Synthese (2008) 165:179–202
Knowledge, Rationality & Action 143–166
DOI 10.1007/s11229-008-9369-8

Probabilistic dynamic belief revision

Alexandru Baltag · Sonja Smets

Received: 14 January 2008 / Accepted: 8 July 2008 / Published online: 8 October 2008

Abstract We investigate the discrete (finite) case of the Popper–Renyi theory of conditional probability, introducing *discrete conditional probabilistic models for knowledge and conditional belief*, and comparing them with the more standard plausibility models. We also consider a related notion, that of *safe belief*, which is a weak (non-negatively introspective) type of "knowledge". We develop a probabilistic version of this concept ("degree of safety") and we analyze its role in games. We completely axiomatize the logic of conditional belief, knowledge and safe belief over conditional probabilistic models. We develop a theory of probabilistic dynamic belief revision, introducing probabilistic "action models" and proposing a notion of probabilistic update product, that comes together with appropriate reduction laws.

Keywords Belief revision · Conditional belief · Dynamic-epistemic logic · Popper functions

A. Baltag
Oxford University Computing Laboratory, University of Oxford,
Parks Road, Oxford OX1 3QD, UK
e-mail: alexandru.baltag@comlab.ox.ac.uk

A. Baltag
GPI, University of Hertfordshire, Hatfield, UK

S. Smets (✉)
Center for Logic and Philosophy of Science, Vrije Universiteit Brussel,
Brussels, Belgium
e-mail: sonsmets@vub.ac.be

S. Smets
IEG, University of Oxford, Oxford, UK

1 Introduction

In this paper, we present an original semantical setting for belief dynamics, by combining three commonly-used approaches to belief change: (1) the *Bayesian approach in its extended Popper–Renyi* (*-de Finetti*) *version* (based on Popper functions, allowing conditionalization on events of probability zero), (2) the *classical AGM-style Belief Revision approach* in its *semantic* presentation (based on plausibility models and plausibility ranking) and (3) the "*Dynamic Epistemic Logic*" (*DEL*) *approach*[1] (based on the distinction between "static" belief revision, represented using conditional belief operators, and "dynamic" revision, represented using epistemic/doxastic action models, dynamic modalities and some notion of product update).

Some connections between the first two approaches were already studied, most thoroughly in VanFraassen (1995), Boutilier (1995), Halpern (2003), and Arlo-costa and Parikh (2005), where it was shown that a correct probabilistic understanding of belief revision and conditional beliefs requires an extension of classical probability theory, along the lines of the Popper–Renyi axioms. The connections between the second and the third approach are investigated in a number of recent papers (Aucher 2003; van Ditmarsch 2005; van Benthem 2006; Baltag and Smets 2006a,b,c, 2008), while van Benthem (2003), van Benthem et al. (2006a), and Kooi (2003) relate the first and the third approach only briefly (using only classical probabilistic models).

Combining these approaches into one, we introduce *a* "*qualitative*" *dynamic logic of conditional beliefs, knowledge, safe belief and belief-updating actions*, which is *decidable and complete with respect to* (*Popper-style*) *conditional-probabilistic models*. The syntax and the proof system for this logic are the same we introduced in Baltag and Smets (2006b,c, 2008), but *the semantics is probabilistic* (instead of using plausibility models). We develop a theory of dynamic belief revision over probabilistic models, by introducing "action models" and a notion of *conditional-probabilistic product update*, which generalizes to a belief-revision context the corresponding notion introduced in Kooi (2003), and van Benthem et al. (2006a) for probabilistic epistemic actions. One can also extract from the probabilistic update product a corresponding qualitative update notion for plausibility models; this last notion coincides with the "*Action-Priority Update*",[2] introduced in Baltag and Smets (2006b,c, 2008), as a way of combining the "update product" from Baltag et al. (1998), and Baltag and Smets (2006a) with ideas from Belief Revision theory.

This paper assumes the general distinction, made in van Ditmarsch (2005), Baltag and Smets (2006a), and van Benthem (2006), between "*dynamic*" *and* "*static*" *belief revision*. To summarize it: "static" belief revision, corresponding closely to the classical AGM theory (Alchourrón et al. 1985; Gärdenfors 1988) and embodied in our setting by the conditional belief operators $B_a^P Q$, captures *the agent's changing beliefs about an unchanging world*. But since in fact, in a modal logic setting, the world is *always changed by our changes of beliefs*, the best way to understand a doxastic

[1] I.e. in the tradition of Gerbrandy (1999), Baltag et al. (1998), Baltag and Moss (2004), Baltag (2002) and of the work of J. van Benthem and the "Amsterdam school".

[2] This name was proposed by J. van Benthem. In Baltag and Smets (2006b,c), this was called "anti-lexicographic product update".

conditional $B_a^P Q$ is as saying that *after learning P, agent a believes 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 believes 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" beliefs (about ontic facts) from a non-modal perspective, neglecting any higher-order "beliefs about beliefs". As a result, it completely misses the changes induced in the world (including in the other agents' epistemic states) by our belief-updating actions (e.g. the learning of a Moore sentence). This is shown by the acceptance in Katsuno and Mendelzon (1992) of the AGM "Success Axiom": in dynamic notation, the setting in Katsuno and Mendelzon (1992) validates the axiom $[P!]B_a P$ (which cannot accommodate Moore sentences). Instead, the authors of Katsuno and Mendelzon (1992) exclusively concentrate on the possible changes of (ontic) facts that may have occurred during 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 (1992): we completely neglect here the ontic changes,[3] considering a world in which the only changes are induced by *"purely doxastic" actions* (such as learning, discovery, communication etc.).

As in Baltag and Smets (2006b) (but now with a probabilistic semantics), we introduce a "weak" (non-negatively introspective) notion of "knowledge", notion we call *safe belief*, to distinguish it from the standard S5-type "knowledge" (Aumann's partition-based knowledge). Safe belief corresponds to Stalnaker's "knowledge" (Stalnaker 1996), itself a modal formalization of Lehrer's conception of "knowledge as in-defeasible belief" (Lehrer 1990). In the context of probabilistic models, we refine this notion by introducing a quantitative scale of "*degrees of safety*" *of a given belief*, giving examples from Game Theory to illustrate the usefulness of this notion.

2 Conditional doxastic logic over probabilistic models

It is well known that simple probability measures yield problems in the context of describing an agent's beliefs and how they can be revised.

First, it seems natural to assume (and it is usually assumed) that *beliefs are closed under finitary conjunctions* (and that moreover the belief operator satisfies Kripke's axiom K). But then the so-called Lottery Paradox (van Fraassen 1995) shows that *no probability other than* 1 *can capture this notion of belief*.

The paradox goes as follows. In a fair lottery composed of 1,000 tickets, an agent assigns probability 0.999 to the event that any particular ticket is *not* the winning one. *If* we identify "belief" with having subjective probability ≥ 0.999, then it follows that, for any given ticket, the agent "*believes*" *that this ticket is not the winning one*; *if* moreover we accept that beliefs are closed under finite conjunctions, it follows that the agent "*believes*" *that no ticket is the winning one*! But this is obviously absurd (and

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

it contradicts the above probabilistic reading of belief): obviously, the agent will *not* assign probability 0.999 to the belief that *no* ticket is the winning one! On the contrary, he should assign probability 0 to this belief: some ticket will definitely be winning.

The same argument applies to any other probability $p < 1$. What this means is that (if we accept the closure of beliefs under finitary conjunctions, then) *we cannot identify* "*belief*" *with* "*subjective probability* $\geq p$" for *any* $p < 1$: none of these is closed under finitary conjunctions.

The *only* probability left to be assigned to events that are "believed" is *probability 1*. So it seems that "*belief*" *must imply* "(subjective) *probability* 1". But then, assuming that *beliefs can sometimes be false* (which seems natural and unavoidable if we want to distinguish "belief" from "knowledge"), it follows that Bayesian updating is *not* an appropriate model for "learning", at least not in the case of agents having any such prior false "beliefs" (in the sense of subjective probability 1).

Indeed, the received wisdom in Bayesianism is that *learning new information corresponds to probabilistic conditionalization*. But once "belief" is accepted to imply "probability 1 ", any non-trivial belief revision (triggered by learning that one of the agent's beliefs was false) will correspond to *conditioning on events of measure* 0: an impossible task in the classical (Kolmogorovian) theory, in which conditional probability is defined as a ratio (whose denominator will be 0 in this case). In probabilistic applications, e.g. in Game Theory, this problem is sometimes preempted by requiring that only impossible events are assigned probability 0. But this, in effect, is a way of eluding the problem by simply stipulating that *agents never have any wrong beliefs*. In fact, this *collapses belief into knowledge*: in a finite discrete probabilistic space[4] satisfying this stipulation,[5] the "belief" operator becomes equivalent to a quantifier over all the states of the space; but this is the standard definition of *knowledge* in Game Theory! The unavoidable conclusion is that Bayesian belief update, based on standard Probability Theory, simply cannot deal with any non-trivial belief revision.

There are several possible solutions to this problem. In this paper we adopt the Popper–Renyi theory of conditional probabilities (Popper 1968; van Fraassen 1976, 1995; Renyi 1964, 1955; Halpern 2001) which takes conditional probability as basic instead of simple probability, and which was already applied to belief revision in van Fraassen (1995), Arlo-Costa and Parikh (2005), and Halpern (2003). We focus here on the *discrete finite* case, which gives us a simplified, unique, "canonical" setting,[6] that can be easily compared with the qualitative (plausibility-based) settings for belief revision.

A *discrete conditional probability space* (*dcps*, for short) is a pair (S, μ), where S is a *finite* set of states and $\mu: \mathcal{P}(S) \times \mathcal{P}(S) \rightarrow [0, 1]$ is a so-called "Popper function", i.e. it satisfies the following axioms:

[4] I.e. a finite state space such that all its subsets are measurable.

[5] In the finite discrete case, this says that every state has a non-zero probability.

[6] The various axiomatic settings proposed in Popper (1968), Renyi (1964), van Fraassen (1976) become equivalent in the discrete case, and moreover they are equivalent in this case with the *conditional lexicographic probability spaces* proposed by researches in Game Theory.

1. $\mu(A \mid A) = 1$,
2. $\mu(A \cup B \mid C) = \mu(A \mid C) + \mu(B \mid C)$, if $A \cap B = \emptyset$, $C \neq \emptyset$,
3. $\mu(A \cap B \mid C) = \mu(A \mid B \cap C) \cdot \mu(B \mid C)$.

In fact, a discrete Popper function μ on a finite space is completely characterized by its behavior on *pairs of states*, i.e. by all the quantities

$$(s, t)_\mu := \mu(\{s\} \mid \{s, t\}),$$

with $s, t \in S$. We *skip the subscript* when the measure is *understood*, and call (s, t) the *priority degree* of s with respect to t.

Observation 1 To see that *priority degrees do capture indeed all the information about the original Popper function* μ, it is enough to observe that, for every A and every $B \neq \emptyset$, we have

$$\mu(A|B) = \sum_{s \in A \cap B} \frac{1}{\sum_{t \in B} \frac{(t,s)}{(s,t)}}$$

(where we use the usual conventions: $\frac{1}{0} = \infty$, $\frac{1}{\infty} = 0$, $\infty + \infty = \infty$ and $\infty + x = \infty$ for all real numbers x).

This gives us an *alternative description* of dcps's as *priority spaces*:

A *priority space* is a pair $(S, (\bullet, \bullet))$, where S is a finite set of states and $(\bullet, \bullet) : S \times S \to [0, 1]$ is a probabilistic assignment on S, satisfying the axioms:

$$\begin{aligned} (s, s) &= 1, \\ (t, s) &= 1 - (s, t) \quad \text{for } s \neq t, \\ (s, w) &= \frac{(s, t) \cdot (t, w)}{(s, t) \cdot (t, w) + (w, t) \cdot (t, s)} \quad \text{for } s \neq w \text{ and denominator } \neq 0. \end{aligned}$$

Proposition 2.1 *Every dcps* (S, μ) *gives rise to a* (*unique*) *priority space* $(S, (\bullet, \bullet))$, *satisfying* $(s, t) = (s, t)_\mu$ *for all* $s, t \in S$. *Conversely, every priority space* $(S, (\bullet, \bullet))$ *uniquely determines a dcps* (S, μ) *satisfying* $(s, t) = (s, t)_\mu$, *for all* $s, t \in S$.

Proof For the first direction, we can easily check that, if we are given a dcps, then the operation defined by $(s, t) := (s, t)_\mu = \mu(\{s\}|\{s, t\})$ satisfies the axioms of a priority space. (Indeed, each of the axioms follows from the corresponding dcps axiom.)

For the converse, given a priority space $(S, (\bullet, \bullet))$, we put $\mu(A|B) := \sum_{s \in A \cap B} \frac{1}{\sum_{t \in B} \frac{(t,s)}{(s,t)}}$, and we verify that this satisfies the axioms of a dcps and that in addition we have $(s, t)_\mu = (s, t)$. □

The (*non-strict*) *priority relation* $\leq \subseteq S \times S$, defined on states by putting $s \leq t$ iff $(s, t) \neq 0$, is a special case of (the converse of) the "superiority" relation introduced in van Fraassen (1995) (following De Finetti). In fact, the *strict* priority relation, which can be easily seen to be given by

$$s < t \quad \text{iff } (t, s) = 0 \quad \text{iff both } (s, t) = 1 \quad \text{and } s \neq t,$$

was introduced in Arlo-Costa and Parikh (2005) under the name of "ranking ordering". It is easy to see that (in the finite discrete case to which we are confined here) *the priority relation is a total preorder*.[7] By arranging the states of a finite space of size n in a list of non-decreasing priority $s_1 \leq s_2 \leq \cdots \leq s_n$, we see that *in order to specify a discrete space of size n it is enough to give* $n-1$ *independent* (*non-zero*) *conditional probabilities, namely the priority degrees* (s_i, s_{i+1}) *for* $1 \leq i \leq n-1$.

In a dcps, there is a straightforward way to define (conditional) belief and knowledge, by simply identifying "belief" with "probability 1"[8] and "knowledge" with "true in all states". In other words, for a "proposition" $P \subseteq S$, we put:

$$B^P Q \text{ iff } \mu(Q|P) = 1,$$
$$K P \text{ iff } P = S.$$

"Belief" is defined as "belief conditional on a tautology", i.e. we put

$$B P := B^S P.$$

But these definitions assume that the state space S is already *restricted* to all the states that the (implicit) agent considers as *epistemically possible*. In a more general context (as the multi-agent case considered below), *knowledge has to be defined by quantifying only over epistemically possible states*, while in the definition of belief we have to conditionalize μ on the set of epistemically possible states.

Conditional probabilistic frames. Given a (finite) set $\mathcal{A}$ of "agents", a *discrete conditional probability frame* (or *dcpf*, for short) is a structure $(S, \mu_a, \Pi_a)_{a\in\mathcal{A}}$, such that, for each $a \in \mathcal{A}$, (S, μ_a) is a discrete conditional probability space and Π_a is a partition (the "information partition") of S. Equivalently, we can of course use equivalence relations $\sim_a$ instead of partitions Π_a. For a state s, denote by $s(a)$ *the information cell* of s in the partition Π_a (or the $\sim_a$*−equivalence class* of s). *Knowledge* and (*conditional*) *belief* become now dependent on the (information the agent possesses about the) state:

$$B_a^P Q := \{s \in S : \mu(Q|P \cap s(a)) = 1\},$$
$$K_a P := \{s \in S : s(a) \subseteq P\}.$$

We interpret the conditional belief statement $s \in B_a^P Q$ in the following way: if the actual state is s, then after "learning" that P *is* the case (in the state s), agent a will believe that Q *was* the case (at the same state s, i.e. before the learning). We abbreviate $(s,t)_{\mu_a}$ as $(s,t)_a$. We denote by $\leq_a$ the induced priority relation.

Conditional probabilistic models. For a given set Φ of *atomic sentences* (intuitively denoting "ontic facts" about the world), a *discrete conditional probability*

[7] Arlo-Costa and Parikh (2005) extends this to countably additive probability measures.

[8] Van Fraassen (1995) and Arlo-Costa and Parikh (2005) consider more subtle distinctions (such as "full belief" and "plain belief"), but it is easy to see that in the case of discrete spaces all these notions become equivalent with "belief" as defined here.

model (*dcpm*, for short) is, as usually in modal logic, a structure $\mathbf{S} = (S, \mu_a, \Pi_a, \|\bullet\|)$ consisting of a dcpf (S, μ_a, Π_a) together with a valuation $\|\bullet\| : \Phi \to \mathcal{P}(S)$.

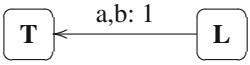
In fact, for the doxastic-epistemic logic (i.e. for computing K_a, B_a and B_a^P), *only the priority degrees between distinct, but epistemically indistinguishable states are relevant*. So it is enough to know $\{(s,t)_a : s \sim_a t\}$. One can thus consider the *local priority* relation $\trianglelefteq_a$, defined as the intersection of the relations $\leq_a$ and $\sim_a$. We denote by $\triangleleft_a$ the corresponding *strict* relation. It is easy to see that, for $s \neq t$, we have

$$s \triangleleft_a t \quad \text{iff } t \in B_a^{\{s,t\}}\{s\}.$$

In other words, *the* (*local*) *priority relation actually captures a notion of priority of beliefs*: given the information that the actual state is one of two different states s or t, agent a will believe it is s (with probability 1) iff $s \triangleleft_a t$. So in representing a dcpm, we will only give the priority degrees between successive distinct states in the same partition cell (listed in non-decreasing local priority order). This provides a way to encode the information partition itself into the probabilistic information, so we *do not* have to represent the indistinguishability relations as well.

Example 1 Alice, Bob and Charles play the following simple game: (it is common knowledge that) one million dollars is put in a box and Bob is first invited in the room. He has *only two options*: he can either *take the money* (**T**) or *leave it* (**L**). If he takes the money, then he gets to keep it, otherwise he ends the game with no money. The box is then covered, and Alice and Bob are invited in the room. They can see each other and see the covered box, so they see that none of them can see inside the box. Separately and independently, each of them has guess if Charles took the money or not. If either one guesses correctly, then he or she is also awarded one million dollars. We assume Alice and Bob are a team, so they are interested to maximize their joint income (the sum of their profits). E.g. if they both make correct guesses, they end up with a joint income of two million dollars; in contrast, if none of them makes a correct guess, they end up with nothing.

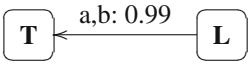
In such a game, it is clearly very important what each player knows or believes about the other players, including about the other players' beliefs etc. Let us assume Alice and Bob announced their beliefs before entering the room (since they are a team, so it is in their common interest to reach common knowledge!), and that it is now *common knowledge that they both believe* that Charles is a "rational" player (in the sense of Game Theory): *they believe he took the money*. In fact, they're right: Charles *did take* the money. We *model* this situation as a *dcpm* with two possible states, one in which Charles takes the money (so that the atomic sentence **T** is true) and one in which Charles leaves the money in the box (so that the sentence **L** is true):



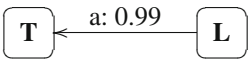
Let us denote by t the state on the left (in which Charles took the money) and by l the state on the right (in which Charles left the money in the box). So the valuation map has $\|\mathbf{T}\| = \{t\}$ and $\|\mathbf{L}\| = \{l\}$. The arrows represent the doxastic/epistemic

information. As announced, we only represent the agents' priority degrees between distinct indistinguishable states. So the fact that there are no c-arrows means that the *all states are distinguishable for Charles*: Charles' information partition Π_c has only single singletons $t(a) = \{t\}$ and $l(b) = \{l\}$. This captures the fact that Charles *knows* whether or not he took the money. In contrast, Alice and Bob *don't know* for sure what happened, but they *only have beliefs* about this: this is captured by the existence of a-arrows and b-arrows between the two states. This means they cannot distinguish epistemically between the two states (since they cannot see inside the box), so the information partitions of Alice and Bob consist of only one information cell: the whole state space $t(a) = t(a) = l(b) = l(b) = \{t, l\}$. Finally, the fact that both the a-arrows and b-arrows are labeled with the number 1 and they go from state l to state t means that the priority degree of state t with respect to l is 1 for both Alice and Bob: $(t, l)_a = (t, l)_b = 1$, so they indeed *believe* (*with probability* 1) *that Charles took the money*. It is not necessary to represent the converse arrows (from t to l), since their numerical values can be deduced from the axioms of a priority space: $(l, t)_a = 1 - (t, l)_a = 0$, and similarly for $(l, t)_b$.

Example 2 Compare this with the case in which Alice and Bob are *not completely certain* that Charles is "rational". Instead, let us suppose that it is now *common knowledge that they both assign a probability of* 0.99 *to the event of Charles taking the money* (and a probability of 0.01 to the event of Charles leaving the money in the box). As before, such a situation could be realized by Alice and Bob first announcing what they believe about Charles, and with what probability. The resulting model is the following:



Example 3 The situation is as in Example 2, except that it is now *common knowledge that, after entering the room, Bob can privately see* inside the box (since he's allowed to briefly lift the cover and take a look inside). Not only his probabilities become trivial (0 or 1, depending on what he sees), but his information partition will now consist of singletons. So, as for Charles, we will not even have to represent Bob's priority degrees anymore: in each case, he *knows* the real state of the system.



3 Relating probabilistic and relational models of conditional belief

To compare probabilistic models with *qualitative* (*relational*) ones, we introduce Kripke models for knowledge and conditional belief based on plausibility relations. A *finite* (*epistemic-doxastic*) *plausibility frame*[9] is a structure $(S, \leq_a, \sim_a)_a$, where S

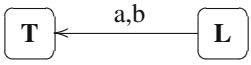
[9] The notion here is the one we introduced in Baltag and Smets (2006a), Baltag and Smets (2006b,c), but it is closely related to other notions in the literature: "Grove models", "Lewis spheres", "Spohn ordinal plausibility ranking".

is finite and, for each a, $\leq_a$ is a *total* (*i.e.* "*connected*", *or* "*complete*") *preorder* and $\sim_a$ is an equivalence relation on S. A plausibility model is a plausibility frame together with a valuation. In a plausibility model, knowledge K_a is defined in the standard way (as a Kripke modality) using $\sim_a$, while conditional belief is defined as:

$$B_a^P Q := \{s \in S : Min_{\leq_a} P \cap s(a) \subseteq Q\}$$

where we used the notations $Min_{\leq_a} T := \{s \in T : s \leq_a t \text{ for all } t \in T\}$ and $s(a) := \{t \in S : s \sim_a t\}$.

Note that (as in the case of probability models) only the plausibility relation between states in the same information cell are relevant; in other words, only the "local" plausibility relation $\trianglelefteq_a := \leq_a \cap \sim_a$ is needed. In fact, this relation encodes the epistemic relations $\sim_a$ as well. So, as before, we only represent $\trianglelefteq_a$, and for convenience we skip all the loops (since $\trianglelefteq$ is reflexive anyway). Now, Example 1 above becomes:



For both the ("global") plausibility relations $s \leq_a t$ and for their "local" correspondent $s \trianglelefteq_a t$, we can also consider their "strict" versions $s <_a t$ and $s \triangleleft_a t$. Finally, the relation of "equi-plausibility" is the equivalence relation $\cong_a$ induced by the preorder $\trianglelefteq_a$: $s \cong_a t$ iff $s \trianglelefteq_a t$ and $t \trianglelefteq_a s$.

One usually reads $s \leq_a t$ as saying that *state s is "at least as plausible" as state t*. But, as shown by the counterexample below, we *cannot* identify "more plausible" with "higher probability": if we accept the identification of (conditional) belief with "(conditional) probability 1", then we *cannot* have $s \leq_a t$ iff $\mu_a(s) \geq \mu_a(t)$. For the same reason, the plausibility preorder is *not* given by the order of *conditional degrees of belief*: we *do not* have $s \leq_a t$ iff $(s,t)_a \geq (t,s)_a$.

Counterexample Consider for instance the situation in Example 2 above, where we have $\mu_a(t) = (t,l)_a = 0.99$ and $\mu_a(l) = (l,t)_a = 0.01$. If higher probability (or higher conditional degree of belief) would imply "more plausible", then we would have $t <_a l$ (since $\mu_a(t) > \mu_a(l)$ and $(t,l)_a > (l,t)_a$), then by the definition of belief in a plausibility model we would have that $B_a\mathbf{T}$ is *true* in both states t and l (since $\mathbf{T}$ is true in all the "most plausible" states, i.e. in t). But this *contradicts* the probabilistic definition of belief (as "probability 1 "): indeed, $\mu_a(\mathbf{T}|t(a)) = \mu_a(\mathbf{T}|\{t,l\} = \mu_a(\mathbf{T}) = 0.99 \neq 1$, and hence (according to the probabilistic definition of belief) the sentence $B_a\mathbf{T}$ is *not* true at state t.

So, since the plausibility relation cannot be identified with the relation of "having higher probability than" (nor with "having higher conditional degree of belief than"), the question is: *how can we relate the two* (*probabilistic and plausibility-based*) *accounts for belief*?

The answer is given by the following result:

Proposition 3.1

1. *Every discrete conditional probability model* $\mathbf{S} = (S, \mu_a, \Pi_a, \| \bullet \|)$ *gives rise to a plausibility model* $\overline{\mathbf{S}} = (S, \leq_a, \sim_a, \| \bullet \|)$, having *the same state space, the*

same valuation and the same notions of knowledge and (*conditional*) *belief* as the original model **S**.

2. Conversely, *every finite plausibility model* $\overline{\mathbf{S}}$ *can be "probabilized"*: we can define conditional probability measures μ_a for each agent at, that will give rise to the same conditional beliefs as $\overline{\mathbf{S}}$.

For the *proof*, we need a useful preliminary result:

Lemma 3.2 If μ is a *discrete Popper function on a finite set* S and $\leq$ is the *corresponding priority relation* (defined as above by putting $s \leq t$ iff $(s,t)_\mu \neq 0$), then the following holds for all events $A, B \subseteq S$:

$$\mu(A|B) = 1 \Leftrightarrow Min_{\leq} B \subseteq A.$$

Proof of Lemma 3.2 For all events A, B, we have the following chain of equivalencies: $\mu(A \mid B) = 1$ iff $\mu(B \setminus A|A) = 0$ iff $\sum_{s \in B \setminus A} \frac{1}{\sum_{t \in B} \frac{(t,s)}{(s,t)}} = 0$ (by Observation 1 in the previous section) iff $\forall s \in B \setminus A \frac{1}{\sum_{t \in B} \frac{(t,s)}{(s,t)}} = 0$ iff $\forall s \in B \setminus A \sum_{t \in B} \frac{(t,s)}{(s,t)} = \infty$ iff $\forall s \in B \setminus A \exists t \in B \frac{(t,s)}{(s,t)} = \infty$ (since $B \subseteq S$ is finite) iff $\forall s \in B \setminus A \exists t \in B\, (s,t) = 0$ (since $(t,s) \neq \infty$ for any s,t) iff $\forall s \in B\, ((\forall t \in B (s,t) \neq 0) \Rightarrow s \in A)$ iff $\forall s \in B\, ((\forall t \in B\, s \leq t) \Rightarrow s \in A)$ iff $Min_{\leq} B \subseteq A$. □

Proof of Proposition 3.1

1. For the first direction, given a dcpm $\mathbf{S} = (S, \mu_a, \Pi_a, \| \bullet \|)$, we take the corresponding priority order as our plausibility: i.e., for all agents and all states, we put

$$s \leq_a t \quad \text{iff } (s,t)_a \neq 0.$$

   We keep the same valuation, and take the epistemic indistinguishability relation $\sim_a$ induced by the partition Π_a:

$$s \sim_a t \quad \text{iff } s(a) = t(a).$$

   This gives us a plausibility model $\overline{\mathbf{S}}$, which by construction has the same state space, same valuation and same notion of knowledge as the original dcpm. To check that the notions of conditional belief are also the same, we note that: a sentence $B_a^P Q$ is true at a state s in the dcpm **S** iff $\mu_a(Q|P \cap s(a)) = 1$ iff $Min_{\leq_a} P \cap s(a) \subseteq Q$ (by Lemma 3.2) iff $B_a^P Q$ is true at the corresponding state s in the plausibility model $\overline{\mathbf{S}}$.
2. For the converse, given a plausibility model $\overline{\mathbf{S}} = (S, \leq_a, \sim_a, \| \bullet \|)_a$, define binary maps $(\bullet, \bullet)_a : S \times S \to [0,1]$, by putting: $(s,t)_a = 1$ if either $s \leq_a t$ or $s = t$; $(s,t)_a = 0$ if $s \geq_a t$; and $(s,t)_a = 0.5$ otherwise. It is straightforward to check that each such map satisfies the axioms of a priority space. By Proposition 2.1, this gives us a Popper measure μ_a on S. So we obtain a dcpm $\mathbf{S} = (S, \mu_a, \Pi_a, \| \bullet \|)$,

where the valuation $\| \bullet \|$ is the same as in the plausibility model $\overline{\mathbf{S}}$ and the partition Π_a is given by the indistinguishability relations $\sim_a$ of the model $\overline{\mathbf{S}}$.
By construction, the dcpm $\mathbf{S}$ has the same state space, same valuation and same notion of knowledge as the original plausibility model $\overline{\mathbf{S}}$. To check that the notions of conditional belief are also the same, we simply note that *the priority order corresponding to the Popper function μ_a coincides with the original plausibility relation $\leq_a$*: in other words, we have $s \leq_a t$ iff $(s,t)_a \neq 0$, for all s, t. Using now the first part of our proof (the proof of part 1), we conclude that the notions of conditional belief are the same. □

Proposition 3.1 is *the main result* of this section, showing that "plausibility" is indeed a *qualitative* notion, which can only capture "*firm*"[10] (*though conditional*) *beliefs*, but *no degrees of belief*: all intermediary binary degrees can be assumed to be equal to 0.5. An immediate consequence of Proposition 3.1 is the following *completeness theorem*:

Corollary 3.3 *The* (*decidable*) *logic CDL introduced in* Baltag and Smets (2006a,b,c, 2008)*, and presented in Appendix* 1 *is sound and complete for dcpm's.*

The *proof* is in Appendix 1.

4 Safe belief and degrees of safety

The *defeasibility analysis* of knowledge (Lehrer 1990), formalized in Stalnaker (1996), is based on the idea that "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). Lehrer and Stalnaker interpret "evidence" as "true information", and thus their "knowledge" differs from ours. Moreover their notion has various *non-standard features*, e.g. negative introspection fails: a rational agent may *believe that she "knows" something* (in their sense), *without knowing it*! This non-standard conception may be common among philosophers, but it is unfamiliar to logicians.

In contrast, our notion of knowledge K_a (as defined in Sect. 2) is just *the standard*, *partition-based* concept of knowledge ("Aumann knowledge"), as it is *commonly used in Logic, Artificial Intelligence and Game Theory*. In particular, our knowledge *is fully introspective* (S5-like), and it is more *robust* than the one of Lehrer and Stalnaker: it is an *"absolute", un-revisable* knowledge, that cannot be defeated by *any* evidence (including false evidence), and it thus satisfies a *stronger* version of the defeasibility analysis (obtained by interpreting "evidence" in the above quote as meaning "any information, be it truthful or not"). Nevertheless, we consider Lehrer's weaker concept to be equally important, and so in Baltag and Smets (2006b,c, 2008) we introduced it in the context of plausibility models, under the name of *"safe belief"*.

[10] I.e. *believed with* (*conditional*) *probability* 1.

Since dcpm's are plausibility models (with "priority" as the plausibility relation), we can use the same definition: the "safe belief" operator is the Kripke modality $\Box_a$ associated to the converse $\trianglerighteq_a$ of the local priority relation,[11] i.e. given by

$$\Box_a Q := [\trianglerighteq_a]Q = \{s \in S : \forall t \in S(t \trianglelefteq_a s \rightarrow t \in Q)\},$$

for all **S**-propositions $Q \subseteq S$. We read $s \in \Box_a Q$ as saying that: *at state s, agent a's belief in Q* (*being the case*) *is safe*; or at *state s, a safely believes that Q*. An important observation is that this notion *does indeed capture the Lehrer-Stalnaker non-standard concept of "knowledge"*:

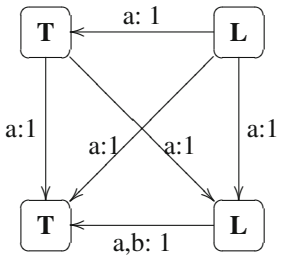
$$s \in \Box_a Q \quad \text{iff } s \in B_a^P Q \quad \text{for all } P \subseteq S \quad \text{such that } s \in P.$$

So *safe beliefs are precisely the beliefs which are persistent under revision with any true information*.

Another important observation, made in Baltag and Smets (2006b) and (2008), is that *conditional belief can be defined only in terms of knowledge and safe belief*: if $\tilde{K}_a P = \neg K_a \neg P$ is the Diamond modality for K, then

$$B_a^P Q \;=\; \tilde{K}_a P \rightarrow \tilde{K}_a(P \wedge \Box_A(P \rightarrow Q)).$$

Example 4 (*Dangerous Learning*) This starts with the situation in Example 1, but involves a form of "cheating". When Alice doesn't pay attention, Bob quickly raises the cover of the box and takes a peek inside, seeing that Charles took the money. Alice doesn't notice this, and she doesn't even suspect this can happen: say, because taking a peek is against the rules of the game, and so she trusts Bob not to do that. The ensuing situation is given by the following model $\mathbf{S}'$



in which we'll denote by t and l the lower nodes (representing the situations in which no "cheating" occurs), and we'll denote by t' and l' the corresponding upper nodes (representing the situations in which Bob "cheats" as described above).

In both Examples 1 and 4 above, Alice holds a *true belief* (at the real state) that Charles took the money: the actual state satisfies $\mathbf{T} \wedge B_a \mathbf{T}$. In both cases, this true

[11] A similar notion was defined in van Benthem and Liu (2004) in a different context, under the name of "preference modality".

belief is *not knowledge* (since Alice doesn't know for sure that he took the money); 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, note that any *new* truthful information would reveal to Alice the real state s, thus confirming her belief that Charles took the money. So in the model $\mathbf{S}$ from Example 1, we have $s \models \Box_a \mathbf{T}$. In contrast, in Example 4, Alice's belief, though true, is *not safe*. There is some piece of correct information which, if learned by Alice, would make her change this belief: we can represent this piece of correct information as the doxastic proposition $\mathbf{T} \rightarrow K_b\mathbf{T}$. Since $\mathbf{T} \rightarrow K_b\mathbf{T}$ is true only at states l, t', l', we have $Min_{\leq_a} t'(a) \cap (\mathbf{T} \rightarrow K_b\mathbf{T}) = Min_{\leq_a}\{l, t', l'\} = \{l\}$; so $B_a^{\mathbf{T}\rightarrow K_b\mathbf{T}}\mathbf{L}$ holds at state t'; i.e., at state t', *if given the information* that "if Charles took the money, then Bob knows it", *Alice would come to wrongly believe that the state l is the real one*, i.e. *that Charles left the money in the box*! This is an example of a *dangerous truth*: a piece of true information whose learning can lead to wrong beliefs.

Degree of safety of a belief. In dcpm's, we can use the probabilistic information to refine our analysis of safe belief, by defining "degrees of safety" of a belief (similarly to the probabilistic degrees of belief). For any number $x \in [0, 1]$, we say that a's belief in Q has a *degree of safety of (at least) x at state s*, and write $s \in \Box_a^x Q$, if a's degree of conditional belief in Q given P is at least x for all true conditions P:

$$\Box_a^x Q = \{s \in S : \mu_a(Q|P \cap s(a)) \geq x \text{ for all } P \text{ such that } s \in P\}.$$

If we define the *degree of safety of a's belief in Q at state s* by

$$d_a^s(Q) := \min_{s\in P\subseteq s(a)} \mu_a(Q|P),$$

then we have:

$$\Box_a^x Q = \{s : d_a^s(Q) \geq x\}.$$

Note that "safe belief" is the same as belief with degree of safety=1.

Weak safety. A belief is "*weakly safe*" if it has degree of safety > 0: such a belief might be lost due to truthful learning, but it is never reversed (into believing the opposite). Indeed, it is easy to see that a's belief in Q is weakly safe iff $\neg B_a^P \neg Q$ for all true propositions P.

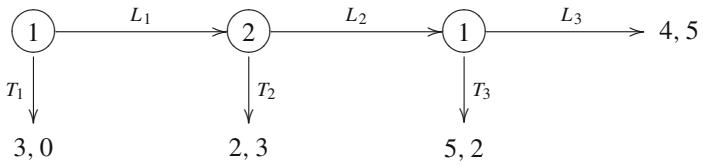
Strongly unsafe beliefs are the ones which (even if true) are not weakly safe (but have a null degree of safety). It is easy to see that, in Example 4 above, Alice's belief that Charles took the money is strongly unsafe. In many situations, it is enough to have a *high enough degree of safety* to pre-empt the potential dangers of learning.

Degree of common safe belief. The standard notion of common belief can be extended to define a concept of "common safe belief of degree x " (for $0 \leq x \leq 1$):

$$C\Box^x P = \bigwedge_{a_1,a_2,\ldots,a_n} \Box_{a_1}^x \Box_{a_2}^x \ldots \Box_{a_n}^x P$$

(where the infinite conjunction ranges over all sequences $a_1, \ldots, a_n$ of agents). As for safe belief, we define *common safe belief* $C\Box P$ simply as "common safe belief of degree 1" $C\Box^1 P$. One could argue that the above notions should play an important role in games: for instance, Aumann's theorem (Aumann 1995) about backwards induction still holds if we weaken his condition of "common knowledge of rationality" to "common safe belief of rationality". Aumann's celebrated result says that, if a state s satisfies *common knowledge of rationality* CkR, then s is the backwards induction solution. It is well-known that the theorem does *not* hold in the weaker assumption of *common true belief in rationality*. The reason is that players' beliefs about the other players' rationality *may change* during the game: they may lose belief in others' rationality if some "surprising" moves are made.

Example (*The Centipede Game*) As an example of what can go wrong, consider the following Centipede Game:



We denote the nodes by the sequences of moves leading to them, e.g. the original node is $\emptyset$, where player 1 is to move. A *model for a game* G is a dcpm $\mathbf{S}$, whose states are strategy profiles for G. Thus, a state s uniquely determines the set of nodes G_s that are reachable during a play at state s. We say that *in state s, player i will come to prefer (another strategy)* t_i *(to his current one* s_i*) at node* v if there is a reachable node $v' \in G_s$, $v' \leq v$ such that for all reachable nodes $v'' \in G_s$ with $v' \leq v'' \leq v$, the expected utility of playing t_i at v conditioned by the information that v'' is reached is bigger than the expected utility of playing the current strategy s_i at v, conditioned by the same information. A player is *rational* in state s if he knows that he will not come to prefer any other strategy (to his current one) at any node. "Rationality" is the sentence R saying that all players are rational.

Board provides a good analysis of the above game in terms of conditional beliefs (Board 2004). In a nutshell, there exist models for this game such that the state (L_1T_3, L_2) satisfies in the same time B_2T_1, $B_2^{L_1L_2}L_3$ and common true belief in rationality. Intuitively, this is because player 2's original belief that player 1 will play T_1 (belief which is fully consistent with common true belief in rationality) is challenged by the surprising move L_1. To respond to this challenge, 2 must revise his beliefs, and it is perfectly possible (and consistent with 2's rationality) that after this revision, 2 starts to believe that 1 is so "irrational" that he will play L_3, if the node L_1L_2 is reached. Given this belief, 2's best response at node L_1 is to play L_2. This explains the failure of backwards induction solution at the state (L_1T_3, L_3).

To warrant the backwards induction solution, we would need common belief in rationality at all nodes that are *actually reachable*. To make this concept robust, the only uniform assumption that we can make at the original node $\emptyset$ seems to be *common safe belief of rationality*. Using degrees of safety, we can refine this somewhat:

Proposition 4.1 *If common safe belief in rationality holds at the initial state of a game of perfect information, then the backwards induction solution is played. Moreover, for any game of perfect information G, there exists a number $x > 0$ such that: if at the initial state of game G the degree of common safe belief in rationality is strictly bigger than x, then the backwards induction solution is played.*

The *proof* is given in detail in the Masters Thesis (Mihalache 2007) of the first author's graduate student Dan Mihalache. Here, we only mention this result as an example of application of our notions.

Safety level. The smallest such x is called the *safety level* of game G, and is computable in terms of the game tree G. For instance, common belief in rationality corresponds to a safety level of 1. In the Centipede Game above, the safety level is $2/3$: it is enough for player 2's initial belief in 1's rationality to have a degree of safety of more than $2/3$ (i.e. he would always be cautious enough to assign only a probability of less than $1/3$ to other player's irrationality, no matter what new information he learns). This would give player 2 an expected utility strictly less than $4/3+5/3 = 3$ if he plays L_2, i.e. less then the expected utility for playing T_2, thus motivating him to play the backwards induction solution. (But, for all he knows, player 2 might still be *wrong*: player 1 might actually *be irrational*, in which case player 2 misses a good opportunity to make a lot of money!)

Proposition 4.2 *The logic $K\Box$ of knowledge and safe belief,* introduced in Baltag and Smets (2006c, 2008) and presented in Appendix 2*, is sound and complete with respect to dcpm's.*

The *proof* is sketched in Appendix 2.

5 Action models and conditional probabilistic update

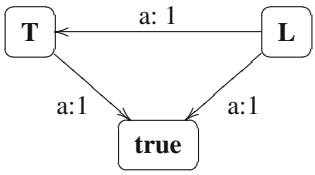
We now improve on the work in Aucher (2003), van Ditmarsch (2005), Baltag and Smets (2006b,c) by introducing *action models*, of both the conditional-probabilistic and the plausibilistic type, in order to represent *uncertain forms of multi-agent belief-updating actions*:

A (*discrete conditional-probabilistic, or finite plausibility*) *action model* is just a (discrete conditional-probabilistic, or finite plausibility) frame Σ, together with a *precondition map*

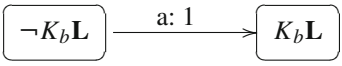
$$pre\colon \Sigma \to Prop$$

associating to each element of Σ some doxastic sentence pre_σ. As in Baltag and Smets (2006b,c), we call the elements of Σ (*basic*) *doxastic actions*, and we call pre_σ *the precondition of action* σ. 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 basic actions $\sigma \in \Sigma$ are taken to represent some *deterministic* actions of a particularly simple nature. As mentioned in the Introduction, 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). The conditional probabilities μ_a, or the plausibility pre-orderings $\trianglelefteq_a$, give *the agent's* (*probabilistic, conditional*) *beliefs about the current action*.

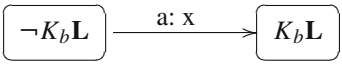
Example 4′ Let us revisit Example 4, and think of the *action* leading to it: Bob takes a peek inside the box, when Alice doesn't pay attention. In the Dynamic Epistemic Logic literature, this action is usually called *a private announcement to a subgroup*: the "insider" (Bob) learns what is in the box, while the outsider Alice *believes that "nothing" is happening*. In the action model $\mathbf{\Sigma}$, the two nodes on top represent the actions of Bob taking a peek into an empty box or him peeking at the million dollars. These actions have as preconditions the sentences **T** (saying that Charles took the money), and respectively **L**. The node on the bottom represents the action in which "nothing is happening" (whose precondition is any tautology **true**):



Example 5 (*Fully Successful Lying*) Suppose now that, *after* Bob secretly took a peek in the box (i.e. in the situation from Example 4), Bob sneakily announces: "Look, I took a peek and *saw the money inside*". For our purposes, we can formalize the content of this announcement as $K_b\mathbf{L}$ ("Bob knows Charles left the money in the box"). This is a *public, but un-truthful announcement*: a lie! Let's assume that it is a *fully successful lie*: (it is common knowledge that) Bob's speech act is so *persuasive* that Alice believes him. This action is given by the *left node* in the model Σ' below:



Example 5′ (*Partially Successful Lying*) In contrast, if Bob's speech is *not fully persuasive*, so that Alice only assigns probability $0 < x < 1$ to him telling the truth, then the action model is:



Probabilistic update product. To compute the output state model from the original (input) state model and the model of the action, we need a binary *'update' operation* $\otimes$, taking any state model $\mathbf{S} = (S, \trianglelefteq_a, \| \bullet \|)_{a\in\mathcal{A}}$ and any action model $\mathbf{\Sigma} = (\Sigma, \trianglelefteq_a, pre)_{a\in\mathcal{A}}$ into a *new state model* $\mathbf{S}\otimes\mathbf{\Sigma}$, representing the possible output-states of executing some action from $\mathbf{\Sigma}$ on some input-state from **S**. We call this the *update product* of the two models. As in Baltag and Moss (2004), we take *the set of states of the new model*[12] to be a subset of the Cartesian product of the two models, given by the consistent pairs:

[12] The justification is that: (1) basic actions are deterministic, so we can identify their outputs with pairs (s, σ) of an input and an action; and (2) an action is executable only on inputs that satisfy its precondition.

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

For simplicity, we denote by $s\sigma$ the pair (s, σ) seen as an output-state in $S \otimes \Sigma$. As in Baltag and Moss (2004), the *valuation* is left unchanged[13]:

$$s\sigma \models p \quad \text{iff } s \models p\,;$$

the *new indistinguishability relation* is the *product* of the two old relations[14]

$$s\sigma \sim_a s'\sigma' \quad \text{iff } \sigma \sim_a \sigma', s \sim_a s';$$

and the *new conditional probability* is given by putting

$$(s\sigma, t\tau) \;=\; \lim_{x \to (s,t)} \frac{x \cdot (\sigma, \tau)}{x \cdot (\sigma, \tau) + (1 - x) \cdot (\tau, \sigma)},$$

where the limit is taken over x's such that the denominator is $\neq 0$.

Justification. This last clause can be justified by thinking of what doxastic actions *do*: they are actions that are meant to *change* the prior beliefs, via new, independent evidence that may *possibly override the prior beliefs* (when they are contradicted by the new evidence). The prior probability is not to be necessarily kept unchanged, since it reflects past beliefs, while the action probability represents the agent's *current beliefs*. Independence implies multiplication of probabilities, except that in the case of contradiction (denominator 0), the prior beliefs are not assumed to be firmly held, but are prone to small errors: so we extend the definition by continuity to the case that the denominator is 0. As in the AGM theory, an agent keeps as much as possible of his prior (certain) beliefs, as long as they are not contradicted by the new (certain) beliefs: *prior certainty can only be overridden if it is contradicted by current certainty; in addition, current certainty always overrides prior uncertainty; while prior uncertainty is "weighted" using the current uncertainty*.[15]

Spelling out the effect of the last clause in detail, we obtain the following equivalent definition by cases:

If $(\sigma, \tau) = 0$, then

$$(s\sigma, t\tau) \;=\; 0;$$

if $(\sigma, \tau) = 1, \sigma \neq \tau$, then

$$(s\sigma, t\tau) \;=\; 1;$$

[13] This is because we only consider "purely doxastic" actions, so the ontic "facts" are left unchanged.

[14] This encodes the intuition that the initial uncertainty about actions is *independent* of the initial uncertainty about states.

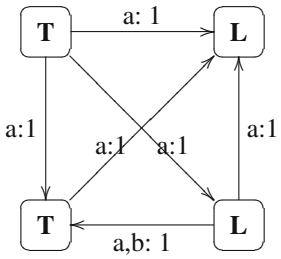
[15] A different justification can be provided using a generalization of Jeffrey's Rule to Popper probabilities.

otherwise

$$(s\sigma, t\tau) = \frac{(s,t)\cdot(\sigma,\tau)}{(s,t)\cdot(\sigma,\tau)+(1-(s,t))\cdot(1-(\sigma,\tau))}.$$

In particular, the first two cases above give us the *Action-Priority Update* from Baltag and Smets (2008) (also called *anti-lexicographic product update* in Baltag and Smets (2006b)): this says that the plausibility relation on the Cartesian product $\mathbf{S}\times\mathbf{\Sigma}$ is the anti-lexicographic preorder induced by the two plausibility preorders.

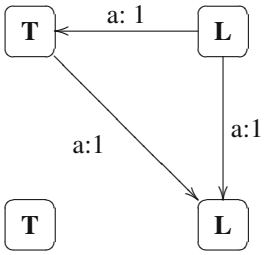
Example 5, 5′ revisited We can see the qualitative difference between "*fully successful lying*" (which may completely overturn the agent's prior beliefs) and the only "*partially successful lying*" (which only "weights" these prior beliefs), by comparing the way the actions in Examples 5 and 5′ update the model $\mathbf{S}'$ in Example 4: the first action yields



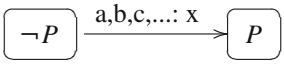
while the second action leaves model $\mathbf{S}'$ essentially unchanged! So a "partially successful" lie can enhance the hearer's doubts *only when she already had some doubts* to start with. *It cannot override the hearer's prior certainty; but a "fully successful" lie always does*!

Other examples of update products. It is easy to see that the update product of the state model in Example 1 and the action model in Example 4′ is indeed (as expected) the state model in Example 4.

Public announcements of " hard facts"= conditionalization. A *truthful public announcement* $P!$ *of some "hard fact"* $\mathbf{P}$ establishes *common knowledge* that P was the case. The action model consists of only one node, whose precondition is P. Its effect on a state model $\mathbf{S}$ (via the update product) is to *delete all the non-P states, keep the indistinguishability relations between the surviving states and change the probabilities by conditionalizing with P*: i.e. $\mu'_a(Q|R) := \mu_a(Q|R\cap P_{\mathbf{S}})$, where $P_S = \{s\in S : s\models_{\mathbf{S}} P\}$. A *concrete example* of this is *publicly announcing*, in the situation from Example 4, that *if Charles then Bob knows it*. This corresponds to $(\mathbf{T}\rightarrow K_b\mathbf{T})!$, and the updated model is:

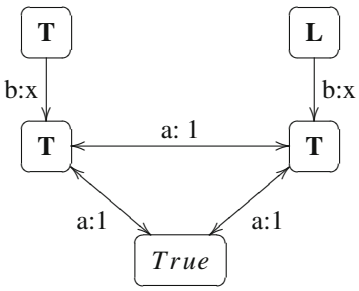


Public announcement of "soft" facts. Suppose an announcement $P!?_x$ is made, in such a way that all the agents *believe* with probability x it is truthful, although they *don't know* for sure that it is truthful.



Note that the effect of such a "soft" announcement is different from the previous "hard" announcement, *even when* $x = 1$. The case $x = 1$ has in fact been considered by other authors, who proposed a notion of "soft update" for it. But it is easy to see that its effect matches what we get by updating any given state model $\mathbf{S}$ with the action $P!?_1$ using our notion of product update: The new state model $\mathbf{S} \otimes P!?$ can be thought of as being obtained from $\mathbf{S}$ by keeping the same information cells, and keeping the same priority order $s \leq t$ between any two states $s, t \in P$, and similarly between states $s, t \notin P$, while in the same time giving to all P-states priority with respect to all non-P states.

Discovery of deceit. Suppose that, in fact, when Bob *thinks* that he is "secretly" taking a peek, Alice *does* pay attention, so that she *notices* that Bob is taking a peek. Suppose it is common knowledge that Bob *doesn't know that she noticed his peeking*, but that he *does* consider this as *a* (*very remote*) *possibility*. More precisely, he believes with some high probability x that there is *Alice didn't notice anything*, and so he assigns the very small probability $1 - x$ to the possibility of her noticing his peeking. The action model is:



Interception of messages. If secret learning (Bob taking a peek in the box) is replaced by a *secret communication* (from Charles to Bob, telling him that he took the money), then the above action model for "discovery of deceit" by Alice (as above) can also be interpreted as a *secret interception (wiretapping) by Alice of the secret message*.

Proposition 5.1 The *dynamic logic of belief-changing actions* presented in Appendix 3, and having the same syntax and proof system as the one in Baltag and Smets (2006b, 2008), *is sound and complete with respect to dcpm's.*

The *proof* is briefly sketched in Appendix 3.

Future work. We list here only three important open problems: (1) Axiomatize the corresponding logics for *infinite* conditional probability models. (2) Study the logics obtained by adding *quantitative* modal operators $\Box_a^x Q$ (and $B_a^{P,x} Q$) expressing that the degree of safety of the belief in Q (or the degree of conditional belief in Q given P) is at least x. (3) Axiomatize the *logic of common safe belief* $C\Box Q$ and its quantitative version $C\Box^x Q$.

Acknowledgements This paper is an extended version of a LORI'07 presentation (Baltag and Smets 2007), itself an extension of a TARK'07 abstract. We thank the organizers of LORI'07, and in particular Fenrong Liu, for their help in preparing this paper for publication. We thank Johan van Benthem, Jelle Gerbrandy, Aviad Heifetz, Dan Mihalache and the anonymous referees for their useful comments. Sonja Smets' contribution to this research was made possible by a post-doctoral fellowship from the Flemish Fund for Scientific Research.

Appendix 1: A complete proof system for *CDL*

The *syntax* of *CDL* is:

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

while the *semantics* over dcpm's is given by the obvious compositional clauses (using the operators $B_a^P Q$ and $K Q$ defined in the paper). Note that here, the *knowledge modality* is a derived operator, defined by putting $K_a\varphi := B_a^{\neg\varphi}\varphi$.

A *doxastic proposition* is a map **P** assigning to each dcpm **S** some **S**-proposition, i.e. a set of states $\mathbf{P_S} \subseteq S$. So the interpretation map for the logic *CDL* associates to each sentence φ of *CDL* a doxastic proposition $||\varphi||$. We denote by *Prop* the family of all doxastic propositions.

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: $\vdash B_a^{\theta}\varphi \to K_a B_a^{\theta}\varphi, \vdash \neg B_a^{\theta}\varphi \to K_a\neg B_a^{\theta}\varphi$
Hypotheses are (hypothetically) accepted: $\vdash B_a^{\varphi}\varphi$
Minimality of revision: $\vdash \neg B_a^{\varphi}\neg\psi \to (B_a^{\varphi\wedge\psi}\theta \leftrightarrow B_a^{\varphi}(\psi \to \theta))$

Proof of Corollary 3.3 (Soundness and Completeness) In Baltag and Smets (2008), we proved the soundness, completeness and finite model property of *CDL* with respect to *plausibility models*, and hence completeness over *finite plausibility models*. This, together with part 2 of Proposition 3.1, gives us completeness with respect to dcpm's, while part 1 of Proposition 3.1 gives us soundness on dcpm's. □

Appendix 2: The logic of knowledge and safe belief

The *syntax* of the logic $K\Box$ of knowledge and safe belief is:

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

while the *semantics* over dcpm's is given by the obvious compositional clauses. Note that in this logic, *(conditional) belief is a derived operator*, defined as $B_a^{\varphi}\psi := \tilde{K}_a\varphi \rightarrow \tilde{K}_a(\varphi \wedge \Box_A(\varphi \rightarrow \psi))$, where $\tilde{K}_a\varphi := \neg K_a\neg\varphi$ is the Diamond modality for K.

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

Necessitation Rule for K and $\Box$: From $\vdash \varphi$ infer $\vdash K_a\varphi$ and $\vdash \Box_a\varphi$;
K-axiom for K_a *and* $\Box_a$;
S5-axioms for K_a;
S4-axioms for $\Box_a$;
$K_a P \rightarrow \Box_a P$;
$K_a(P \vee \Box_a Q) \wedge K_a(Q \vee \Box_a P) \rightarrow K_a P \vee K_a Q$.

Proof of Proposition 4.2 (Soundness and Completeness) Soundness results from putting together the soundness proof on plausibility models in Baltag and Smets (2008) with part 1 of Proposition 3.1. Completeness similarly follows from the results in Baltag and Smets (2008) (completeness and finite model property for plausibility models) together with part 2 of Proposition 3.1 above. □

Appendix 3: The dynamic logic of belief-changing actions

Dynamic Modalities. Given a doxastic action σ (living in some action model $\mathbf{\Sigma}$, we can define a corresponding dynamic modality, capturing the *weakest precondition* of σ: for every proposition $\mathbf{P}$, the proposition $[\sigma]\mathbf{P}$ is given by

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

Syntax of dynamic logic of doxastic actions. This was briefly sketched in Baltag and Smets (2006c): As in Baltag and Moss (2004), we consider a doxastic *signature*, i.e. a *finite* (*fixed*) *plausibility frame* $\mathbf{\Sigma}$, together with an *ordered list without repetitions* $(\sigma_1, \ldots, \sigma_n)$ of some of the elements of $\mathbf{\Sigma}$. Each signature gives rise to a dynamic-doxastic *logic* $L(\mathbf{\Sigma})$, as in Baltag and Moss (2004): one defines by double recursion a set of *sentences* φ and a set of *program terms* π; the *basic programs* are of the form $\pi = \sigma\vec{\varphi} = \sigma\varphi_1 \ldots \varphi_n$, where $\sigma \in \Sigma$ and φ_i are sentences in our logic; program terms are generated from basic programs using *non-deterministic sum (choice)* $\pi \cup \pi'$ and *sequential composition* $\pi; \pi'$. Sentences are built using the operators of the logic $K\Box$ above, and in addition a dynamic modality $[\pi]\varphi$, taking program terms and sentences into other sentences. As in Baltag and Moss (2004), the plausibility preorders on the signature $\mathbf{\Sigma}$ induce in a natural way plausibility preorders on basic programs

in $CDL(\mathbf{\Sigma})$: we put $(\sigma\vec{\varphi})_a^{\Pi\vec{\varphi}} := \{\sigma'\vec{\varphi} : \sigma' \in \sigma_a^{\Pi}\}$. The given listing can be used to assign syntactic preconditions for basic programs, by putting: $pre(\sigma_i\vec{\varphi}) := \varphi_i$, and $pre(\sigma\vec{\varphi}) := \top$ (the trivially true sentence) if σ is not in the listing. Thus, the basic programs of the form $\sigma\vec{\varphi}$ form a (finite) *syntactic plausibility model*[16] $\mathbf{\Sigma}\vec{\varphi}$. Every given interpretation $||\bullet|| : L(\mathbf{\Sigma}) \to Prop$ of sentences as doxastic propositions will convert this syntactic model into a "real" (semantic) plausibility model, called $\mathbf{\Sigma}||\vec{\varphi}||$.

To give the *semantics*, choose any dcpf $(\mathbf{\Sigma}, \mu)$ whose priority frame is isomorphic to $\mathbf{\Sigma}$. We define by induction two *interpretation maps*, one taking any sentence φ to a doxastic proposition $||\varphi|| \in Prop$, the second taking any program term α to a (possibly non-deterministic) doxastic "program", i.e. a *set* of basic actions in some dcpf. The definition uses the obvious semantic clauses and is completely similar to the one in Baltag and Moss (2004).

The *proof system* is obtained by adding to the logic $K\Box$ the following "Reduction Axioms":

$$
\begin{aligned}
[\alpha]p &\leftrightarrow pre_\alpha \to p \\
[\alpha]\neg\varphi &\leftrightarrow pre_\alpha \to \neg[\alpha]\varphi \\
[\alpha](\varphi \wedge \psi) &\leftrightarrow pre_\alpha \to [\alpha]\varphi \wedge [\alpha]\psi \\
[\alpha]K_a\varphi &\leftrightarrow pre_\alpha \to \bigwedge_{\beta \sim_a \alpha} K_a[\beta]\varphi \\
[\alpha]\Box_a\varphi &\leftrightarrow pre_\alpha \to \bigwedge_{\alpha \lhd_a \beta} K_a[\beta]\varphi \wedge \bigwedge_{\alpha \cong_a \gamma} \Box_a[\gamma]\varphi \\
[\pi \cup \pi']\varphi &\leftrightarrow [\pi]\varphi \wedge [\pi']\varphi \\
[\pi;\pi']\varphi &\leftrightarrow [\pi][\pi']\varphi
\end{aligned}
$$

where p is any atomic sentence, π, π' are programs and α is an *action*, i.e. a *basic* program in $L(\mathbf{\Sigma})$, $\sim_a$ is epistemic indistinguishability between actions, $\lhd_a$ is *strict plausibility* order on actions, while $\cong_a$ is *equi-plausibility* of (indistinguishable) actions: $\alpha \cong_a \beta$ iff both $\alpha \trianglelefteq_a \beta$ and $\beta \trianglelefteq_a \alpha$.

Proof of Proposition 5.1 (*Soundness and Completeness*) Soundness is an easy exercise. Completeness follows from the completeness of the $K\Box$ logic (Proposition 4.2) together with the results in Baltag and Smets (2008), where the above Reduction Laws were used inductively to show that any formula in the logic $L(\mathbf{\Sigma})$ can be "reduced" (i.e. it is provably equivalent) to a formula in the $K\Box$-logic. □

References

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

[16] A *syntactic plausibility model* is just a plausibility frame endowed with a *syntactic* precondition map, associating sentences to basic actions. For justification and examples, in the context of *epistemic* action models, see Baltag and Moss (2004).

Arlo-Costa, H., & Parikh, R. (2005). Conditional probability and defeasible inference. *Journal of Philosophical Logic, 34*, 97–119.

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

Aumann, R. (1995). Backwards induction and common knowledge of rationality. *Games and Economic Behavior, 8*, 6–19.

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. (2004). Logics for epistemic programs. *Synthese, 139*, 165–224. Knowledge, Rationality & Action (pp. 1–60).

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

Baltag, A., & Smets, S. (2006a). 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 W. van der Hoek & M. Wooldridge (Eds.), *Proceedings of LOFT'06* (pp. 11–24). Liverpool.

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

Baltag, A., & Smets, S. (2007). Probabilistic dynamic belief revision. In J. van Benthem, S. Ju, & F. Veltman (Eds.), *Proceedings of LORI'07* (pp. 21–39). London: College Publications.

Baltag, A., & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, & M. Wooldridge (Eds.), *Texts in logic and games* (vol. 3). Amsterdam: Amsterdam University Press.

Board, O. (2004). Dynamic interactive epistemology. *Games and Economic Behaviour, 49*, 49–80.

Boutilier, C. (1995). On the revision of probabilistic belief states. *Notre Dame Journal of Formal Logic, 36*(1), 158–183.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: Bradford Books, MIT Press.

Gerbrandy, J. (1999). *Dynamic epistemic logic. Logic, language and information* (Vol. 2). Stanford: CSLI Publications, Stanford University.

Halpern, J. (2001). Lexicographic probability, conditional probability, and nonstandard probability. In *Proceedings of the Eighth Conference on Theoretical Aspects of Rationality and Knowledge (TARK 8)* (pp. 17–30).

Halpern, J. (2003). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.

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

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

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

Mihalache, D. (2007). *Safe belief, rationality and backwards induction in games*. Masters Thesis in Computer Science. Oxford: Oxford University.

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

Popper, K. (1968). *The logic of scientific discovery* (*revised Edition*) (1st ed., 1934). London Hutchison.

Renyi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungaricae, 6*, 285–335.

Renyi, A. (1964). Sur les espaces simples des Probabilites conditionnelles. *Annales de L'institut Henri Poincare: Section B: Calcul des probabilities et statistique, B1*, 3–21.

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

van Benthem, J. (2003). Conditional probability meets update logic. *Journal of Philosophical Logic, 12*(4), 409–421.

van Benthem, J. (2006). Dynamic logic of belief revision. Technical Report, ILLC, DARE electronic archive, University of Amsterdam. To appear in *Journal of Applied Non-Classical Logics*.

van Benthem, J., Gerbrandy, J., & Kooi, B. (2006a). Dynamic update with probabilities. In W. van der Hoek & M. Wooldridge (Eds.), *Proceedings of LOFT'06*. Liverpool.

van Benthem, J., & Liu, F. (2004). Dynamic logic of preference upgrade. Technical report. ILLC Research Report PP-2005-29.

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

van Ditmarsch, H. (2005). Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action), 147*, 229–275.

van Fraassen, B. (1976). Representations of conditional probabilities. *Journal of Philosophical Logic, 5*, 417–430.

van Fraassen, B. (1995). Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical Logic, 24*, 349–377.