma$ is a possible "deterministic resolution" of Γ. For simplicity, when $\Gamma = \{\gamma\}$ is a singleton, we ambiguously identify the program Γ with the action γ.

Observe that $\mathbf{\Sigma}$-programs $\Gamma \subseteq \Sigma$ are formally the "dynamic analogues" of $\mathbf{S}$-propositions $P \subseteq S$. So the dynamic analogue of the conditional doxastic appearance s_a^P (representing agent a's revised theory about state s, after revision with proposition P) is the set σ_a^Γ.

Interpretation: beliefs about changes encode changes of beliefs. The name "doxastic actions" might be a bit misleading, and from a philosophical perspective Johan van Benthem's term "doxastic events" seems more appropriate. The elements of a plausibility model do not carry information about agency or intentionality and cannot represent "real" actions in all their complexity, but only the *doxastic changes* induced by these actions: each of the nodes of the graph represents a *specific kind of change of beliefs (of all the agents)*. As in Baltag and Moss (2004), we only deal here with pure "belief changes", i.e., actions that do not change the "ontic" facts of the world, but only the agents' beliefs.[17] Moreover, we think of these as *deterministic* changes: there is at most one output of applying an action to a state.[18] Intuitively, the precondition defines the *domain of applicability* of σ: this action can be executed on a state s iff s satisfies its precondition. The plausibility pre-orderings $\leq_a$ give *the agents' conditional beliefs about the current action*. But this should be interpreted as *beliefs about changes*, that *encode changes of beliefs*. In this sense, we use such

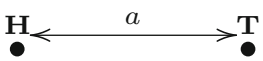
[16] Van Benthem calls this an "event model".

[17] We stress this is a minor restriction, and it is very easy to extend this setting to "ontic" actions. The only reason we stick with this restriction is that it simplifies the definitions, and that it is general enough to apply to all the actions we are interested here, and in particular to all *communication actions*.

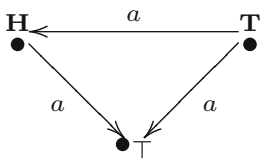
[18] As in Baltag and Moss (2004), we will be able to represent non-deterministic actions as sums (unions) of deterministic ones.

"beliefs about actions" as a way to represent doxastic changes: the information about how the agent changes her beliefs is captured by our action plausibility relations. So we read $\sigma <_a \sigma'$ as saying that: if agent a is informed that either σ or σ' is currently happening, then she cannot distinguish between the two, but she believes that σ is in fact happening. As already mentioned, doxastic programs $\Gamma \subseteq \Sigma$ represent *non-deterministic* changes of belief. Finally, for an action σ and a program Γ, the program σ_a^Γ represents *the agent's revised theory (belief) about the current action σ after "learning" that (one of the deterministic resolutions γ in) Γ is currently happening*.

Example 4 (Private "Fair-Game" Announcements). Let us consider the *action* that produced the situation represented in Example 2 above. In front of Alice, Bob looked at the coin, in such a way that (it was common knowledge that) only he saw the face. In the DEL literature, this is sometimes known as a "fair game" announcement: everybody is commonly aware that an insider (or a group of insiders) privately learns some information. It is "fair" since the outsiders are *not "deceived" in any way*: e.g., in our example, Alice knows that Bob looks at the coin (and he knows that she knows etc.). In other words, Bob's looking at the coin is not an "illegal" action, but one that obeys the (commonly agreed) "rules of the game". To make this precise, let us assume that this is happening in such a way that Alice has no strong beliefs about which of the two possible actions (Bob-seeing-Heads-up and Bob-seeing-Tails-up) is actually happening. Of course, we assumed that before this, she already believed that the coin lies Heads up, but apart from this we now assume that *the way the action (of "Bob looking") is happening gives her no indication of what face he is seeing*. We represent these actions using a two-node plausibility model $\mathbf{\Sigma}_2$ (where as in the case of state models we draw arrows for the converse plausibility relations $\geq_a$, disregarding all the loops):



Example 5 (Fully Private Announcements). Let us consider the *action* that produced the situation represented in Example 3 above. This was the action of Bob taking a peek at the coin, while Alice was away. Recall that we assumed that Alice *believed that nothing was really happening* in her absence (since she assumed Bob was playing by the rules), though obviously she *didn't know* this (that nothing was happening). In the DEL literature, this action is usually called *a fully private announcement*: Bob learns which face is up, while the outsider Alice believes nothing of the kind is happening. To represent this, we consider an action model $\mathbf{\Sigma}_3$ consisting of three "actions": the actual action σ in which Bob takes a peek and sees the coin lying Heads up; the alternative possible action ρ is the one in which Bob sees the coin lying Tails up; finally, the action τ is the one in which "nothing is really happening" (as Alice believes). The plausibility model $\mathbf{\Sigma}_3$ for this action is:



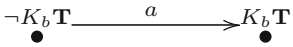
Here, the action σ is the one in the upper left corner, having precondition **H**: indeed, this can happen iff the coin is really lying Heads up; similarly, the action ρ in the upper right corner has precondition **T**, since it can only happen iff the coin is Tails up. Finally, the action τ is the lower one, having as precondition the "universally true" proposition $\top$: indeed, this action can always happen (since in it, nothing is really happening!). The plausibility relations reflect the agents' beliefs: in each case, both Bob and Charles know exactly what is happening, so their local plausibility relations are the identity (and thus we draw no arrows for them). Alice believes nothing is happening, so τ is the most plausible action for her (to which all her arrows are pointing); so she keeps her belief that **H** is the case, thus considering σ as more plausible than ρ.

Examples of doxastic programs. Consider the program $\Gamma = \{\sigma, \rho\} \subseteq \mathbf{\Sigma}_3$ over the action model $\mathbf{\Sigma_3}$ from Example 5. The program Γ represents *the action of "Bob taking a peek at the coin", without any specification of which face he is seeing*. Although expressed in a non-deterministic manner (as a collection of two possible actions, σ and ρ), this program corresponds is in fact *deterministic*, since in each possible state only one of the actions σ or ρ can happen: there is no state satisfying both **H** and **T**. The whole set Σ gives another doxastic program, one that is really non-deterministic: it represents the non-deterministic choice of Bob between taking a peek and not taking it.

Appearance of actions and their revision: Examples. As an example of an agent's "theory" about an action, consider the appearance of action ρ to Alice: $\rho_a = \{\tau\}$. Indeed, if ρ happens (Bob taking a peek and sees the coin is Tails up), Alice believes that τ (i.e., nothing) is happening: this is the "apparent action", as far as Alice is concerned. As an example of a "revised theory" about an action, consider the conditional appearance ρ_a^Γ of ρ to Alice given the program $\Gamma = \{\sigma, \rho\}$ introduced above. It is easy to see that we have $\rho_a^\Gamma = \{\sigma\}$. This captures our intuitions about Alice's revised theory: if, while ρ was happening, she were told that Bob took a peek (i.e., she'd revise with Γ), then she would believe that he saw the coin lying Heads up (i.e., that σ happened).

Example 6 (Successful Lying). Suppose now that, *after* the previous action, i.e., after we arrived in the situation described in Example 3, Bob sneakily announces: "I took a peek and saw the coin was lying *Tails up*". We formalize the content of this announcement as $K_b\mathbf{T}$, i.e., saying that "Bob knows the coin is lying Tails up". This is a *public announcement*, but *not a truthful one* (though it does convey some truthful information): it is a *lie*! We assume it is in fact a *successful lie*: it is common

knowledge that, even after Bob admitted having taken a peek, Alice still believes him. This action is given by the *left node* in the following model $\mathbf{\Sigma}_4$:



The Action-Priority Update

We are ready to define our *update operation*, representing the way an action from a (action) plausibility model $\mathbf{\Sigma} = (\Sigma, \leq_a, \text{pre})_{a\in\mathcal{A}}$ "acts" on an input-state from a given (state) plausibility model $\mathbf{S} = (S, \leq_a, \|\cdot\|)_{a\in\mathcal{A}}$. We denote the updated state model by $\mathbf{S}\otimes\mathbf{\Sigma}$, and call it the *update product* of the two models. The construction is similar to a point to the one in Baltag et al. (1998) and Baltag and Moss (2004), and thus also somewhat similar to the ones in Aucher (2003) and van Ditmarsch (2005). In fact, the set of updated states, the updated valuation and the updated indistinguishability relation are *the same* in these constructions. The main difference lies in our definition of the *updated plausibility relation*, via the *Action Priority Rule*.

Updating Single-Agent Models: The Anti-lexicographic Order

To warm up, let us first define the update product for the single-agent case. Let $\mathbf{S} = (S, \leq, \|\cdot\|)$ be a single-agent plausibility state model and let $\mathbf{\Sigma} = (\Sigma, \leq, \text{pre})$ be a single-agent plausibility action model.

We represent the *states of the updated model* $\mathbf{S}\otimes\mathbf{\Sigma}$ as pairs (s, σ) of input-states and actions, i.e., as elements of the Cartesian product $S \times \Sigma$. This reflects that the basic actions in our action models are assumed to be *deterministic*: For a given input-state and a given action, there can only be at most one output-state. More specifically, we select the pairs which are *consistent, in the sense that the input-state satisfies the precondition of the action*. This is natural: the precondition of an action is a specification of its domain of applicability. So the *set of states* of $\mathbf{S}\otimes\mathbf{\Sigma}$ is taken to be

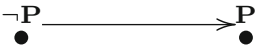
$$S \otimes \Sigma := \{(s, \sigma) : s \models_{\mathbf{S}} \text{pre}(\sigma)\}.$$

The *updated valuation* is essentially given by the *original valuation* from the input-state model: For all $(s, \sigma) \in S \otimes \Sigma$, we put $(s, \sigma) \models p$ iff $s \models p$. This "conservative" way to update the valuation expresses the fact that we only consider here actions that are *"purely doxastic"*, i.e., pure "belief changes", that do not affect the ontic "facts" of the world (captured here by atomic sentences).

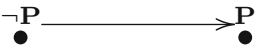
We still need to define the updated plausibility relation. To motivate our definition, we first consider two examples:

Example 7 (A Sample Case). Suppose that we have two states $s, s' \in \mathbf{S}$ such that $s < s'$, $s \models \neg\mathbf{P}$, $s' \models \mathbf{P}$. This means that, if given the supplementary information that the real state is either s or s', the agent believes $\neg\mathbf{P}$:

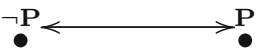
Suppose then an event happens, in whose model there are two actions σ, σ' such that $\sigma > \sigma'$, $\mathrm{pre}_\sigma = \neg\mathbf{P}$, $\mathrm{pre}_{\sigma'} = \mathbf{P}$. In other words, if given the information that either σ or σ' is happening, the agent believes that σ' is happening, i.e., she believes that $\mathbf{P}$ is learnt. This part of the model behaves just like a *soft public announcement* of $\mathbf{P}$:



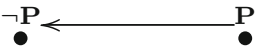
Naturally, we expect the agent to *change her belief* accordingly, i.e., her updated plausibility relation on states should now go the other way:



Example 8 (A Second Sample Case). Suppose the initial situation was the same as above, but now the two actions σ, σ' are assumed to be equi-plausible: $\sigma \cong \sigma'$. This is a *completely unreliable announcement of* $\mathbf{P}$, in which the veracity and the falsity of the announcement are equally plausible alternatives:



In the AGM paradigm, it is natural to expect the agents to *keep their original beliefs* unchanged after this event:



The anti-lexicographic order. Putting the above two sample cases together, we conclude that the updated plausibility relation should be the *anti-lexicographic preorder relation* induced on pairs $(s, \sigma) \in S \times \Sigma$ by the preorders on $\mathbf{S}$ and on $\mathbf{\Sigma}$, i.e.:

$$(s, \sigma) \leq (s', \sigma') \quad \text{iff:} \quad \text{either } \sigma < \sigma', \text{ or else } \sigma \cong \sigma' \text{ and } s \leq s'.$$

In other words, the updated plausibility order gives *"priority" to the action plausibility relation, and apart from this it keeps as much as possible the old order*. This reflects our commitment to an AGM-type of revision, in which the new information has priority over old beliefs. The "actions" represent here the "new information", although (unlike in AGM) this information comes in *dynamic form* (as action plausibility order), and so it is not fully reducible to its propositional content (the action's precondition). In fact, this is a generalization of one of the belief-revision policies encountered in the literature (the so-called "*maximal-Spohn revision*"). But, in the context of our qualitative (conditional) interpretation of plausibility models, we will argue below that this is essentially the only reasonable option.

Updating Multi-agent Models: The General Case

In the multi-agent case, the construction of *the updated state space and updated valuation is the same as above*. But for the updated plausibility relation we need to take into account *a third possibility*: the case when either the initial states or the actions are *distinguishable*, belonging to *different information cells*.

Example 9 (A Third Sample Case). Suppose that we have two states $s, s' \in \mathbf{S}$ such that $s \models \neg\mathbf{P}$, $s' \models \mathbf{P}$, but $s \not\sim_a s'$ are *distinguishable* (i.e., non-comparable):

This means that, if given the supplementary information that the real state is either s or s', the agent immediately *knows* which of the two is the real states, and thus *she knows whether* $\mathbf{P}$ *holds or not*. It is obvious that, after any of the actions considered in the previous two examples, a perfect-recall agent *will continue to know* whether $\mathbf{P}$ held or not, and so *the output-states after* σ *and* σ' *will still be distinguishable (non-comparable)*.

The "Action-Priority" Rule. Putting this together with the other sample cases, we obtain our update rule, in full generality:

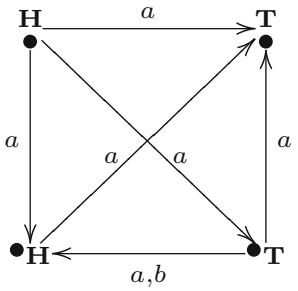
$$(s, \sigma) \leq_a (s', \sigma') \text{ iff either } \sigma <_a \sigma' \text{ and } s \sim_a s', \text{ or else } \sigma \cong_a \sigma' \text{ and } s \leq_a s'$$

We regard this construction as the most natural analogue in a belief-revision context of the similar notion in Baltag and Moss (2004) and Baltag et al. (1998). Following a suggestion of Johan van Benthem, we call this the Action-Priority Update Rule.

Sanity check: Examples 2 and 3 revisited. To check the correctness of our update operation, take first the update product $\mathbf{S} \otimes \mathbf{\Sigma}_2$ of the model $\mathbf{S}$ in Example 1 from the previous section with the action model $\mathbf{\Sigma}_2$ in Example 4 from the previous section. As predicted, the resulting state model is isomorphic to the model $\mathbf{W}$ from

Example 2. Similarly, if $\mathbf{\Sigma}_3$ is the action model from Example 5, then we can see that the product $\mathbf{S}\otimes\mathbf{\Sigma}_3$ is isomorphic to the state model $\mathbf{S}'$ from Example 3.

"In-sanity check": Successful lying. Applying the action model $\mathbf{\Sigma}_4$ in Example 6, representing the "successful lying" action, to the state model $\mathbf{S}'$ from Example 3, we obtain indeed the intuitively correct output of "successful lying", namely the following model $\mathbf{S}'\otimes\mathbf{\Sigma}_4$:



Interpretation. As its name makes explicit, the Action-Priority Rule gives "priority" to the *action* plausibility relation. This is not an arbitrary choice, but it is motivated by our specific interpretation of action models, as embodied in our Motto above: *beliefs about changes* (i.e., the action plausibility relations) *are nothing but ways to encode changes of belief* (i.e., reversals of the original plausibility order). So *the (strict) order on actions encodes changes of order on states*. The Action-Priority Rule is a consequence of this interpretation: it just says that a strong plausibility order $\sigma <_a \sigma'$ on actions corresponds indeed to a change of ordering, (from whatever the ordering was) between the original (indistinguishable) input-states $s \sim_a s'$, to the order $(s,\sigma) <_a (s',\sigma')$ between output-states; while equally plausible actions $\sigma \cong_a \sigma'$ will leave the initial ordering unchanged: $(s,\sigma) \leq_a (s',\sigma')$ iff $s \leq_a s'$. Giving priority to action plausibility does not in any way mean that the agent's belief in actions is stronger than her belief in states; it just captures the fact that, at the time of updating with a given action, *the belief about the action is what is actual, it is the current belief about what is going on, while the beliefs about the input-states are in the past*.[19]

In a nutshell: *the doxastic action is the one that changes the initial doxastic state, and not vice-versa*. The belief update induced by a given action is nothing but an update with the (presently) believed action. If the believed action σ requires the agent to revise some past beliefs, then so be it: this is the whole point of believing σ, namely to use it to revise one's past beliefs. For example, in a successful lying, the action plausibility relation makes the hearer believe that the speaker is telling

[19]Of course, *at a later moment*, the above-mentioned belief about action (*now* belonging to the past) might be itself revised. But this is another, *future update*.

the truth; so she'll accept this message (unless contradicted by her knowledge), and change her past beliefs appropriately: this is what makes the lying successful.

Action-priority update generalizes product update. Recall the definition of the epistemic indistinguishability relation $\sim_a$ in a plausibility model: $s \sim_a s'$ iff either $s \leq_a s'$ or $s' \leq_a s$. It is easy to see that the Action Priority Update implies the familiar update rule from Baltag et al. (1998) and Baltag and Moss (2004), known in Dynamic Epistemic Logic as the "product update":

$$(s, \sigma) \sim_a (s', \sigma') \text{ iff } s \sim_a s' \text{ and } \sigma \sim_a \sigma'.$$

Program transitions. For every state model $\mathbf{S}$, every program $\Gamma \subseteq \Sigma$ over an action model $\boldsymbol{\Sigma}$ induces a transition relation $\overset{\Gamma}{\rightarrow}_{\mathbf{S}} \subseteq \mathbf{S} \times (S \otimes \Sigma)$ from $\mathbf{S}$ to $\mathbf{S} \otimes \boldsymbol{\Sigma}$, given by:

$$s \overset{\Gamma}{\rightarrow}_{\mathbf{S}} (s', \gamma) \text{ iff } s = s', (s, \gamma) \in S \otimes \Sigma \text{ and } \gamma \in \Gamma .$$

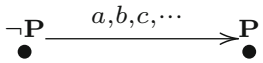
Simulating Various Belief-Revision Policies

We give here three examples of *multi-agent belief-revision policies* that can be simulated by our product update: *truthful public announcements of "hard facts"*, *"lexicographic update"* and *"conservative upgrade"*. They were all introduced by van Benthem in (2007), as multi-agent versions of revision operators previously considered by Rott (1989) and others.

Public announcements of "hard facts". A *truthful public announcement* $!\mathbf{P}$ *of some "hard fact"* $\mathbf{P}$ is not really about belief revision, but about the learning of *certified true information*: it establishes *common knowledge* that $\mathbf{P}$ was the case. This is the action described in van Benthem (2007) as (public) "belief change under hard facts". As an operation on models, this is described in van Benthem (2007) as taking any state model $\mathbf{S}$ and *deleting all the non-*$\mathbf{P}$ *states, while keeping the same indistinguishability and plausibility relations between the surviving states*. In our setting, the corresponding action model consists of only one node, labeled with $\mathbf{P}$. It is easy to see that the above operation on models can be exactly "simulated" by taking the anti-lexicographic product update with this one-node action model.

Public announcements of "soft facts": The "lexicographic upgrade". To allow for "soft" belief revision, an operation $\Uparrow\mathbf{P}$ was introduced in van Benthem (2007), essentially adapting to public announcements the 'lexicographic' policy for belief revision described in Rott (1989). This operation, called "lexicographic update" consists of changing the current plausibility order on any given state model as follows: *every* $\mathbf{P}$*-world becomes "better" (more plausible) than all* $\neg\mathbf{P}$*-worlds in*

*the same information cell, and within the two zones (***P** *and* ¬**P***), the old ordering remains*. In our setting, this action corresponds to the following local plausibility action model:



Taking the anti-lexicographic update product with this action will give an exact "simulation" of the lexicographic upgrade operation.

"Conservative upgrade". The operation ↑**P** of "conservative upgrade", also defined in van Benthem (2007), changes any model as follows: *in every information cell, the best* **P***-worlds become better than all the worlds in that cell* (i.e., in every cell the most plausible **P**-states become the most plausible overall in that cell), *and apart from that, the old order remains*. In the case of a system *with only one agent*, it is easy to see that we have $\uparrow \mathbf{P} = \Uparrow(*_a \mathbf{P})$, where $*_a$ is the unary "revision modality" introduced in the previous section. In the case of a set $\mathcal{A} = \{1, \cdots, n\}$ with $n > 1$ agents, we can simulate ↑**P** using a model with 2^n actions $\{\uparrow_I \mathbf{P}\}_{I \subseteq \mathcal{A}}$, with

$$\mathrm{pre}_{\uparrow_I \mathbf{P}} = \bigwedge_{i \in I} *_i \mathbf{P} \wedge \bigwedge_{j \notin I} \neg *_j \mathbf{P},$$

$$\uparrow_I \mathbf{P} \leq_k \uparrow_J \mathbf{P} \quad \text{iff} \quad J \cap \{k\} \subseteq I.$$

Operations on Doxastic Programs

First, we introduce *dynamic modalities*, capturing *the "weakest precondition" of a program* Γ. These are the natural analogues of the PDL modalities for our program transition relations $\xrightarrow{\Gamma}$ between models.

Dynamic modalities. Let **Σ** be some action plausibility model and $\Gamma \subseteq \Sigma$ be a doxastic model over **Σ**. For every doxastic proposition **P**, we define a doxastic proposition [Γ]**P** given by

$$([\Gamma]\mathbf{P})_{\mathbf{S}} := [\xrightarrow{\Gamma}_{\mathbf{S}}]\mathbf{P}_{\mathbf{S}} = \{s \in S : \forall t \in S \otimes \Sigma \, (s \xrightarrow{\Gamma}_{\mathbf{S}} t \Rightarrow t \models_{\mathbf{S} \otimes \mathbf{\Sigma}} \mathbf{P})\}.$$

For *basic doxastic actions* $\sigma \in \Sigma$, we define the dynamic modality $[\sigma]$ via the above-mentioned identification of actions σ with singleton programs $\{\sigma\}$:

$$([\sigma]\mathbf{P})_{\mathbf{S}} := ([\{\sigma\}]\mathbf{P})_{\mathbf{S}} = \{s \in S : \text{ if } (s, \sigma) \in \mathbf{S} \otimes \mathbf{\Sigma} \text{ then } (s, \sigma) \in \mathbf{P}_{\mathbf{S} \otimes \mathbf{\Sigma}}\}.$$

The dual (Diamond) modalities are defined as usually: $\langle \Gamma \rangle \mathbf{P} := \neg[\Gamma]\neg\mathbf{P}$.

We can now introduce operators on doxastic programs that are the analogues of the *regular operations* of PDL.

Sequential composition. The *sequential composition* $\mathbf{\Sigma};\mathbf{\Delta}$ of two action plausibility models $\mathbf{\Sigma} = (\Sigma, \leq_a, \text{pre})$, $\mathbf{\Delta} = (\Delta, \leq_a, \text{pre})$ is defined as follows:

- the set of basic actions is the Cartesian product $\Sigma \times \Delta$
- the preconditions are given by $\text{pre}_{(\sigma,\delta)} := \langle\sigma\rangle\, \text{pre}_\delta$
- the plausibility order is given by putting $(\sigma, \delta) \leq_a (\sigma', \delta')$ iff: either $\sigma <_a \sigma'$ and $\delta \sim_a \delta'$, or else $\sigma \cong_a \sigma'$ and $\delta \leq_a \delta'$.

We think of (σ, δ) as the action of *performing first* σ *then* δ, and thus use the notation

$$\sigma;\delta := (\sigma, \delta).$$

We can extend this notation to doxastic programs, by defining the *sequential composition of programs* $\Gamma \subseteq \Sigma$ and $\Lambda \subseteq \Delta$ to be a program $\Gamma;\Lambda \subseteq \Sigma;\Delta$ over the action model $\mathbf{\Sigma};\mathbf{\Delta}$, given by:

$$\Gamma;\Lambda := \{(\gamma, \lambda) : \gamma \in \Gamma, \lambda \in \Lambda\}.$$

It is easy to see that this behaves indeed like a sequential composition:

Proposition 12. *For every state plausibility model* $\mathbf{S}$*, action plausibility models* $\mathbf{\Sigma}$ *and* $\mathbf{\Delta}$*, and programs* $\Gamma \subseteq \Sigma$*,* $\Lambda \subseteq \Delta$*, we have the following:*

1. *The state plausibility models* $(\mathbf{S} \otimes \mathbf{\Sigma}) \otimes \mathbf{\Delta}$ *and* $\mathbf{S} \otimes (\mathbf{\Sigma};\mathbf{\Delta})$ *are isomorphic, via the canonical map* $F : (S \otimes \Sigma) \otimes \Delta \to S \otimes (\Sigma;\Delta)$ *given by*

$$F(\,((s, \sigma), \delta)\,) \;:=\; (s, (\sigma, \delta)).$$

2. *The transition relation for the program* $\Gamma;\Delta$ *is the relational composition of the transition relations for* Γ *and for* Δ *and of the isomorphism map F:*

$s \overset{\Gamma;\Delta}{\to}_{\mathbf{S}} s'$ iff there exist $w, t \in S \otimes \Sigma$ such that

$$s \overset{\Gamma}{\to}_{\mathbf{S}} w \overset{\Delta}{\to}_{\mathbf{S}\otimes\mathbf{\Sigma}} t \text{ and } F(t) = s'.$$

Union (non-deterministic choice). If $\mathbf{\Sigma} = (\Sigma, \leq_a, \text{pre})$ and $\mathbf{\Delta} = (\Delta, \leq'_a, \text{pre}')$ are two action plausibility models, their *disjoint union* $\mathbf{\Sigma} \sqcup \mathbf{\Delta}$ is simply given by taking as set of states the disjoint union $\Sigma \sqcup \Delta$ of the two sets of states, taking as plausibility order the disjoint union $\leq_a \sqcup \leq'_a$ and as precondition map the disjoint union $\text{pre} \sqcup \text{pre}'$ of the two precondition maps. If $\Gamma \subseteq \Sigma$ and $\Lambda \subseteq \Delta$ are doxastic programs over the two models, we define their *union* to be the program over the model $\mathbf{\Sigma} \sqcup \mathbf{\Delta}$ given by the disjoint union $\Gamma \sqcup \Lambda$ of the sets of actions of the two programs.

Again, it is easy to see that *this behaves indeed like a non-deterministic choice operator*:

Proposition 13. *Let* $i_1 : \Sigma \to \Sigma \sqcup \Delta$ *and* $i_2 : \Delta \to \Sigma \sqcup \Delta$ *be the two canonical injections. Then the following are equivalent:*

- $s \overset{\Gamma \sqcup \Delta}{\to}_{\mathbf{S}} s'$
- *there exists* t *such that:*

$$\text{either } s \overset{\Gamma}{\to}_{\mathbf{S}} t \text{ and } i_1(t) = s', \text{ or else } s \overset{\Delta}{\to}_{\mathbf{S}} t \text{ and } i_2(t) = s'.$$

Other operators. *Arbitrary unions* $\bigsqcup_i \Gamma_i$ can be similarly defined, and then one can define *iteration* $\Gamma^* := \bigsqcup_i \Gamma^i$ (where $\Gamma^0 = !\top$ and $\Gamma^{i+1} = \Gamma; \Gamma^i$).

The Laws of Dynamic Belief Revision

The "laws of dynamic belief revision" are the fundamental equations of Belief Dynamics, allowing us *to compute future doxastic attitudes from past ones*, given the doxastic events that happen in the meantime. In modal terms, these can be stated as "reduction laws" for inductively computing dynamic modalities $[\Gamma]\mathbf{P}$, by reducing them to modalities $[\Gamma']\mathbf{P}'$ in which either the propositions $\mathbf{P}'$ or the programs Γ' have *lower complexity*.

The following immediate consequence of the definition of $[\Gamma]\mathbf{P}$ allows us to reduce modalities for non-deterministic programs Γ to the ones for their deterministic resolutions $\gamma \in \Gamma$:

Deterministic Resolution Law. For every program $\Gamma \subseteq \Sigma$, we have

$$[\Gamma]\mathbf{P} \;=\; \bigwedge_{\gamma \in \Gamma} [\gamma]\mathbf{P}\,.$$

So, for our other laws, we can restrict ourselves to *basic actions* in Σ.

The Action-Knowledge Law. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]K_a\mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma' \sim_a \sigma} K_a[\sigma']\mathbf{P}\,.$$

This Action-Knowledge Law is essentially the same as in Baltag et al. (1998) and Baltag and Moss (2004): *a proposition* $\mathbf{P}$ *will be known after a doxastic event iff, whenever the event can take place, it is known that* $\mathbf{P}$ *will become true after all events that are indistinguishable from the given one*.

The Action-Safe-Belief Law. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]\Box_a \mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma' <_a \alpha} K_a[\sigma']\mathbf{P} \wedge \bigwedge_{\sigma'' \cong_a \sigma} \Box_a[\sigma'']\mathbf{P}\,.$$

This law embodies the essence of the Action-Priority Rule: *a proposition* **P** *will be safely believed after a doxastic event iff, whenever the event can take place, it is known that* **P** *will become true after all more plausible events and in the same time it is safely believed that* **P** *will become true after all equi-plausible events*.

Since we took knowledge and safe belief as the basis of our static logic $K\Box$, the above two laws are the "fundamental equations" of our theory of dynamic belief revision. But note that, as a consequence, one can obtain *derived laws for (conditional) belief* as well. Indeed, using the above-mentioned characterization of conditional belief in terms of K and $\Box$, we obtain the following:

The Derived Law of Action-Conditional-Belief. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]B_a^{\mathbf{P}}\mathbf{Q} \;=\; \mathrm{pre}_\sigma \to \bigvee_{\Gamma \subseteq \Sigma} \Big(\bigwedge_{\gamma \in \Gamma} \tilde{K}_a\langle\gamma\rangle\mathbf{P} \wedge \bigwedge_{\gamma' \notin \Gamma} \neg\tilde{K}_a\langle\gamma'\rangle\mathbf{P} \wedge B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]\mathbf{Q} \Big)\,.$$

This derived law, a version of which was first introduced in Baltag and Smets (2006c) (where it was considered a fundamental law), allows us to predict future conditional beliefs from current conditional beliefs.

To explain the meaning of this law, we re-state it as follows: For every $s \in \mathbf{S}$ and $\sigma \in \mathbf{\Sigma}$, we have:

$$s \models [\sigma]B_a^{\mathbf{P}}\mathbf{Q} \;\text{ iff }\; s \models \mathrm{pre}_\sigma \to B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]\mathbf{Q}\,,$$
$$\text{where }\; \Gamma = \{\gamma \in \Sigma : s \models_{\mathbf{S}} \tilde{K}_a\langle\gamma\rangle\mathbf{P}\}\,.$$

It is easy to see that this "local" (state-dependent) version of the reduction law is equivalent to the previous (state-independent) one. The set Γ encodes the extra information about the current action that is given to the agent by the context s and by the post-condition **P**; while σ_a^Γ is the action's *post-conditional contextual appearance*, i.e., the way it appears to the agent in the view of this extra-information Γ. Indeed, a given action might "appear" differently in a given context (i.e., at a state s) than it does in general: the information possessed by the agent at the state s might imply the negation of certain actions, hence their impossibility; this information will then be used to revise the agent's beliefs about the actions, obtaining her contextual beliefs. Moreover, in the presence of further information (a "post-condition" **P**), this appearance might again be revised. The "post-conditional contextual appearance" is the result of this double revision: the agent's belief about action σ is revised with the information given to her by the context s and the post-condition **P**. This information

is encoded in a set $\Gamma = \{\gamma \in \Sigma : s \models_{\mathbf{S}} \tilde{K}_a\langle\gamma\rangle\mathbf{P}\}$ of "admissible" actions: the actions for which the agent considers epistemically possible (at s) that they can be performed and they can achieve the post-condition $\mathbf{P}$. The "post-conditional contextual appearance" σ_a^Γ of action σ captures the agent's revised theory about σ after revision with the relevant information Γ.

So the above law says that: *the agent's future conditional beliefs* $[\sigma]B_a^{\mathbf{P}}$ *can be predicted, given that action* σ *happens, by her current conditional beliefs* $B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]$ *about what will be true after the apparent action* σ_a^Γ (as it appears in the given context and in the view of the given post-condition $\mathbf{P}$), *beliefs conditioned on the information* $(\langle\sigma_a^\Gamma\rangle\mathbf{P})$ *that the apparent action* σ_a^Γ *actually can lead to the fulfillment of the post-condition* $\mathbf{P}$.

Special cases. As special cases of the Action-Conditional-Belief Law, we can derive *all the reduction laws* in van Benthem (2007) for (conditional) belief after the events $!\mathbf{P}$, $\Uparrow\mathbf{P}$ and $\uparrow\mathbf{P}$:

$$[!\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \mathbf{P} \to B_a^{\mathbf{P}\wedge[!\mathbf{P}]\mathbf{Q}}[!\mathbf{P}]\mathbf{R}\,,$$

$$[\Uparrow\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \big(\tilde{K}_a^{\mathbf{P}}[\Uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{\mathbf{P}\wedge[\Uparrow\mathbf{P}]\mathbf{Q}}[\Uparrow\mathbf{P}]\mathbf{R}\big) \vee \big(\neg\tilde{K}_a^{\mathbf{P}}[\Uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{[\Uparrow\mathbf{P}]\mathbf{Q}}[\Uparrow\mathbf{P}]\mathbf{R}\big)\,,$$

$$[\uparrow\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \big(\tilde{B}_a^{\mathbf{P}}[\uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{\mathbf{P}\wedge[\uparrow\mathbf{P}]\mathbf{Q}}[\uparrow\mathbf{P}]\mathbf{R}\big) \vee \big(\neg\tilde{B}_a^{\mathbf{P}}[\uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{[\uparrow\mathbf{P}]\mathbf{Q}}[\uparrow\mathbf{P}]\mathbf{R}\big)\,,$$

where

$$K_a^{\mathbf{P}}\mathbf{Q} := K_a(\mathbf{P} \to \mathbf{Q})\,, \qquad \tilde{K}_a^{\mathbf{P}}\mathbf{Q} := \neg K_a^{\mathbf{P}}\neg\mathbf{Q}\,, \qquad \tilde{B}_a^{\mathbf{P}}\mathbf{Q} := \neg B_a^{\mathbf{P}}\neg\mathbf{Q}\,.$$

Laws for other doxastic attitudes. The *equi-plausibility modality behaves dynamically "almost" like knowledge*, while *the strict plausibility modality behaves like safe belief*, as witnessed by the following laws:

$$[\sigma][\cong_a]\mathbf{P} = \mathrm{pre}_\sigma \to \bigwedge_{\sigma'\cong_a\sigma} [\cong_a][\sigma']\mathbf{P}\,,$$

$$[\sigma][>_a]\mathbf{P} = \mathrm{pre}_\sigma \to \bigwedge_{\sigma'<_a\alpha} K_a[\sigma']\mathbf{P} \wedge \bigwedge_{\sigma''\cong_a\sigma} [>_a][\sigma'']\mathbf{P}\,.$$

From these, we can *derive laws for all the other doxastic attitudes* above.

The Logic of Doxastic Actions

The problem of finding a general syntax for action models has been tackled in various ways by different authors. Here we use the *action-signature approach* from Baltag and Moss (2004).

Signature. A doxastic *action signature* is a *finite plausibility frame* $\mathbf{\Sigma}$, together with an *ordered list without repetitions* $(\sigma_1, \ldots, \sigma_n)$ of some of the elements of $\mathbf{\Sigma}$. The elements of $\mathbf{\Sigma}$ are called *action types*. A type σ is called *trivial* if it is *not* in the above list.

Example 10. The *"hard" public announcement signature* HardPub is a singleton frame, consisting of one action type !, identity as the order relation, and the list (!).

The *"soft" public announcement signature* SoftPub is a two-point frame, consisting of types $\Uparrow$ and $\Downarrow$, with $\Downarrow <_a \Uparrow$ for all agents a, and the list $(\Uparrow, \Downarrow)$.

Similarly, one can define the signatures of *fully private announcements with n alternatives*, *private "fair-game" announcements, conservative upgrades* etc. As we will see below, *there is no signature of "successful (public) lying": public lying actions fall under the type of "soft" public announcements*, so they are generated by that signature.

Languages. For each action signature $(\mathbf{\Sigma}, (\sigma_1, \ldots, \sigma_n))$, the language $L(\mathbf{\Sigma})$ consists of a set of *sentences* φ and a set of *program terms* π, defined by simultaneous recursion:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid \Box_a\varphi \mid [\pi]\varphi$$
$$\pi ::= \sigma\varphi_1 \ldots \varphi_n \mid \pi \sqcup \pi \mid \pi;\pi$$

where $p \in \Phi$, $a \in \mathcal{A}$, $\sigma \in \Sigma$, and $\sigma\varphi_1 \ldots \varphi_n$ is an expression consisting of σ and a string of n sentences, where n is the length of the list $(\sigma_1, \ldots, \sigma_n)$.

Syntactic action model. The expressions of the form $\sigma\vec{\varphi}$ are called *basic programs*. The preorders on $\mathbf{\Sigma}$ induce in a natural way preorders on the basic programs in $L(\mathbf{\Sigma})$:

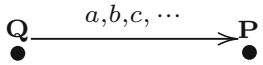
$$\sigma\vec{\varphi} \leq_a \sigma\vec{\psi} \text{ iff } \sigma \leq_a \sigma' \text{ and } \vec{\varphi} = \vec{\psi}.$$

The given listing can be used to assign syntactic preconditions for basic programs, by putting: $\mathrm{pre}_{\sigma_i\vec{\varphi}} := \varphi_i$, and $\mathrm{pre}_{\sigma\vec{\varphi}} := \top$ (the trivially true sentence) if σ is not in the listing. Thus, the basic programs of the form $\sigma\vec{\varphi}$ form a *"syntactic plausibility model"* $\mathbf{\Sigma}\vec{\varphi}$; i.e., every given interpretation $\|\cdot\| : L(\mathbf{\Sigma}) \to$ Prop of sentences as doxastic propositions will convert this syntactic model into a "real" (semantic) plausibility model, called $\mathbf{\Sigma}\|\vec{\varphi}\|$.

Action models induced by a signature. For a given signature $\mathbf{\Sigma}$, let $(\sigma_1, \ldots, \sigma_n)$ be its list of non-trivial types, and let $\vec{\mathbf{P}} = (\mathbf{P}_1, \ldots, \mathbf{P}_n)$ be a matching list of doxastic propositions. The *action model generated by the signature* $\mathbf{\Sigma}$ *and the list of propositions* $\vec{\mathbf{P}}$ is the model $\mathbf{\Sigma}\vec{\mathbf{P}}$, having $\mathbf{\Sigma}$ as its underlying action frame and having a precondition map given by: $\mathrm{pre}_{\sigma_i} = \mathbf{P}_i$, for non-trivial types σ_i; and $\mathrm{pre}_{\sigma} = \top$ (the trivially true proposition), for trivial types σ. When referring to σ as an *action* in $\mathbf{\Sigma}\vec{\mathbf{P}}$, we will denote it by $\sigma\vec{\mathbf{P}}$, to distinguish it from the action *type* $\sigma \in \Sigma$.

We can obviously extend this construction to *sets of action types*: given a signature $\mathbf{\Sigma}$ and a list $\vec{\mathbf{P}} = (\mathbf{P}_1, \ldots, \mathbf{P}_n)$, every set $\Gamma \subseteq \Sigma$ gives rise to a doxastic program $\Gamma\vec{\mathbf{P}} := \{\sigma\vec{\mathbf{P}} : \sigma \in \Sigma\} \subseteq \Sigma\vec{\mathbf{P}}$.

Example 11. The action model of a hard public announcement $!\,\mathbf{P}$ is generated as $!(\mathbf{P})$ by the hard public announcement signature $\mathsf{HardPub} = \{!\}$ and the list $(\mathbf{P})$. Similarly, the action model $\mathsf{SoftPub}(\mathbf{P})$ generated by the *soft* public announcement signature $\mathsf{SoftPub}$ and a list $(\mathbf{P}, \mathbf{Q})$ of two propositions consists of two actions $\Uparrow(\mathbf{P}, \mathbf{Q})$ and $\Downarrow(\mathbf{P}, \mathbf{Q})$, with $\Uparrow(\mathbf{P}, \mathbf{Q}) <_a \Downarrow(\mathbf{P}, \mathbf{Q})$, $\mathrm{pre}_{\Uparrow(\mathbf{P},\mathbf{Q})} = \mathbf{P}$ and $\mathrm{pre}_{\Downarrow(\mathbf{P},\mathbf{Q})} = \mathbf{Q}$:



This represents an event during which all agents share a common belief that $\mathbf{P}$ is announced; but they might be wrong and maybe $\mathbf{Q}$ was announced instead. However, it is common knowledge that either $\mathbf{P}$ or $\mathbf{Q}$ was announced.

Successful (public) lying $\mathrm{Lie}\,\mathbf{P}$ (*by an anonymous agent, falsely announcing* $\mathbf{P}$) can now be expressed as $\mathrm{Lie}\,\mathbf{P} := \Downarrow(\mathbf{P}, \neg\mathbf{P})$. The *truthful* soft announcement is $\mathrm{True}\,\mathbf{P} := \Uparrow(\mathbf{P}, \neg\mathbf{P})$. Finally, the soft public announcement (lexicographic update) $\Uparrow\mathbf{P}$, as previously defined, is given by the non-deterministic union $\Uparrow\mathbf{P} := \mathrm{True}\,\mathbf{P} \sqcup \mathrm{Lie}\,\mathbf{P}$.

Semantics. We define by simultaneous induction two *interpretation maps*, one taking sentences φ into doxastic propositions $\|\varphi\| \in \mathsf{Prop}$, the second taking program terms π into doxastic programs $\|\pi\|$ over some plausibility frames. The inductive definition uses the obvious semantic clauses. For programs: $\|\sigma\vec{\varphi}\|$ is the action $\sigma\|\vec{\varphi}\|$ (or, more exactly, the singleton program $\{\sigma\|\vec{\varphi}\|\}$ over the frame $\mathbf{\Sigma}\|\vec{\varphi}\|$), $\|\pi \sqcup \pi'\| := \|\pi\| \sqcup \|\pi'\|$, $\|\pi; \pi'\| := \|\pi\|; \|\pi'\|$. For sentences: $\|p\|$ is as given by the valuation, $\|\neg\varphi\| := \neg\|\varphi\|$, $\|\varphi \wedge \psi\| := \|\varphi\| \wedge \|\psi\|$, $\|K_a\varphi\| := K_a\|\varphi\|$, $\|\Box_a\varphi\| := \Box_a\|\varphi\|$, $\|[\pi]\varphi\| := [\,\|\pi\|\,]\,\|\varphi\|$.

Proof system. In addition to the axioms and rules of the logic $K\Box$, the logic $L(\mathbf{\Sigma})$ includes the following Reduction Axioms:

$$
\begin{aligned}
[\alpha]p &\leftrightarrow \mathrm{pre}_\alpha \to p\\
[\alpha]\neg\varphi &\leftrightarrow \mathrm{pre}_\alpha \to \neg[\alpha]\varphi\\
[\alpha](\varphi \wedge \psi) &\leftrightarrow \mathrm{pre}_\alpha \to [\alpha]\varphi \wedge [\alpha]\psi\\
[\alpha]K_a\varphi &\leftrightarrow \mathrm{pre}_\alpha \to \bigwedge_{\alpha' \sim_a \alpha} K_a[\alpha']\varphi\\
[\alpha]\Box_a\varphi &\leftrightarrow \mathrm{pre}_\alpha \to \bigwedge_{\alpha' <_a \alpha} K_a[\alpha']\varphi \wedge \bigwedge_{\alpha'' \cong_a \alpha} \Box_a[\alpha'']\varphi\\
[\pi \sqcup \pi']\varphi &\leftrightarrow [\pi]\varphi \wedge [\pi']\varphi\\
[\pi; \pi']\varphi &\leftrightarrow [\pi]\,[\pi']\varphi
\end{aligned}
$$

where p is any atomic sentence, π, π' are program terms, α is a *basic* program term in $L(\mathbf{\Sigma})$, pre is the syntactic precondition map defined above, and $\sim_a, <_a, \cong_a$ are respectively the (syntactic) epistemic indistinguishability, the strict plausibility order and the equi-plausibility relation on basic programs.

Theorem 16. *For every signature* $\mathbf{\Sigma}$*, the above proof system for the dynamic logic* $L(\mathbf{\Sigma})$ *is* complete, decidable *and* has the finite model property. *In fact, this dynamic logic has the same expressive power as the "static" logic* $K\Box$ *of knowledge and safe belief.*

Proof (Sketch). The proof is similar to the ones in Baltag and Moss (2004), Baltag et al. (1998), and van Ditmarsch et al. (2007). We use the reduction laws to inductively simplify any formula until it is reduced to a formula of the $K\Box$-logic, then use the completeness of the $K\Box$ logic. Note that this is *not* an induction on subformulas, but (as in Baltag et al. 1998) on an appropriate notion of "complexity" ordering of formulas. Q.E.D.

Current and Future Work, Some Open Questions

In our papers Baltag and Smets (2007a,b), we present a *probabilistic version* of the theory developed here, based on *discrete (finite) Popper-Renyi conditional probability spaces* (allowing for conditionalization on events of non-zero probability, in order to cope with non-trivial belief revisions). We consider subjective probability to be the proper notion of "degree of belief", and we investigate its relationship with the qualitative concepts developed here. We develop a probabilistic generalization of the Action Priority Rule, and show that the logics presented above are *complete for the (discrete) conditional probabilistic semantics*.

We mention here a number of open questions: (1) Axiomatize the full (static) logic of doxastic attitudes introduced in this paper. It can be easily shown that they can all be reduced to the modalities K_a, $[>_a]$ and $[\cong_a]$. There are a number of obvious axioms for the resulting logic $K[>][\cong]$ (note in particular that $[>]$ satisfies the Gödel-Löb formula!), but the completeness problem is still open. (2) Axiomatize the logic of *common safe belief and common knowledge*, and their *dynamic versions*. More generally, explore the logics obtained by adding *fixed points*, or at least "epistemic regular (PDL-like) operations" as in van Benthem et al. (2006b), on top of our doxastic modalities. (3) Investigate the *expressive limits* of this approach *with respect to belief-revision policies*: what policies can be simulated by our update? (4) Extend the work in Baltag and Smets (2007a,b), by investigating and axiomatizing doxastic logics on *infinite* conditional probability models. (5) Extend the logics with *quantitative (probabilistic)* modal operators $B^P_{a,x}Q$ (or $\Box_{a,x}Q$) expressing that *the degree of conditional belief in* Q *given* P (or the *degree of safety* of the belief in Q) is at least x.

Acknowledgements Sonja Smets' contribution to this research was made possible by the post-doctoral fellowship awarded to her by the Flemish Fund for Scientific Research. We thank Johan van Benthem for his insights and help, and for the illuminating discussions we had with him on the topic of this paper. His pioneering work on dynamic belief revision acted as the "trigger" for our own. We also thank Larry Moss, Hans van Ditmarsch, Jan van Eijck and Hans Rott for their most valuable feedback. Finally, we thank the editors and the anonymous referees of the LOFT7-proceedings for their useful suggestions and comments.

During the republication of this paper in 2015, the research of Sonja Smets was funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no.283963.

References

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*(2), 510–530.

Aucher, G. (2003). *A combined system for update logic and belief revision*. Master's thesis, University of Amsterdam. ILLC Publications MoL-2003-03.

Aumann, R. J. (1999). Interactive epistemology I: Knowledge. *International Journal of Game Theory, 28*(3), 263–300.

Baltag, A. (2002). A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research, 54*(1), 1–46.

Baltag, A., & Moss, L. S. (2004). Logics for epistemic programs. *Synthese, 139*(2), 165–224.

Baltag, A., Moss, L. S., & Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In I. Gilboa (Ed.), *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, (pp. 43–56).

Baltag, A., & Sadrzadeh, M. (2006). The algebra of multi-agent dynamic belief revision. *Electronic Notes in Theoretical Computer Science, 157*(4), 37–56.

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

Baltag, A., & Smets, S. (2006b) Dynamic belief revision over multi-agent plausibility models. In Bonanno et al. (2006) (pp. 11–24).

Baltag, A., & Smets, S. (2006c). The logic of conditional doxastic actions: A theory of dynamic multi-agent belief revision. In S. Artemov, & Parikh, R. (Eds.), *Proceedings of ESSLLI Workshop on Rationality and Knowledge*, (pp. 13–30). ESSLLI.

Baltag, A., & Smets, S. (2007a). From conditional probability to the logic of doxastic actions. In D. Samet (Ed.), *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, Brussels (pp. 52–61). UCL Presses Universitaires de Louvain.

Baltag, A., & Smets, S. (2007b). Probabilistic dynamic belief revision. In J. F. A. K. van Benthem, S. Ju, & F. Veltman (Eds.), *A Meeting of the Minds: Proceedings of the Workshop on Logic, Rationality and Interaction*, Beijing, 2007 (Texts in computer science, Vol. 8). London: College Publications.

Battigalli, P., & Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory, 105*(2), 356–391.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal logic* (Cambridge tracts in theoretical computer science, Vol. 53). Cambridge: Cambridge University Press.

Board, O. (2002). Dynamic interactive epistemology. *Games and Economic Behaviour, 49*(1), 49–80.

Bonanno, G. (2005). A simple modal logic for belief revision. *Synthese, 147*(2), 193–228.

Bonanno, G., van der Hoek, W., & Wooldridge, M. (Eds.). (2006). *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT7)*, University of Liverpool UK.

Friedmann, N., & Halpern, J. Y. (1994). Conditional logics of belief revision. In *Proceedings of the of 12th National Conference on Artificial Intelligence (AAAI-94)*, Seattle, 31 July–4 Aug 1994 (pp. 915–921). Menlo Park: AAAI.

Gärdenfors, P. *Knowledge in flux: Modelling the dynamics of epistemic states*. Gardenfors. 1988, MIT Press, Cambridge/London.

Gerbrandy, J. (1999). Dynamic epistemic logic. In L. S. Moss, J. Ginzburg, & M. de Rijke (Eds.), *Logic, language and information* (Vol. 2, p. 67–84). Stanford: CSLI Publications/Stanford University.

Gerbrandy, J., & Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language and Information, 6*(2), 147–169.

Gerbrandy, J. D. (1999). *Bisimulations on planet Kripke*. PhD thesis, University of Amsterdam. ILLC Publications, DS-1999-01.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121–123.

Gochet, P., & Gribomont, P. (2006). Epistemic logic. In D. M. Gabbay & J. Woods (Eds.), *Handbook of the history of logic* (Vol. 7, p. 99–195). Oxford: Elsevier.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic, 17*(2), 157–170.

Hintikka, J. (1962). *Knowledge and belief*. Ithaca: Cornell University Press.

Katsuno, H., & Mendelzon, A. O. (1992). On the difference between updating a knowledge base and revising it. In P. Gärdenfors (Ed.), *Belief revision* (Cambridge tracts in theoretical computer science, pp. 183–203). Cambridge/New York: Cambridge University Press.

Klein, P. (1971). A proposed definition of propositional knowledge. *Journal of Philosophy, 68*(16), 471–482.

Kooi, B. P. (2003). Probabilistic dynamic epistemic logic. *Journal of Logic, Language and Information, 12*(4), 381–408.

Lehrer, K. (1990). *Theory of knowledge*. London: Routledge.

Lehrer, K., & Paxson, T. Jr. (1969). Knowledge: Undefeated justified true belief. *Journal of Philosophy, 66*(8), 225–237.

Meyer, J.-J. Ch. & van der Hoek, W. (1995). *Epistemic logic for AI and computer science* (Cambridge tracts in theoretical computer science, Vol. 41). Cambridge: Cambridge University Press.

Pappas, G., & Swain, M. (Eds.). (1978). *Essays on knowledge and justification*. Ithaca: Cornell University Press.

Plaza, J. A. (1989). Logics of public communications. In M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, & Z. W. Ras (Eds.), *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems Poster Session Program* (pp. 201–216). Oak Ridge National Laboratory, ORNL/DSRD-24.

Rott, H. (1989). Conditionals and theory change: Revisions, expansions, and additions. *Synthese, 81*(1), 91–113.

Rott, H. (2004). Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis, 61*(2–3), 469–493.

Ryan, M., & Schobbens, P.-Y. (1997). Counterfactuals and updates as inverse modalities. *Journal of Logic, Language and Information, 6*(2), 123–146.

Segerberg, K. (1998). Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic, 39*(3), 287–306.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change, and statistics* (Vol. II, pp. 105–134). Dordrecht/Boston: Kluwer Academic

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (APQ monograph series, Vol. 2). Oxford: Blackwell.

Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–163.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies, 128*(1), 169–199.

van Benthem, J. F. A. K. (2007). Dynamic logic for belief revision. *Journal of Applied Non-classical Logics, 17*(2), 129–155.

van Benthem, J. F. A. K., Gerbrandy, J., & Kooi, B. (2006a) Dynamic update with probabilities. In Bonanno et al. (2006) (pp. 237–246).

van Benthem, J. F. A. K., & Liu, F. (2004). Dynamic logic of preference upgrade. Technical report, University of Amsterdam. ILLC Publications, PP-2005-29.

van Benthem, J. F. A. K., van Eijck, J., & Kooi, B. P. (2006b). Logics of communication and change. *Information and Computation, 204*(11), 1620–1662.

van der Hoek, W. (1993). Systems for knowledge and beliefs. *Journal of Logic and Computation, 3*(2), 173–195.

van Ditmarsch, H. P. (2000). *Knowledge games*. PhD thesis, University of Groningen. ILLC Pubications, DS-2000-06.

van Ditmarsch, H. P. (2002). Descriptions of game actions. *Journal of Logic, Language and Information, 11*(3), 349–365.

van Ditmarsch, H. P. (2005) Prolegomena to dynamic logic for belief revision. *Synthese, 147*(2), 229–275.

van Ditmarsch, H. P., & Labuschagne, W. (2007). My beliefs about your beliefs: A case study in theory of mind and epistemic logic. *Synthese, 155*(2), 191–209.

van Ditmarsch, H. P., van der Hoek, W., & Kooi, B. P. (2007). *Dynamic epistemic logic* (Synthese library, Vol. 337). Dordrecht: Springer.

Voorbraak, F. P. J. M. (1993). *As far as I know*. PhD thesis, Utrecht University, Utrecht (Quaestiones infinitae, Vol. VII).

Williamson, T. (2001). Some philosophical aspects of reasoning about knowledge. In J. van Benthem (Ed.), *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge (TARK'01)* (p. 97). San Francisco: Morgan Kaufmann.