Epistemic Proactivity via Reasoning about Beliefs and Expectations using Plausibility Models

Thomas Bolander[1*] and Hermine J. Grosinger[2*†]

[1*]Technical University of Denmark, Copenhagen, Denmark.
[2*]Örebro University, Örebro, Sweden.

*Corresponding author(s). E-mail(s): tobo@dtu.dk;
hermine.grosinger@oru.se;
†These authors contributed equally to this work.

Abstract

We present a formalism that allows to distinguish between, on the one hand, *beliefs* about actions that have already occurred and their resulting states (*a posteriori* beliefs), and, on the other hand, *expectations* about actions that have not yet occurred and their resulting future states (*a priori* beliefs). We use plausibility models within Dynamic Epistemic Logic (DEL) to model beliefs, expectations and judgments of plausibility. Having such a formalism in place, we can reason about an agent's false beliefs and false expectations, and when to make belief updates (*relevant* announcements) so future undesirable situations may be avoided. A potential application of our framework is human-robot interaction. Based on reasoning about the human's false expectation, a *proactive* robot can autonomously decide when and what to announce to help avoiding that the human ends up in an undesirable state.

Keywords: Dynamic Epistemic Logic, Plausibility models, Expectations, Belief Update, Proactivity, Human-Robot Interaction

Acknowledgements

This research is funded by the Swedish Research Council (Vetenskapsrådet), by projects 2021-05542 and 2023-04349, on H.J. Grosinger's part. Thomas Bolander is supported by the Independent Research Fund Denmark (grant no. 10.46540/4258-00060B). We are very grateful to the reviewers for suggesting improvements, and to

Alexandru Baltag for assistance with some of the crucial steps of the completeness proof.

1 Introduction

We want to model agents' expectations about future states, including potentially false expectations which should be correctable by belief updates. These expectations concern what is believed to hold in the future, after a sequence of actions have occurred. Actions can be deterministic or non-deterministic. For a non-deterministic action, an agent might expect some outcomes to be more plausible than others. For instance, she might believe a certain coin to be biased towards heads, so when tossed, she expects it to end up showing heads. For a deterministic action, the outcome is given, but it might depend on facts that an agent doesn't know about, but only has beliefs about.

The paper focuses on potential application scenarios within human-robot interaction where a robot ('Rob') reasons about the beliefs and knowledge of a human ('Hanna'). The role of Rob is to *proactively* help Hanna in situations where she has false beliefs or false expectations that might result in undesirable outcomes. For example, consider Hanna having a false belief that a sugar dispenser contains *sugar* (it contains *salt*) and that the car parked behind her house is *red* (it is *blue*). Should Rob inform her about *all* her false beliefs, or only some of them, and then which? If Hanna is going to put the content of the dispenser into her coffee, then it might be *relevant* to correct the false belief of *sugar*. But if Hanna skips coffee, it is not (and might even be regarded meddlesome). Correcting *red* might never be relevant. Hence, for Rob to decide which announcements to make to Hanna, it is not only important to reason about whether they correct false beliefs (or false expectations), but also whether those false beliefs would lead to *undesirable* outcomes if not corrected. Timeliness is crucial here. If the belief update is done when Hanna tastes salty coffee, it's not helpful as she then already knows *salt*, and the undesirability can no longer be prevented.

In this paper, we will introduce a framework for modeling all the mentioned types of actions as well as agents' beliefs and expectations about them (including higher-order beliefs and expectations). A very general framework for modeling actions and knowledge/beliefs about actions is the event models of *Dynamic Epistemic Logic (DEL)* [1, 2]. Since we are in this paper mainly interested in modeling beliefs, expectations and judgments of plausibility, we will be using the particular framework of *plausibility models* within DEL. Plausibility models were introduced by Baltag and Smets [3]. Their formal framework has almost all the ingredients we need for our purposes. However, to be able to distinguish between an agent's current beliefs about a possible future state (a priori beliefs) and the knowledge and beliefs that the agent will end up having in that future state (a posteriori beliefs), we need to extend and generalize their formalism. We will present our generalized formalism, and illustrate how it can be used to model expectations (including false expectations) about future outcomes in a natural way.

As mentioned, our focus is to apply our formalism to human-robot interaction scenarios, more precisely, *proactive* robot behaviour in such scenarios. We base our understanding of AI proactivity on Grosinger [4], who defines it as the ability to

"*autonomously initiate anticipatory action based on reasoning, meant to impact people and/or their environments*". Proactivity is also a prominent property of humans [5]. Implementing it in AI systems may therefore make collaboration with humans more natural. Recent works [6–8] aim at achieving proactive AI systems by reasoning about the false belief of humans, which witness the relevance of this research direction.

We make the following contributions: (i) We introduce a variant of DEL with plausibility models having both *a priori* and *a posteriori* belief operators (our a priori belief operator will be referred to as an *expectation operator*). (ii) We axiomatize the logic, proving soundness and completeness. (iii) We use the introduced logic to formally define what it means for an announcement to be *relevant* (now or in the future). (iv) We illustrate how to use the formalism to decide which announcements to make to prevent *undesirable* outcomes. (v) We prove that deciding whether a relevant announcement exists, and computing one if it does, can both be done in polynomial time in the size of the updated model.

2 Formal framework

We will use a version of the *multiagent plausibility models* by Baltag and Smets [3, 9]. Both our language and semantics will differ somewhat from theirs in order to fit our purposes. In particular, as we will use the models to capture both a priori beliefs about future action outcomes as well as a posteriori knowledge gained in those outcomes, we will include both an a priori plausibility relation and an (a posteriori) indistinguishability relation. This was also originally considered in [3], but they quickly turned to only consider a posteriori knowledge and belief.

2.1 Plausibility models

All our languages and models will be defined relative to a set P of (propositional) atoms and a set Ag of agents. Whenever no confusion can arise, the dependency on P and Ag will be left implicit. Given a set X and a relation $\leq$ on X, the set of *least* elements of X is defined by $\mathrm{Min}_{\leq} X := \{x \in X \mid x \leq x' \text{ for all } x' \in X\}$ [9].[1] A *well-preorder* on X is a reflexive, transitive relation $\leq$ s.t. every non-empty subset has least elements, i.e., for all non-empty $Y \subseteq X$, $\mathrm{Min}_{\leq} Y \neq \emptyset$. We write $x < y$ when $x \leq y$ and $y \not\leq x$, and $x \simeq y$ when both $x \leq y$ and $y \leq x$. We write $x \lessgtr y$ when x and y are comparable by $\leq$, that is, when either $x \leq y$ or $y \leq x$. Let $\leq \subseteq X \times X$ be a well-preorder and suppose $x, y \in X$. Since every non-empty subset of X has least elements, also the subset $\{x, y\}$ must have least elements, i.e., we must either have $x \leq y$ or $y \leq x$, and hence $x \lessgtr y$. That is, on well-preorders, all elements are pairwise comparable, i.e., well-preorders are *connected*.

Definition 2.1. *A* plausibility model *is* $\mathcal{M} = (W, \sim_i, \leq_i, L)_{i \in Ag}$ *where*

- W *is a finite set of* worlds.

[1] We use the notation Min for the least elements as in the cited paper. In the cited paper, they call the elements of $\mathrm{Min}_{\leq} X$ the *minimal* elements of X, but we are here following the more standard terminology of calling them the *least* elements. We also do this to avoid confusion when later considering minimal elements in the standard sense: x is *minimal* when there exists no $x' < x$. We however still use $\mathrm{Min}_{\leq} X$ to denote the set of least elements of X, to be consistent with the existing literature on plausibility models. When working on well-preorders, as we will do here almost exclusively, the set of least and minimal elements coincide.

- $\leq_i$ *is a reflexive relation on* W *called the* (a priori) plausibility relation. *We further require that* $\leq_i$ *is a union of mutually disjoint well-preorders.*
- $\sim_i$ *is an equivalence relation on* W *called the* (a posteriori) indistinguishability relation. *We further require that* $w \sim_i v$ *implies* $w \leqslant_i v$.
- $L : W \to 2^P$ *is a* labelling function *mapping each world to a finite set of atoms: the atoms true at the world.*

A pair $s = (\mathcal{M}, W_d)$ *with* $W_d \subseteq W$ *is called a* (doxastic) state*, with* W_d *being the set of* designated worlds. *A doxastic state is called* single-pointed *if* $|W_d| = 1$*, and* multi-pointed *if* $|W_d| > 1$. *Single-pointed states* $(\mathcal{M}, \{w\})$ *are often written* $(\mathcal{M}, w)$*, and we call* w *the* designated world. *We use* $w \in \mathcal{M}$ *to denote that* w *is a world of* $\mathcal{M}$.

Note that our plausibility models are finite by definition, which is required for our later computational complexity results. Note also that we allow W to be the empty set, which is slightly non-standard, but it helps simplifying some of the later definitions and proofs. We will use doxastic states to represent possible future states, and what is believed and expected about those states. We use single-pointed states $(\mathcal{M}, w)$ to represent the situation where the actual future world is assumed to be w. We will be using multi-pointed states $(\mathcal{M}, W_d)$ to represent the perspective on a future state by an agent; in this case, the set W_d represents the set of worlds that the agent believes the actual future world will belong to. Multi-pointed states can be used to represent uncertainty about future outcomes, for instance a state representing the possible outcomes of a coin toss, where both the world representing *heads* and the world representing *tails* are designated to represent that it is not a priori known which outcome will be the actual. We read $w \leq_i v$ as "w is at least as plausible as v for agent i", $w <_i v$ as "w is (strictly) more plausible than v for agent i", and $w \simeq_i v$ as "w is equi-plausible to v for agent i". Given $W' \subseteq W$ and $i \in Ag$, the elements of $\mathrm{Min}_{\leq_i} W'$ are hence called the "most plausible worlds" of W' for agent i. As in [3], the relation $\leq_i$ is intended to capture the a priori beliefs of agent i (but note that in almost all other papers on plausibility models, $\leq_i$ is intended to capture the a posteriori (current) beliefs of agent i).

We read $w \sim_i v$ as "w is (a posteriori) indistinguishable from v for agent i". If we have $w \sim_i v$ in a doxastic state s representing a possible future situation, it means that even when arriving at that future situation, w and v are still indistinguishable to agent i. It could e.g. be that we use s to represent what is known and believed by the agent Rob (denoted by r) about the consequences of tossing a coin. If w represents the outcome *heads* and v represents the outcome *tails*, we would use $w <_r v$ to represent that w is considered a priori more plausible than v to Rob, representing that Rob believes the coin to be biased towards heads. We would use $w \sim_r v$ to represent that even after the coin has been tossed, the two outcomes are not distinguishable to Rob (e.g. if the coin is tossed inside a dice cup). As $\sim_i$ is an equivalence relation, we can use standard notation $[w]_{\sim_i}$ for the $\sim_i$-equivalence class (*epistemic equivalence class*) of the world w, i.e., we have $[w]_{\sim_i} = \{v \in W \mid v \sim_i w\}$. Since $\leq_i$ is a union of mutually disjoint well-preorders on W, and since on well-preorders all elements are pairwise comparable, the relation $\leqslant_i$ must be an equivalence relation on W. We define $[w]_{\leqslant_i} = \{v \in W \mid v \leqslant_i w\}$. Since we have required that $w \sim_i v$ implies $w \leqslant_i v$, we have $[w]_{\sim_i} \subseteq [w]_{\leqslant_i}$. Since the relation $\leq_i$ intended to capture the a priori perspective

on a future situation, the set $[w]_{\leqslant_i}$ contains the worlds that agent i finds *a priori* indistinguishable from w. Conversely, since the relation $\sim_i$ is intended to capture the a posteriori perspective on a future situation, the set $[w]_{\sim_i}$ contains the worlds that i finds *a posteriori* indistinguishable from w. The fact that $[w]_{\sim_i} \subseteq [w]_{\leqslant_i}$ represents that agents can distinguish at least as much a posteriori (after having made observations) as they can a priori (before having made observations). In a doxastic state representing the situation after the coin has been tossed inside the dice cup, we would naturally have both $w \leqslant_r v$ and $w \sim_r v$: Rob does not a priori know the outcome, and he will also not know this a posteriori (as he will not observe the outcome). In a doxastic state representing the situation after having first tossed the coin and then lifted the cup to observe the outcome, we would naturally have $w \leqslant_r v$ and $w \not\sim_r v$: Rob can still not a priori predict the actual outcome, but a posteriori (after sensing), he will be able to distinguish *heads* from *tails*.

Let us briefly compare our formal setting with the one that it matches most closely, by Baltag and Smets [2006]. In that paper, $\leq_i$ is a single well-preorder, not a union of mutually disjoint well-preorders. This implies that in their setting, $\leqslant_i$ is the universal relation on W, and hence all worlds are a priori indistinguishable to all agents. Here we need a bit more. Say for instance that our coin tossing agent Rob actually *does* a priori know that the outcome of the coin toss will be heads, e.g. if Rob knows it to be a trick coin. We would formalize this case as having $w \not\leqslant_r v$, but that is not possible in the setting of the aforementioned paper, and we can't just omit the world v altogether, as some other agents might still consider it possible.

In other work by Baltag and Smets [9], they actually do allow $\leq_i$ to be a union of mutually disjoint well-preorders rather than a single well-preorder. However, in these settings, $\leq_i$ is formalizing what [3] refers to as a "local plausibility relation", which in our setting would correspond to an a posteriori rather than an a priori plausibility relation. Technically speaking, the consequence is that their $\sim_i$ relation becomes the symmetric closure of their $\leq_i$ relation, i.e., $\sim_i = \leqslant_i$, since then both $\leq_i$ and $\sim_i$ represent the a posteriori perspective. In our logic, $\leq_i$ represents the a priori perspective, and $\sim_i$ the a posteriori perspective, and hence we only require $\sim_i \subseteq \leqslant_i$.

Example 2.1. Let $Ag = \{r, h\}$, with r for Rob and h for Hanna (recall that Rob is supposed to be a robot and Hanna a human). Consider a possible future situation where Hanna has poured herself a cup of coffee and sees a sugar dispenser on the table. Let *sugar* denote that the dispenser contains sugar, and *salt* that it contains salt. Letting $L(w) = \{sugar\}$ and $L(v) = \{salt\}$, we can use $w <_h v$ to represent that she a priori considers it more plausible that the dispenser contains sugar than salt. Suppose that in reality the dispenser contains salt, Rob knows this, and he also knows about the false belief of Hanna. Rob's representation of this situation is given by the state s_1 of Figure 1, which also has $w \sim_h v$: the two worlds are a posteriori indistinguishable to Hanna, representing that she will not know whether *sugar* or *salt* is true. Hanna might later try to find out which one is true by tasting the content of the dispenser. Rob is aware of this, and might try to reason about the effect of Hanna tasting the content. Rob can represent the 'after tasting' situation by the state s_2 of Figure 1. It represents that Hanna comes to know whether it's sugar or salt (since $w \not\sim v$). Note that the model still contains the plausiblity ordering $w <_h v$ between the two worlds. At first,

Fig. 1 Two simple doxastic states, $s_1 = (\mathcal{M}_1, v)$ and $s_2 = (\mathcal{M}_2, v)$, both with $W = \{w, v\}$, $L(w) = \{sugar\}$ and $L(v) = \{salt\}$. General conventions for illustrations of plausibility models: Each node (circle) represents a world, labelled by its name followed by the list of true atoms at the world. The solid circles represent the designated worlds. When for two worlds w, v and some agent i we have $w <_i v$, we put a directed edge labelled i from v to w, except for reflexive loops that are never shown. When $w \simeq_i v$, the edge is bidirectional. The edge is solid if we also have $w \sim_i v$, otherwise it is dashed. The two states s_1 and s_2 only differ by having $w \sim_h v$ in s_1 and $w \not\sim_h v$ in s_2: in s_2, the two worlds have become distinguishable.

it might sound counter-intuitive to talk about considering *sugar* to be more plausible than *salt* and at the same time knowing which one it is. However, we are using these models to reason about possible future states, including what the agents know and believe about these future states, as well as what they will come to believe and know when reaching those states. This is similarly to what is done in epistemic planning based on plausibility models [10, 11]. It is also similar to the distinction between *plan time* and *execution time* knowledge and uncertainty considered by Bacchus and Petrick [12]. Their plan time uncertainty would be represented by our $\leqslant_i$ relation: It is the future outcomes that the agent a priori (at plan time) cannot distinguish. Their execution time uncertainty would be represented by our $\sim_i$ relation: It is the future outcomes that are still going to be indistinguishable at execution time (a posteriori), i.e., after having received the additional information that comes from observing the outcomes of the executed actions. In this specific example, s_2 represents Rob reasoning about the outcome of Hanna tasting the content: she will come to know whether it is salt or sugar (since $w \not\sim_h v$), but a priori she considers it most plausible that it turns out to be sugar (since $w <_h v$).

Consider 'standard' plausibility models as defined by Baltag and Smets [9]. These are models $\mathcal{M}^{std} = (W, \sim_i^{std}, \leq_i^{std}, L)_{i \in Ag}$ where the $\sim_i^{std}$ are equivalence relations, where $\sim_i^{std} = \leqslant_i^{std}$, and where the restriction of $\leq_i^{std}$ to each $\sim_i^{std}$-equivalence class is a well-preorder [9]. These standard plausibility models are clearly also plausibility models according to our Definition 2.1, but the opposite doesn't hold in general, as our models don't require $\sim_i^{std} = \leqslant_i^{std}$. However, as soon as one of our models $\mathcal{M} = (W, \sim_i, \leq_i)_{i \in Ag}$ satisfy $\sim_i = \leqslant_i$ for all i, it's a standard plausibility model. Thus we can formally define:

Definition 2.2. *A plausibility model $\mathcal{M} = (W, \sim_i, \leq_i)$ is called* standard *if $\sim_i = \leqslant_i$ for all agents $i \in Ag$. Similarly, a doxastic state $s = (\mathcal{M}, W_d)$ is called* standard *if $\mathcal{M}$ is standard.*

As mentioned, not every plausibility model of our type is standard. However, we have that any given plausibility model $\mathcal{M} = (W, \sim_i, \leq_i, L)$ of our type (i.e., a model satisfying Definition 2.1) gives rise to two distinct standard plausibility models:

1. $\mathcal{M}^{apri} = (W, \leqslant_i, \leq_i, L)_{i \in Ag}$
2. $\mathcal{M}^{apos} = (W, \sim_i, \leq_i \cap \sim_i, L)_{i \in Ag}$

It is easy to check that these are indeed standard plausibility models: 1) in both models, the epistemic relation is the symmetric closure of the plausibility relation (for the model $\mathcal{M}^{apos}$, this follows from the fact that any plausibility model of our type satisfies $\sim_i \subseteq \leqslant_i$); 2) in both models, the plausibility relation is a well-preorder on each indistinguishability equivalence class (for the model $\mathcal{M}^{apos}$, this follows from well-preorders being preserved on subsets). As the names suggest, these two models capture the a priori and a posteriori perspective, respectively. This also clearly illustrates how to move from the a priori to the a posteriori perspective: The a posteriori perspective is achieved by taking the intersection of an agent's a priori relations with its (a posteriori) indistinguishability relation. The intuition is that to move from what we now know about a future situation to what we will know in that future situation is to integrate the consequences of the observations we are going to make, encoded by $\sim_i$. Consider $\mathcal{M}_2$ of Figure 1 representing the future situation where Hanna has tasted the content of the dispenser. In the standard plausibility model $\mathcal{M}_2^{apri}$, w and v are indistinguishable to h, and w is strictly more plausible than v. This represents Hanna's a priori perspective, where she doesn't know whether she will taste sugar or salt, but she expects sugar. In the standard plausibility model $\mathcal{M}_2^{apos}$, w and v have become distinguishable and hence not related by the plausibility relation either (since it's a standard plausibility model). This represents Hanna's a posteriori perspective, where she will know whether it's salt or sugar (depending on whether w or v becomes the actual world).

Note that the special case where $\mathcal{M}^{apri} = \mathcal{M}^{apos}$ occurs exactly when $\leqslant_i = \sim_i$, i.e., if and only if $\mathcal{M}$ is standard. Thus, a plausibility model of our type is a standard plausibility model iff the induced a priori and a posteriori models are the same, i.e., iff the model doesn't distinguish the a priori and a posteriori perspectives. Standard plausibility models can hence also naturally be used to represent what is currently the case, since they don't have the a priori/a posteriori distinction. The model $\mathcal{M}_1$ of Figure 1 is standard, so we could also use it to represent what *currently* holds in some situation: the situation where Hanna has poured herself a cup of coffee and sees the sugar dispenser on the table. However, $\mathcal{M}_2$ is not standard, it represents the future situation after Hanna has tasted the content of the dispenser.

The 'a priori' vs 'a posteriori' distinction suggests that there is a temporal perspective involved. The a posteriori relations are supposed to capture what will be believed and known when reaching a particular future state. So the time point we are referring to is defined by the time point of that future state, e.g. after tossing a coin or after tasting the content of a dispenser. However, what is then the 'now' that we refer to when we talk about the a priori relations? It is some state at a given initial time point. Important is only that we build the model so that the information gained by the agent between the initial time point and the modeled future time point is encoded by $\sim_i$.

Note.

We introduce our framework here suggesting a temporal perspective. We use *a priori* to refer to the agent's beliefs about a future state, and *a posteriori* to refer to the agent's beliefs when it is in that "future" state. Note, that the temporal interpretation is not the only one that can be used for our framework, where we do not have any explicit modeling of time. We only model that what is known and believed when being

in a state (a posteriori) and what is known and believed about the state when not being in the state. This allows interpretations such as *imagined* or *virtual* states, or also *conditional* states, or indeed *counterfactual* states. However, here for simplicity and easier understanding, we will only use the temporal interpretation of our framework.

2.2 Logical language

The language $\mathcal{L}$ used to reason about plausibility models is:

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi \wedge \phi \mid K_i\phi \mid B_i\phi \mid E_i\phi \mid [a]\phi \qquad (\mathcal{L})$$

where $p \in P$, $i \in Ag$ and a is a doxastic action (defined later). We read $K_i\phi$ as "agent i knows ϕ", $B_i\phi$ as "agent i believes ϕ", $E_i\phi$ as "agent i expects ϕ" and $[a]\phi$ as "after a happens, ϕ". Since we are using doxastic states to represent possible future states, we also sometimes read $K_i\phi$ as "agent i will know ϕ" and $B_i\phi$ as "agent i will believe ϕ". As we will later see, we could e.g. have $s \models [toss](K_h heads \wedge E_h tails)$: After the coin has been tossed in state s, Hanna will know heads, but expects tails (so she has a *false expectation* in this case, a concept that we will introduce more formally in Section 4). The dual operator $\langle a\rangle\phi$ is defined by abbreviation: $\langle a\rangle\phi := \neg[a]\neg\phi$. Similarly, we define $\bot := \neg\top$. The *propositional language*, $\mathcal{L}_{prop}$, is the language induced by the first 4 of the above clauses. The semantics for the propositional connectives is standard (where $\top$ denotes a tautology, hence, true in any world of any model). The semantics for multi-pointed states is defined in terms of single-pointed states as follows, where $(\mathcal{M}, W_d)$ is a state and ϕ a formula of $\mathcal{L}$:

$$(\mathcal{M}, W_d) \models \phi \quad \text{iff} \quad (\mathcal{M}, w) \models \phi \text{ for all } w \in W_d$$

Note that if $W_d = \emptyset$, then $(\mathcal{M}, W_d) \models \bot$ (a model without designated worlds makes any formula true, including $\bot$). We now provide the semantics for the knowledge and belief modalities; the expectation and action modalities will only be introduced later. The semantics are as follows, where $(\mathcal{M}, w)$ is a (single-pointed) state and ϕ a formula of $\mathcal{L}$:

$$(\mathcal{M}, w) \models K_i\phi \quad \text{iff} \quad (\mathcal{M}, v) \models \phi \text{ for all } v \in [w]_{\sim_i}$$
$$(\mathcal{M}, w) \models B_i\phi \quad \text{iff} \quad (\mathcal{M}, v) \models \phi \text{ for all } v \in \mathrm{Min}_{\leq_i}[w]_{\sim_i}$$

According to these semantics, what is known is what is true in all worlds a posteriori indistinguishable from the designated world (by the relevant agent), and what is believed is what is true in the most plausible of these (by the relevant agent). In s_1 of Figure 1, Hanna does not know whether the dispenser contains sugar or salt, but she believes it's sugar: $s_1 \models \neg K_h sugar \wedge \neg K_h salt \wedge B_h sugar$.[2] In s_2, she will know that it is salt: $s_2 \models K_h salt$. So the move from s_1 to s_2 corresponds to Hanna tasting the content and getting her false belief corrected. In our setting, belief is meant to capture "firm belief", something that the agent is willing to act and depend upon. In the salt and sugar example, the meaning of Hanna believing the dispenser to contain sugar is that she wouldn't hesitate to pour some of the content into her coffee cup and

[2] As s_1 is a standard plausibility model, it could equally well be referring to the present state as an imagined future state, cf. the discussion in Section 2.1.

take a sip, even if strongly disliking salty coffee (as most people do!). If we wished to express that she is feeling uncertain about whether *sugar* is true, then despite her maybe believing 'somewhat more' in *sugar* than *salt*, we would still model this as equiplausibility ($w \simeq_h v$).

Note that both the K_i and B_i operators take the a posteriori perspective, since they limit the points of evaluation to what is a posteriori indistinguishable from the world of evaluation w (note that we have $v \in [w]_{\sim_i}$ and $v \in \mathrm{Min}_{\leq_i}[w]_{\sim_i}$ in the semantics of K_i and B_i, respectively).

2.3 Event models and product update

To represent actions or other events happening in the environment (e.g. tasting the content of the dispenser), we are going to use *event models*, and the result of applying an action in a plausibility model is defined via the *product update* operator [3], here using the version including postconditions to also be able to model ontic/physical change [10]. As usual, we define a *literal* as either an atom (element of P) or its negation. We often identify conjunctions of literals with the set of literals they contain, so for literals $l_1, \ldots, l_n$, we identify $l_1 \wedge \cdots \wedge l_n$ with $\{l_1, \ldots, l_n\}$. We also allow the empty conjunction of literals, which is identified with $\top$.

Definition 2.3. *A* (plausibility) event model *is* $\mathcal{E} = (E, \sim_i, \leq_i, pre, post)_{i \in Ag}$*, where*

- E *is a finite set of* events.
- $\leq_i$ *is a* plausibility relation *satisfying the same conditions as for plausibility models (Definition 2.1).*
- $\sim_i$ *is an* indistinguishability relation *satisfying the same conditions as for plausibility models (Definition 2.1).*
- $pre : E \to \mathcal{L}_{prop}$ *assigns to each event* e *a* precondition $pre(e)$*, being a formula of the propositional language.*[3]
- $post : E \to \mathcal{L}_{prop}$ *assigns to each event* e *a* postcondition $post(e)$ *being a propositionally satisfiable conjunction of literals.*

A (doxastic) action *is a pair* $(\mathcal{E}, E_d)$*, where* $\mathcal{E} = (E, \sim_i, \leq_i, pre, post)_{i \in Ag}$ *is an event model and* $E_d \subseteq E$ *is a set of* designated events*. A doxastic action is called* single-pointed *if* $|E_d| = 1$*, and* multi-pointed *if* $|E_d| > 1$*. Single-pointed actions* $(\mathcal{E}, \{e\})$ *are often written* $(\mathcal{E}, e)$*, and we call* e *the* designated event*.*

Multi-pointed actions were also introduced by Baltag and Smets [9] (they called them 'doxastic programs'). By pointing out multiple events, we can for instance represent non-deterministic actions [13] or the 'appearance' of an action to a specific agent (defined later). Given a state s and an action a, the result of applying a in s is the product update $s \otimes a$ defined next.

Definition 2.4. *Given a plausibility model* $\mathcal{M} = (W, \sim_i, \leq_i, L)_{i \in Ag}$ *and an event model* $\mathcal{E} = (E, \sim_i, \leq_i, pre, post)_{i \in Ag}$*, we define the* product update *of* $\mathcal{M}$ *with* $\mathcal{E}$ *as* $\mathcal{M} \otimes \mathcal{E} = (W', \sim'_i, \leq'_i, L')_{i \in Ag}$ *given by*[4]

- $W' = \{(w, e) \in W \times E \mid (\mathcal{M}, w) \models pre(e)\}$
- $(w, e) \leq'_i (v, f)$ *iff* $(e <_i f$ *and* $w \lesseqgtr_i v)$ *or* $(e \simeq_i f$ *and* $w \leq_i v)$
- $(w, e) \sim'_i (v, f)$ *iff* $w \sim_i v$ *and* $e \sim_i f$

[3] Usually arbitrary modal formulas are allowed as precondition, but they are not needed here, so we omit them for simplicity, and only consider propositional preconditions.

- $p \in L'(w,e)$ *iff* $p \in post(e)$ *or* $\big((\mathcal{M},w) \models p$ *and* $\neg p \notin post(e)\big)$

Given a state $s = (\mathcal{M}, W_d)$ *and an action* $a = (\mathcal{E}, E_d)$*, we define the product update of* s *with* a *as* $s \otimes a = (\mathcal{M} \otimes \mathcal{E}, W' \cap (W_d \times E_d))$*. If for all* $w \in W_d$ *there exists* $e \in E_d$ *with* $(\mathcal{M}, w) \models pre(e)$*, we say that* a *is* applicable *(or* executable*) in* s*.*

The intention of the above definition is of course that $\mathcal{M} \otimes \mathcal{E}$ becomes a plausibility model and $s \otimes a$ becomes a state, i.e., that they satisfy the conditions of Definition 2.1. This was not proved, but only implicitly assumed, in the original papers on plausibility models [3, 9]. Even though the proof is not difficult, we find it instructive, and provide it here for completeness. Also note that since our plausibility models are defined slightly differently than in the aforementioned papers, we in any case ought to ensure that our product update preserves the necessary properties.

Proposition 2.1. *The product update* $\mathcal{M} \otimes \mathcal{E}$ *of a plausibility model* $\mathcal{M}$ *with an event model* $\mathcal{E}$ *is a plausibility model. The product update* $s \otimes a$ *of a single- or multi-pointed state* s *with an applicable action* a *is a single- or multi-pointed state.*

Proof. The second claim of the theorem follows immediately from the first using the definition of applicability, so we only prove the first. We need to show that $\mathcal{M} \otimes \mathcal{E} = (W', \sim_i', \leq_i', L')_{i \in Ag}$ as defined in Definition 2.4 is a plausibility model. Consider first the relation $\sim_i'$ (for any $i \in Ag$). We need to show that it is an equivalence relation on W' and that $(w,e) \sim_i' (v,f)$ implies $(w,e) \lesssim_i' (v,f)$. Since the $\sim_i$ relations on both W and E are equivalence relations, the definition of $\sim_i'$ immediately implies that it is an equivalence relation. To prove that $(w,e) \sim_i' (v,f)$ implies $(w,e) \lesssim_i' (v,f)$, suppose $(w,e) \sim_i' (v,f)$. Then $w \sim_i v$ and $e \sim_i f$. Thus $w \lesssim_i v$ and $e \lesssim_i f$. From $e \lesssim_i f$, we get that either $e <_i f$, $f <_i e$ or $e \simeq_i f$. If $e <_i f$, then since $w \lesssim_i v$, we get $(w,e) \leq_i' (v,f)$, by definition of $\leq_i'$. This implies $(w,e) \lesssim_i' (v,f)$. The case of $f <_i e$ is symmetric. If $e \simeq_i f$, then since $w \lesssim_i v$, we either have $w \leq_i v$ or $v \leq_i w$. Thus either $(w,e) \leq_i' (v,f)$ or $(v,f) \leq_i' (w,e)$, by definition of $\leq_i'$. Again this implies $(w,e) \lesssim_i' (v,f)$, as required. It now only remains to show that $\leq_i'$ is reflexive and a union of mutually disjoint well-preorders on W'. Reflexivity of $\leq_i'$ is inherited from the reflexivity of the $\leq_i$ relations on $\mathcal{M}$ and $\mathcal{E}$: since $w \leq_i w$ and $e \simeq_i e$, by definition of $\leq_i'$ we get $(w,e) \leq_i' (w,e)$. To show that $\leq_i'$ is a union of mutually disjoint well-preorders, we first we show that it is a preorder (reflexive and transitive). We already know that it is reflexive, so we turn to transitivity. Suppose $(w,e) \leq_i' (v,f)$ and $(v,f) \leq_i' (u,g)$. We then have four cases to consider:

- $e <_i f$, $w \lesssim_i v$, $f <_i g$, $v \lesssim_i u$. By transitivity of $<_i$, and since $\lesssim_i$ is an equivalence relation, we immediately get $e <_i g$ and $w \lesssim_i u$, hence $(w,e) \leq_i' (u,g)$.
- $e <_i f$, $w \lesssim_i v$, $f \simeq_i g$, $v \leq_i u$. From $e <_i f$ and $f \simeq_i g$, we get $e <_i g$. From $w \lesssim_i v$ and $v \leq_i u$, we get $w \lesssim_i u$. Hence $(w,e) \leq_i' (u,g)$.
- $e \simeq_i f$, $w \leq_i v$, $f <_i g$, $v \lesssim_i u$. Symmetric to the previous case.
- $e \simeq_i f$, $w \leq_i v$, $f \simeq_i g$, $v \leq_i u$. This gives $e \simeq_i g$ and $w \leq_i u$, hence $(w,e) \leq_i' (u,g)$.

We now turn to proving that $\leq_i'$ is a union of mutually disjoint well-preorders. This amounts to showing that $\leq_i'$ is a well-preorder on each $\lesssim_i'$ equivalence class, i.e., that any subset X of any $\lesssim_i'$ equivalence class has $\leq_i'$-least elements. Let an arbitrary such set X be given. For any subset $X' \subseteq X$, define sets $\pi_1(X') = \{w \in W \mid (w,e) \in$

[4] Note that we are using the same symbols for the indistinguishability and plausibility relations in the two models, but it will always be clear from the context which one we are talking about.

X' for some $e \in E\}$ and $\pi_2(X') = \{e \in E \mid (w,e) \in X' \text{ for some } w \in W\}$. We now prove two claims to be used in the finding least elements in X. *Claim 1*: If $((w,e),(v,f)) \in \leqslant_i'$ then $w \leqslant_i v$ and $e \leqslant_i f$. *Proof of Claim 1*: Since the $\leqslant_i$ relations are symmetric, it suffices to prove that $((w,e),(v,f)) \in \leq_i'$ implies $w \leqslant_i v$ and $e \leqslant_i f$. If $((w,e),(v,f)) \in \leq_i'$, then either both $e <_i f$ and $w \leqslant_i v$ or else both $e \simeq_i f$ and $w \leq_i v$. In both cases, we immediately get $w \leqslant_i v$ and $e \leqslant_i f$, completing the proof of the claim. *Claim 2*: For any $X' \subseteq X$, $\pi_1(X')$ and $\pi_2(X')$ are well-preordered by $\leq_i$. *Proof of Claim 2*: We only prove it for $\pi_1(X')$, the proof for $\pi_2(X')$ being symmetric. Since $\leq_i$ is a union of disjoint well-preorders, it suffices to prove that all elements of $\pi_1(X')$ are related by $\leqslant_i$. So let $w, v \in \pi_1(X')$. Then by definition, we have $(w,e),(v,f) \in X'$ for some $e, f \in E$. Since X' is a subset of a $\leqslant_i'$ equivalence class, Claim 1 gives us the required conclusion, completing the proof of Claim 2. Now let $X_1 = \{(w,e) \in X \mid e \in \mathrm{Min}_{\leq_i} \pi_2(X)\}$ and $X_2 = \{(w,e) \in X_1 \mid w \in \mathrm{Min}_{\leq_i} \pi_1(X_1)\}$. Claim 2 gives us that these sets are both non-empty. We will now show that the elements of X_2 are least elements of X, which gives the required. So let $(w,e) \in X_2$ and $(v,f) \in X$. We need to show that $(w,e) \leq_i' (v,f)$. First note that since $(w,e),(v,f) \in X$ and X is a subset of a $\leqslant_i'$ equivalence class, Claim 1 gives us that $w \leqslant_i v$ and $e \leqslant_i f$. Hence, if $e <_i f$, then we would immediately get $(w,e) \leq_i' (v,f)$ from the definition of $\leq_i'$. Thus it only remains to consider the case $e \not<_i f$. Since $e \in \mathrm{Min}_{\leq_i} \pi_2(X)$, e is $\leq_i$-least among the events occuring in X, and thus $f \not<_i e$. Since we now have $e \leqslant_i f$, $e \not<_i f$ and $f \not<_i e$, we get $e \simeq_i f$. To prove $(w,e) \leq_i' (v,f)$, it then suffices to show that $w \leq_i v$. Since $e \simeq_i f$ and $e \in \mathrm{Min}_{\leq_i} \pi_2(X)$, we also get $f \in \mathrm{Min}_{\leq_i} \pi_2(X)$, and hence $(v,f) \in X_1$. Since $(w,e) \in X_2$, we have $w \in \mathrm{Min}_{\leq_i} \pi_1(X_1)$, i.e., w is $\leq_i$-least among the worlds occuring in X_1. As $(v,f) \in X_1$, we can conclude $w \leq_i v$. □

The definition of $\leq_i'$ above gives us the so-called *action-priority update* [9]: If the plausibility ordering on worlds and events don't agree, the ordering on events take precedence, with the intuition that beliefs are updated in face of the new information represented by the events. We can now extend the definition of the semantics with the clause for the *dynamic modality* $[a]\phi$, where a is an action and $\phi \in \mathcal{L}$:

$$(\mathcal{M}, w) \models [a]\phi \quad \text{iff} \quad (\mathcal{M}, w) \otimes a \models \phi$$

This definition is slightly non-standard. Usually in DEL, the semantic condition for $(\mathcal{M}, w) \models [(\mathcal{E}, e)]\phi$ to hold is that "$(\mathcal{M}, w) \not\models pre(e)$ or $(\mathcal{M} \otimes \mathcal{E}, (w,e)) \models \phi$" [2]. In our setting, if $(\mathcal{M}, w) \not\models pre(e)$, then $(\mathcal{M}, w) \otimes (\mathcal{E}, e)$ will have an empty set of designated worlds, and hence by our definitions, $(\mathcal{M}, w) \otimes (\mathcal{E}, e) \models \phi$ will trivially hold. This implies that for single-pointed states and actions, the standard semantics and ours coincide, despite the formulation of our semantic condition being simpler. The simplification is made possible by allowing states to have arbitrary sets of designated worlds, even the empty set. The expected properties of our semantic condition are more formally confirmed by the following lemma.

Lemma 2.1. *The following holds for all states s, actions a and formulas ϕ:*

1. $s \models [a]\phi$ *iff* $s \otimes a \models \phi$.
2. $s \models \langle a \rangle \top$ *iff a is applicable in s.*
3. $s \models [(\mathcal{E}, E_d)]\phi$ *iff for all* $e \in E_d$, $s \models [(\mathcal{E}, e)]\phi$.

Proof. Suppose $s = (\mathcal{M}, W_d)$, $a = (\mathcal{E}, E_d)$ and let W' denote the worlds of $\mathcal{M} \otimes \mathcal{E}$. *Item 1*: Using the definitions, we have $s \models [a]\phi$ *iff* $(\mathcal{M}, w) \models [a]\phi$ for all $w \in W_d$ *iff* $(\mathcal{M}, w) \otimes a \models \phi$ for all $w \in W_d$ *iff* $(\mathcal{M} \otimes \mathcal{E}, W' \cap (\{w\} \times E_d)) \models \phi$ for all $w \in W_d$ *iff* $(\mathcal{M} \otimes \mathcal{E}, W' \cap (W_d \times E_d)) \models \phi$ *iff* $s \otimes a \models \phi$. *Item 2*: We have $s \models \langle a \rangle \top$ *iff* $(\mathcal{M}, w) \models \langle a \rangle \top$ for all $w \in W_d$ *iff* $(\mathcal{M}, w) \not\models [a]\bot$ for all $w \in W_d$ *iff* $(\mathcal{M}, w) \otimes a \not\models \bot$ for all $w \in W_d$ *iff* for all $w \in W_d$, the set of designated worlds of $(\mathcal{M}, w) \otimes a$ is non-empty *iff* for all $w \in W_d$, there exists $e \in E_d$ such that $(w, e) \in W'$ *iff* for all $w \in W_d$, there exists $e \in E_d$ such that $(\mathcal{M}, w) \models pre(e)$ *iff* a is applicable in s. *Item 3*: As in the proof of item 1, we have $s \models [(\mathcal{E}, E_d)]\phi$ *iff* $(\mathcal{M} \otimes \mathcal{E}, W' \cap (W_d \times E_d))$. We further get $(\mathcal{M} \otimes \mathcal{E}, W' \cap (W_d \times E_d))$ *iff* $(\mathcal{M} \otimes \mathcal{E}, W' \cap (W_d \times \{e\})) \models \phi$ for all $e \in E_d$ *iff* $s \otimes [(\mathcal{E}, e)] \models \phi$ for all $e \in E_d$ *iff* $s \models [(\mathcal{E}, e)]\phi$ for all $e \in E_d$. □

As in Baltag and Smets [9], we can define sequential compositions of actions. Since Baltag and Smets didn't include postconditions, for the postconditions we rely on the sequential composition defined by Ditmarsch and Kooi [14], although we get simplified expressions, since both our pre- and post-conditions are propositional. Below we use the standard notation $\phi[\gamma/\psi]$ for replacing all occurrences of ψ in ϕ with γ.

Definition 2.5. *Given event models* $\mathcal{E} = (E, \sim_i, \leq_i, pre, post)_{i \in Ag}$ *and* $\mathcal{E}' = (E', \sim'_i, \leq'_i, pre', post')_{i \in Ag}$, *we define their* sequential composition *as the event model* $\mathcal{E}; \mathcal{E}' = (E'', \sim''_i, \leq''_i, pre'', post'')_{i \in Ag}$ *given by*

- $E'' = \{(e, e') \in E \times E' \mid post(e) \wedge pre'(e')$ *is a propositionally satisfiable formula*$\}$
- $(e, e') \leq''_i (f, f')$ *iff* $(e' <'_i f'$ *and* $e \leqslant_i f)$ *or* $(e' \simeq'_i f'$ *and* $e \leq_i f)$
- $(e, e') \sim''_i (f, f')$ *iff* $e \sim_i f$ *and* $e' \sim'_i f'$
- $pre''(e, e') = pre'(e')[\top/p_1] \cdots [\top/p_n][\bot/q_1] \cdots [\bot/q_m] \wedge pre(e)$ *where* $post(e) = p_1 \wedge \cdots \wedge p_n \wedge \neg q_1 \wedge \cdots \wedge \neg q_m$
- $post''(e, e') = post'(e') \cup \{p, \neg p \in post(e) \mid p \notin post'(e')\}$

For actions $a = (\mathcal{E}, E_d)$ *and* $a' = (\mathcal{E}', E'_d)$, *we define their sequential composition as* $a; a' = (\mathcal{E}; \mathcal{E}', E'' \cap (E_d \times E'_d))$.

In [14], the precondition map is defined by $pre''(e, e') = pre(e) \wedge [(\mathcal{E}, e)]pre'(e')$ which would in principle also work here, except then the precondition is no longer a propositional formula. It is still equivalent to one, though, as we will see in the later axiomatization, Theorem 6.1, below.

Proposition 2.2. *For all states* s *and actions* a, a', *the states* $s \otimes a \otimes a'$ *and* $s \otimes (a; a')$ *are isomorphic. For all formulas* ϕ, *we then have* $s \models [a; a']\phi$ *iff* $s \models [a][a']\phi$.

Proof. By Lemma 2.1(1), the second statement follows directly from the first, so we only prove the first. As this result corresponds directly to Proposition 3.7 of [9], we here only prove the crucial step of verifying that the plausibility relation $\leq''_i$ is correctly defined. This amounts to proving that for all worlds w, v of s, all events e, e' of a and all events f, f' of a', we have $(w, (e, e')) \leq''_i (v, (f, f'))$ iff $((w, e), e') \leq''_i ((v, f), f')$. This is proved directly based on Definitions 2.4 and 2.5:

$((w, e), e') \leq_i ((v, f), f') \Leftrightarrow (e' <_i f', (w, e) \leqslant_i (v, f))$ or $(e' \simeq_i f', (w, e) \leq_i (v, f)) \Leftrightarrow$
$(e' <_i f', w \leqslant_i v, e \leqslant_i f)$ or $(e' \simeq_i f', e <_i f, w \leqslant_i v)$ or $(e' \simeq_i f', e \simeq_i f, w \leq_i v) \Leftrightarrow$
$(((e' <_i f', e \leqslant_i f)$ or $(e' \simeq_i f', e <_i f))$ and $w \leqslant_i v)$ or $(e' \simeq_i f', e \simeq_i f, w \leq_i v) \Leftrightarrow$
$((e, e') <_i (f, f')$ and $w \leqslant_i v)$ or $((e, e') \simeq_i (f, f')$ and $w \leq_i v) \Leftrightarrow$
$(w, (e, e')) \leq_i (v, (f, f'))$. □

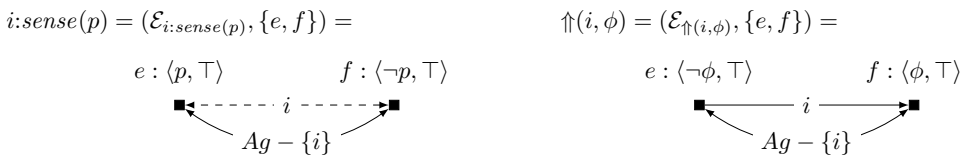

Fig. 2 *Left*: For each $i \in Ag$ and $p \in P$, we define a *sensing action* $i{:}sense(p)$ representing that agent i senses the truth value of p. General conventions for illustration of event models: Each node (square) represents an event e, labeled by its name, pre- and post-conditions, $\langle pre(e), post(e)\rangle$. Edge conventions are the same as for for plausibility models. In $i{:}sense(p)$, there are two events e and f, and we have $pre(e) = p$, $pre(f) = \neg p$, $post(e) = post(f) = \top$. The two events are distinguishable, but equiplausible to agent i. The two events are indistinguishable to all other agents. This means that the action represents the situation in which agent i becomes able to distinguish p-worlds from $\neg p$-worlds (since $e \not\sim_i f$), i.e., learns the truth-value of p, but the other agents don't (since $e \sim_j f$ for $j \neq i$). Both squares are solid, indicating that both events are designated.
Right: For each $i \in Ag$ and $\phi \in \mathcal{L}_{prop}$, a *belief update* $\Uparrow(i, \phi)$ making all the ϕ-worlds strictly more plausible than the $\neg\phi$-worlds to agent i, and not affecting any other agents. This belief update action is standard in the literature on plausibility models, used to model *announcements of soft facts* [3, 9]. Here we will also refer to $\Uparrow(i, \phi)$ as a *(soft) announcement* of ϕ to agent i.

Given Lemma 2.1(2), we can extend the notion of applicability to action sequences: An action sequence $a_1, \ldots, a_n$ is said to be *applicable* in a state s if the formula $s \models \langle a_1\rangle\top \wedge [a_1](\langle a_2\rangle\top \wedge [a_2](\cdots \langle a_n\rangle\top \cdots))$ holds. This is the standard definition of applicability in epistemic planning based on DEL [15–17]. Note that applicability of an action sequence $a_1, \ldots, a_n$ in a state s does *not* reduce to requiring $s \models \langle a_1; \cdots; a_n\rangle\top$, since we need to ensure that every a_i, $i \leq n$, is applicable in the result of *any* execution of $a_1, \ldots, a_{i-1}$ in s [17].

Example 2.2. Consider the sensing action $h{:}sense(sugar)$, cf. Figure 2, left (where we let $i = h$ and $p = sugar$). It represents the action of Hanna sensing whether *sugar* is true, i.e., it corresponds to tasting the content of the dispenser. Note that s_2 of Figure 1 is isomorphic to $s_1 \otimes h{:}sense(sugar)$: sensing cuts the indistinguishability link. This fits with what we explained previously: The model s_2 of Figure 1 represents how the model s_1 of Figure 1 is considered to evolve as the consequence of Hanna sensing (tasting) the content. Consider now instead the belief update action $\Uparrow(h, salt)$ defined by Figure 2, right (with $i = h$ and $\phi = salt$). It could be used to represent Rob telling Hanna that "the dispenser contains salt". Note that the action makes the announced formula most plausible to Hanna, i.e., she considers Rob trustworthy (meaning that she believes him to tell the truth). We have that $s_1 \otimes \Uparrow(h, salt)$ is isomorphic to s_1 with the direction of edge reversed, and hence $s_1 \otimes \Uparrow(h, salt) \models B_h salt$. Thus we have $s_1 \models B_h sugar \wedge [\Uparrow(h, salt)]B_h salt$: In s_1, she believes it is sugar, and after Rob tells her it is salt, she believes it is salt.

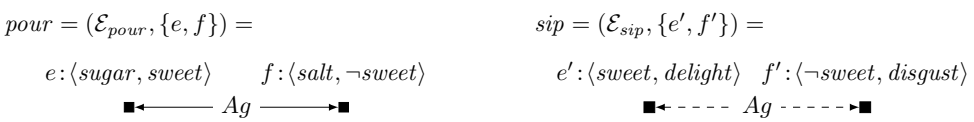

Fig. 3 The action *pour* represents the action of pouring content from the dispenser into the coffee cup and *sip* represents the action of Hanna taking a sip from the cup. Note that the edge is solid for *pour* and dashed for *sip*: pouring doesn't reveal the content, but taking a sip does. Both actions are public, i.e., all agents get to see the same outcome. In the case of *sip*, the intuition is that all agents present can see the facial expression of Hanna after taking a sip.

3 Modeling expectations

As mentioned in the introduction, we are interested in human-robot interaction scenarios where the robot (Rob) can reason about the false expectations of the human (Hanna), in order to 'warn her' in case those false expectations might lead to undesirable outcomes. To be able to do this, it turns out we need to extend our language, which we will carefully argue for in the following.

3.1 Beliefs about actions

Example 3.1. We add actions *pour* for pouring content from the dispenser into the cup of coffee, and *sip* for Hanna taking a sip from the cup, see Figure 3. We use the atom *sweet* to denote that the coffee is sweet (not salty), and *delight* and *disgust* to denote whether Hanna is delighted or disgusted. Consider the action composition *pour*; *sip* corresponding to first pouring and then taking a sip. Since $post(e) \wedge pre(f') = sweet \wedge \neg sweet$ and $post(f) \wedge pre(e') = \neg sweet \wedge sweet$ are not propositionally satisfiable formulas, we get that *pour*; *sip* only contains two events, (e, e') and (f, f'), cf. Definition 2.5. We will abbreviate these events as ee' and ff', respectively. More generally, *from now on, we will often abbreviate tuples* $(\cdots((w, e_1), e_2), \ldots, e_n)$ *by* $we_1 \cdots e_n$. Using again Definition 2.5, we see that $pre(ee') = \top \wedge sugar$, $pre(ff') = \top \wedge salt$, $post(ee') = delight \wedge sweet$, and $post(ff') = disgust \wedge \neg sweet$. We also get that $ee' \simeq_i ff'$ and $ee' \not\sim_i ff'$. In other words, *pour*; *sip* represents the combined action of pouring and sipping, where the outcome will either be that the coffee is sweet and Hanna is delighted or the coffee is not sweet and Hanna is disgusted; and where the two outcomes are a posteriori, but not a priori, distinguishable.

Consider again the situation modeled by s_1 of Figure 1. Hanna might consider to first pour content from the dispenser into her cup and then take a sip, as she believes this will lead to delight: $s_1 \models B_h[pour; sip]delight$. However, Rob knows that Hanna's beliefs are false, and that instead her planned action sequence will lead to disgust: $s_1 \models K_r[pour; sip]disgust$. So it would make sense for Rob to initially inform her about the salt in the dispenser, in order to avoid her getting disgusted. As in Example 2.2, we can use the belief update action $\Uparrow(h, salt)$ to represent that Rob informs Hanna about the content. We get $s_1 \models [\Uparrow(h, salt)]B_h[pour; sip]disgust$: If Rob first informs her about the salt, Hanna will (correctly) believe that pouring and sipping will lead to

Fig. 4 *Left:* The action *toss* for Hanna tossing a coin biased towards heads. The outcome where *heads* becomes true (represented by the event e) is considered strictly more plausible than the outcome where *tails* becomes true (represented by the event f), but the two outcomes are epistemically indistinguishable (as the coin toss takes place inside the dice cup). *Right:* The product update $s_0 \otimes toss$ where s_0 is the state containing a single world w with $L(w) = \emptyset$.

disgust. In that case, she can be expected to come up with an alternative plan to avoid the disgust (the role of Rob is only to provide relevant information to Hanna that she can decide to act upon or not, it's not to control her actions or limit her autonomy).

From the example above it would seem that if we want to express that an agent i a priori believes that some action a will lead to ϕ, we can use the formula $B_i[a]\phi$. However, this is not always the case, as that formula doesn't take into account the beliefs that i has *about* the action a. Let us illustrate this with an example. Hanna has a coin that she believes is biased towards heads, so that if she tosses it inside a dice cup, she expects it to end up showing heads. We can model the coin toss by the action *toss* in Figure 4 left (we here ignore other agents and consider the situation where $Ag = \{h\}$). Suppose *toss* is executed in the state s_0 containing a single world w with $L(w) = \emptyset$, leading to the state $s_0 \otimes toss$ shown in Figure 4 right. Since $s_0 \otimes toss$ contains a designated world not satisfying *heads* (the world wf), we get $s_0 \otimes toss \not\models heads$, and hence $s_0 \not\models [toss]heads$. Since s_0 is a singleton state, for all formulas ϕ we have that $s_0 \models \phi$ iff $s_0 \models B_h\phi$. Thus we can conclude that $s_0 \not\models B_h[toss]heads$: It's not the case that Hanna believes that tossing the coin will result in heads. This clearly doesn't capture the fact that she actually *does* expect the coin will land heads.[5]

One way to analyse the problem is to notice that $[a]$ is a box modality expressing "all executions of a lead to ϕ" and $\langle a \rangle$ is a diamond modality expressing "some execution of a leads to ϕ". When we in natural language say "Hanna believes that tossing the coin will result in heads", we are neither saying that she believes *all* executions to result in heads (corresponding to $B_h[toss]heads$), nor that she believes *some* execution to result in heads (corresponding to $B_h\langle toss \rangle heads$). No, we say that the *most plausible* execution(s), from her perspective, will result in heads. This seems to require a new modality capturing "executions considered most plausible by agent i". Baltag and Smets [9] already considered this and wrote: "we need to be able to model the agent's 'dynamic beliefs', i.e. *their beliefs about the action itself*: the *appearance* of this action (while it is happening) to each of the agents". They then define the *appearance* of a single-pointed action $(\mathcal{E}, e)$ to an agent i as $(\mathcal{E}, e)_i = (\mathcal{E}, \mathrm{Min}_{\leq_i}[e]_{\sim_i})$. We can extend this to multi-pointed actions by letting $(\mathcal{E}, E_d)_i = (\mathcal{E}, \bigcup_{e \in E_d} \mathrm{Min}_{\leq_i}[e]_{\sim_i})$. The appearance of *toss* to Hanna is then $toss_h = (\mathcal{E}_{toss}, e)$: While she knows that either

[5] Note that the issue here is not tied to the fact that the coin toss is modeled as a non-deterministic action with two designated events (meaning that the modeling agent, say Rob, cannot predict the outcome). If only f had been designated in *toss* (meaning that the modeling agent knows that *tails* will happen), $s_0 \otimes toss$ would only have wf as designated world, and again we would then have $s_0 \otimes toss \not\models heads$ and hence $s_0 \not\models B_h[toss]heads$.

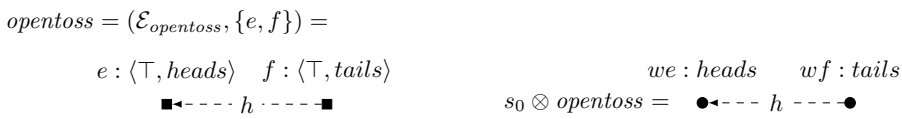

Fig. 5 The action *opentoss* for Hanna tossing the coin in the open (left), and the result of executing this action in s_0 (right).

e or f might happen, she expects it will be e (corresponding to *heads*). It represents her perspective on the *toss* action. We can now correctly express Hanna's belief that tossing will result in heads by the formula $B_h[toss_h]heads$, which indeed holds in s_0 (since the only designated world of $s_0 \otimes toss_h$ is *we* in which *heads* holds). In other words, when we want to express i's a priori belief that a will lead to ϕ, it doesn't suffice to use $B_i[a]\phi$, since it expresses that i believes that *all* executions of a will lead to ϕ. However, the formula $B_i[a_i]\phi$ appears to work, since it expresses that i believes that all *most plausible* executions of a (from her perspective) will lead to ϕ.

3.2 A priori perspectives and subjective applicability

There is still one issue with the discussion above, though: The appearance operator is used to model how an action appears to an agent "while it is happening", cf. the previous quote by Baltag and Smets [9]. In our framework, we are rather modeling expectations of *future* actions, *before* they happen. Consider for instance a new action *opentoss* that is as *toss* except the two events are distinguishable, see Figure 5. This corresponds to Hanna tossing the coin in the open, immediately observing the outcome. We would of course still want to say that Hanna a priori believes that tossing the coin will result in heads. However, note that by definition of appearance, $opentoss_h = toss$ (since $\mathrm{Min}_{\leq_h}[e]_{\sim_h} = \mathrm{Min}_{\leq_h}\{e\} = \{e\}$ and $\mathrm{Min}_{\leq_h}[f]_{\sim_h} = \mathrm{Min}_{\leq_h}\{f\} = \{f\}$), so we would still get $s_0 \not\models B_h[opentoss_h]heads$. The problem is that the appearance operator is defined in terms of the $\sim_i$ relation that captures a posteriori indistinguishability, so what is observed *when* the action happens. We are here interested in modeling a priori indistinguishability concerning a future action, which is instead captured by $\leqslant_i$. Using this relation instead, we now define an operator that gives an agent's 'a priori perspective' on an action.

Definition 3.1. *The* a priori perspective *of agent $i \in Ag$ on the action $(\mathcal{E}, E_d)$ is the action* $(\mathcal{E}, E_d)^i = (\mathcal{E}, \bigcup_{e \in E_d} \mathrm{Min}_{\leq_i}[e]_{\leqslant_i})$.[6]

With this new definition, we finally get $s_0 \models B_h[opentoss^h]heads$ as intended, since we have $opentoss^h = (\mathcal{E}_{opentoss}, e)$ (as $\mathrm{Min}_{\leq_h}[e]_{\leqslant_h} = \mathrm{Min}_{\leq_h}\{e, f\} = \{e\}$ and $\mathrm{Min}_{\leq_h}[f]_{\leqslant_h} = \mathrm{Min}_{\leq_h}\{e, f\} = \{e\}$). We would read it as saying that Hanna a priori believes (or simply 'Hanna expects') that tossing the coin will result in heads.

[6] Note that we are here putting the agent index as a superscript instead of the subscript used for the appearance operator. The superscript notation has also been used previously in DEL, to define *perspective shifts* on epistemic actions [18]. The operator defined here plays the same role in DEL based on plausibility models as those 'perspective shift' operators do in standard DEL (DEL with knowledge operators only).

Let us sum up the discussion so far. When in natural language we say "i believes that doing a will lead to ϕ", we are not only expressing a static belief about the state of affairs before a has taken place, but we are also expressing the agent's beliefs *about* the action a itself. This is why we can't formalize the natural language statement simply as $B_i[a]\phi$. Baltag and Smets [9] handled the issue by introducing their appearance operator, so that one could write $B_i[a_i]\phi$. However, their appearance operator captures how the action appears *while* happening, whereas we are interested in modeling expectations about future actions and their future outcomes. This is then instead captured by $B_i[a^i]\phi$. The differences are modeled via different sets of designated events. The difference between $[(\mathcal{E}, E_d)]$ and $[(\mathcal{E}, E_d)^i]$ is that in the former, it is assumed that the "actual" event is among the ones in E_d, whereas in the latter, it is assumed to be in $\bigcup_{e\in E_d} \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$: the events a priori considered most plausible by i, given that the actual event is among the ones in E_d.

Definition 3.2. *An action a is called* subjectively applicable *to agent i (i-applicable for short) in a state s if $s \models B_i\langle a^i\rangle\top$ holds. Else, it is called a* surprising action *(to agent i in s).*

Compare the definition above to item 2 of Lemma 2.1 which concluded that applicability of a in s can be expressed as $s \models \langle a\rangle\top$. The difference to subjective applicability is that we: 1) add a belief operator in front to express that applicability has to be verified from the subjective perspective of the agent in question, and 2) we replace the action itself by the agent's a priori perspective on it.

Example 3.2. An action can clearly be applicable without being subjectively applicable, e.g. Hanna might falsely believe the box to be locked, so she doesn't consider the action of opening it to be applicable (we can suppose that the *open_box* action has a single event $\langle \neg locked, open\rangle$). So it would be a surprising action for Hanna if she observed someone opening the box. Conversely, an action can also be subjectively applicable without being applicable, e.g. in the reverse situation where the box *is* actually locked, but Hanna falsely believes it is not. In that case, Hanna considers the action of opening the box to be applicable and may *try* to apply the action which fails because opening the box is not objectively applicable.

Example 3.3. Consider again the state s_1 of Figure 1 and the action $\Uparrow(h, salt)$ of Figure 2. We have that $s_1 \models B_h\langle\Uparrow(h, salt)^h\rangle\top$ iff $(\mathcal{M}_1, w) \models \langle\Uparrow(h, salt)^h\rangle\top$ iff $(\mathcal{M}_1, w) \models \langle(\mathcal{E}_{\Uparrow(h,salt)}, f)\rangle\top$. This does not hold, since f has precondition $salt$, and w satisfies only $sugar$. Announcing "it is salt" is objectively applicable, but not subjectively applicable for Hanna and hence a surprising action for her according to the above definition. The point is that Hanna initially, in s_1, believes that the content is sugar, so it would be a surprise to her if someone would announce that it is salt.

3.3 Introducing the expectation operator

The discussion above suggests that if we want to express that agent i believes the action sequence $a_1, \ldots, a_n$ to lead to ϕ, we can use the formula $B_i[a_1^i \cdots a_n^i]\phi$. As a priori beliefs are encoded by the plausibility relation, there might however be a more direct way to formalize expectations: by introducing a separate expectation modality based solely on the plausibility relation. We already introduced an expection modality E_i in our language, we just didn't provide a semantics for it yet. We first wanted

to motivate the need for such a modality in addition to the standard knowledge and belief modalities. The semantics for the expectation modality is:

$$(\mathcal{M}, w) \models E_i\phi \text{ iff } (\mathcal{M}, v) \models \phi \text{ for all } v \in \text{Min}_{\leq_i}[w]_{\leqslant_i}.$$

Note that this is exactly the same semantics as for the belief modality, except we evaluate in the most plausible worlds of $[w]_{\leqslant_i}$ (the worlds that are a priori indistinguishable from w) instead of $[w]_{\sim_i}$ (the worlds that are a posteriori indistinguishable from w). Note that the formula $E_i\phi \rightarrow B_i\phi$ is *not* valid. We can see this in Figure 1: it holds that $(\mathcal{M}_2, v) \models E_h sugar \wedge B_h \neg sugar$. We read this as follows: After Hanna tastes the content of the dispenser, she expects sugar to be true, but will actually end up believing that sugar is false. This view of modeling is similar to previous work on conditional beliefs. Baltag and Smets [3] interpret the conditional belief statement $B_a^P Q$ as follows: "... if the actual state is s, then after coming to believe that P is the case (at this *actual* state), agent a will believe that Q was the case (at the *same* actual state, before his change of belief). In other words, conditional beliefs B_a^P give descriptions of the agent's plan (or commitments) about what he will believe about the current state after receiving new (believable) information ...". Furthermore, van Benthem and Smets [19] state "... conditional beliefs "pre-encode" the beliefs that agents would have if they were to learn certain (new) things ...". We can make the correspondence to conditional beliefs a bit more formally precise. The semantics of the conditional belief operator in standard plausibility models $\mathcal{M} = (W, \leqslant_i, \leq_i)_{i \in Ag}$ is defined by: $(\mathcal{M}, w) \models B^\psi\phi$ iff $(\mathcal{M}, v) \models \phi$ for all $v \in \text{Min}_{\leq_i}([w]_{\leqslant_i} \cap \{u \in W \mid (\mathcal{M}, u) \models \psi\})$. The difference between the semantics of our expection operator and our belief operator is that $E_i\phi$ evaluates in all worlds of $\text{Min}_{\leq_i}[w]_{\leqslant_i}$ whereas $B_i\phi$ only evaluates in the worlds of $\text{Min}_{\leq_i}([w]_{\leqslant_i} \cap [w]_{\sim_i})$. Thus $B_i\phi$ is evaluated as $E_i\phi$, except it is conditioned on the actual world being in $[w]_{\sim_i}$—in exactly the same way as the semantics of the conditional belief formula $B_i^\psi\phi$ is conditioned on the actual world being a ψ-world. Hence, we can think of our belief operator B_i as a "conditional expectation operator": it's an expectation operator conditioned on receiving the additional information that the actual world is in $[w]_{\sim_i}$. So we could read $(\mathcal{M}_2, v) \models E_h sugar \wedge B_h \neg sugar$ as: Hanna expects sugar to be true, but conditional on the additional information that the actual world is in $[v]_{\sim_h} = \{v\}$, she expects sugar to be false.

As earlier mentioned, we are going to reason about future situations, situations that will arise after a sequence of actions has occurred. It is crucial that the expectation modality is applied in the state that represents the possible future situation. If we want to say that Hanna expects the coin toss to lead to heads, it is formalized as $[toss]E_h heads$, not as $E_h[toss]heads$ (the latter wouldn't work for the same reason as $B_h[toss]heads$ didn't work). So in general, we now propose to express agent i's expectation that the action sequence $a_1, \ldots, a_n$ will lead to ϕ as $[a_1 \cdots a_n]E_i\phi$ instead of $B_i[a_1^i \cdots a_n^i]\phi$. Note that the former formula is simpler in not requiring the a priori perspective operator on actions. This allows us to express beliefs, knowledge and expectations about the same action sequence in a more compact way, for instance:

1. $s_0 \models [toss](E_h heads \wedge \neg K_h heads \wedge \neg K_h \neg heads)$
   After *toss* happens in s_0, Hanna expects *heads*, but will not know whether *heads*.

2. $s_0 \models [opentoss](E_h heads \wedge (K_h heads \vee K_h \neg heads))$
   After *opentoss* happens in s_0, Hanna expects *heads*, and she will know whether *heads*.

Trying to express the same two formulas using the $B_i[a_1^i; \cdots ; a_n^i]$ notation, we would get the somewhat more convoluted and less immediately readable:

1. $s_0 \models B_h[toss^h]heads \wedge [toss](\neg K_h heads \wedge \neg K_h \neg heads)$
2. $s_0 \models B_h[opentoss^h]heads \wedge [opentoss](K_h heads \vee K_h \neg heads)$

There are several advantages to having introduced a separate expectation operator. First, when modeling h's a priori beliefs about future action outcomes, we need to consider the actions from her perspective, meaning that without the expectation operator, we would need to write an extra action sequence representing her a priori perspective on the actions (like $toss^h$ and $opentoss^h$ above). With the E-operator, on the other hand, we consider the *actual* action sequence, not the agent's perspective of it. Concerning notation, the E-operator provides a more compact notation than the B-operator—especially when considering several agents whose beliefs about a future state needs to be computed using individual perspectives of the action sequence for each of them. For instance, if Rob expects the coin toss to result in tails, using the expectation operator, we can express both agents' perspectives by:

$$s_0 \models [toss](E_h heads \wedge E_r tails \wedge \neg K_h heads \wedge \neg K_h \neg heads)$$

Expressing the same without the expectation operator would become:

$$s_0 \models B_h[toss^h]heads \wedge B_r[toss^r]tails \wedge [toss](\neg K_h heads \wedge \neg K_h \neg heads)$$

But more importantly, we can model something different using the two operators and action sequences—the belief about the future state after an action sequence from the agent's perspective, vs, the expectation of a state that comes after application of a "neutral" action sequence. A case which can illustrate this difference is *surprise*, or simply the fact that the agent cannot foresee each action in an action sequence leading to a future state. It is straightforward and natural to model this with the E-operator, which uses the actual action sequence to a future state where we cannot assume that all actions can be foreseen by the agent. However, if we were to model surprising actions using the B-notation, which is using the agent's perspective on the action sequence, one could (rightfully) argue: why would the agent consider an action to happen when she does not consider it plausible? Usually, we do not expect a surprise to happen because if we would, it would not be a surprise any more.

3.4 Expectations vs. beliefs about future action outcomes

An obvious question is whether $[a_1; \cdots ; a_n]E_i\phi$ and $B_i[a_1^i; \cdots ; a_n^i]\phi$ always express the same thing. The answer is a conditional 'yes': It holds under certain conditions concerning the initial model and applicability, that we will now explore.

Lemma 3.1. *Let a state $(\mathcal{M}, w)$ and an action $(\mathcal{E}, e)$ be given, and let $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ and $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$. If $we, vf \in \mathcal{M} \otimes \mathcal{E}$, then $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$.*

Proof. To prove that vf is a least element, suppose the opposite. Then there exists some $v'f' <_i vf$ with $v'f' \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$. From the latter we get $v'f' \leqslant_i we$, implying $v' \leqslant_i w$ and $f' \leqslant_i e$, using the product update definition. From $v'f' <_i vf$ we get, also using the product update definition, that one of the following holds: 1) $f' <_i f$ and $v' \leqslant_i v$; 2) $f' \simeq_i f$ and $v' <_i v$. Suppose first that 1) holds. Then we have both $f' <_i f$ and $f' \leqslant_i e$, contradicting that $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$. Suppose instead that 2) holds. Then we have both $v' <_i v$ and $v' \leqslant_i w$, contradicting that $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$. □

Recall the definition of standard states in Definition 2.2. We now have:

Lemma 3.2. *Suppose an action $(\mathcal{E}, e)$ is both applicable and i-applicable in a standard state $(\mathcal{M}, w)$. If $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ then there exists $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$ such that $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$.*

Proof. By subjective applicability, we have $(\mathcal{M}, w) \models B_i\langle(\mathcal{E}, e)^i\rangle\top$. Since $(\mathcal{M}, w)$ is standard, $\sim_i = \leqslant_i$, and hence $(\mathcal{M}, w) \models E_i\langle(\mathcal{E}, e)^i\rangle\top$. By assumption on v, this implies $(\mathcal{M}, v) \models \langle(\mathcal{E}, e)^i\rangle\top$, i.e., $(\mathcal{M}, v) \models \langle(\mathcal{E}, \mathrm{Min}_{\leq_i}[e]_{\leqslant_i})\rangle\top$. From this we get an $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$ such that $(\mathcal{M}, v) \models pre(f)$, using Lemma 2.1(2). From $(\mathcal{M}, v) \models pre(f)$, we get $vf \in \mathcal{M} \otimes \mathcal{E}$. Since $(\mathcal{E}, e)$ is also (plainly) applicable in $(\mathcal{M}, w)$, we get $(\mathcal{M}, w) \models pre(e)$, and hence also $we \in \mathcal{M} \otimes \mathcal{E}$. We now satisfy all conditions of Lemma 3.1, which then gives $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$. □

Lemma 3.3. *Suppose $(\mathcal{E}, e)$ is both applicable and i-applicable in a standard state $(\mathcal{M}, w)$. If $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$ then $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ and $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$.*

Proof. By Lemma 3.2, $\mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$ must contain some element $w'e'$ where $w' \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ and $e' \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$. Since $w'e', vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$, we must have $w'e' \simeq_i vf$. This implies $w' \simeq_i v$ and $e' \simeq_i f$. But as $e' \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$ and $w' \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$, we then also get $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$ and $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$, as required. □

Theorem 3.3. *Suppose an action $a = (\mathcal{E}, e)$ is both applicable and i-applicable in a standard state $s = (\mathcal{M}, w)$. Then for all formulas ϕ, we have*

$$s \models B_i[a^i]\phi \leftrightarrow [a]E_i\phi.$$

Proof. ($\rightarrow$): Suppose $s \models B_i[a^i]\phi$. We need to prove $s \models [a]E_i\phi$, i.e., $s \otimes a \models E_i\phi$. This is equivalent to proving that for all $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$, $(\mathcal{M} \otimes \mathcal{E}, vf) \models \phi$. So let $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$ be chosen arbitrarily. By Lemma 3.3, we have that $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ and $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$. Since s is standard, we also get $v \in \mathrm{Min}_{\leq_i}[w]_{\sim_i}$. Since we have assumed $(\mathcal{M}, w) \models B_i[a^i]\phi$, we then get $(\mathcal{M}, v) \models [a^i]\phi$. Since $a^i = (\mathcal{E}, \mathrm{Min}_{\leq_i}[e]_{\leqslant_i})$ and $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$, we further get $(\mathcal{M}, v) \models [(\mathcal{E}, f)]\phi$, using Lemma 2.1(3). This proves $(\mathcal{M} \otimes \mathcal{E}, vf) \models \phi$, as required.

($\leftarrow$): Suppose $s \models [a]E_i\phi$. We need to prove $s \models B_i[a^i]\phi$. Since s is standard, this is equivalent to proving that for all $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$, $(\mathcal{M}, v) \models [a^i]\phi$. Since $a^i = (\mathcal{E}, \mathrm{Min}_{\leq_i}[e]_{\leqslant_i})$, this is further equivalent to proving that for all $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ and $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$, $(\mathcal{M}, v) \models [(\mathcal{E}, f)]\phi$, using again Lemma 2.1(3). So let $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ and $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$ be chosen arbitrarily. We then need to prove $(\mathcal{M}, v) \models [(\mathcal{E}, f)]\phi$, i.e., $(\mathcal{M} \otimes \mathcal{E}, vf) \models \phi$. From $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ and $f \in \mathrm{Min}_{\leq_i}[e]_{\leqslant_i}$, Lemma 3.1 gives that $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$. Since we have assumed $(\mathcal{M}, w) \models [a]E_i\phi$, we then get $(\mathcal{M} \otimes \mathcal{E}, we) \models E_i\phi$. Since $vf \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$, we get $(\mathcal{M} \otimes \mathcal{E}, vf) \models \phi$. □

The theorem expresses that under suitable applicability conditions, the two formulas express the same. For instance, when we want to express that, in s_0, Hanna believes that the coin toss will result in heads, it doesn't matter whether we do this by $B_h[toss^h]heads$ or by $[toss]E_h heads$. The first formula reads "Hanna believes that all most plausible executions of *toss* will result in heads", whereas the second reads "after any execution of *toss*, Hanna expects heads".

We earlier argued that the expectation operator gives a simpler and more intuitive way to express certain judgments about truth in plausibility models. Still, if $[a]E_i\phi$ and $B_i[a^i]\phi$ were always logically equivalent, one might argue that we could just introduce the former notation as an abbreviation of the latter, and avoid the expectation operator in the language. Theorem 3.3 shows that such an abbreviation is possible when considering actions that are both applicable and i-applicable and when applied in standard states. But it doesn't hold more generally than that, i.e., as soon as we drop either i-applicability or standardness, the biimplication of Theorem 3.3 is no longer generally true. Let us first consider dropping the condition of i-applicability. In Example 3.3, we showed that $\Uparrow(h, salt)$ is a surprising action to Hanna in s_1, i.e., it is not h-applicable in s_1. However, it is still clearly applicable. Intuitively, the point is that Rob can announce "it is salt" (the action is applicable), but if Hanna's beliefs were true, he wouldn't be able to sincerely make this announcement (it is not subjectively applicable to Hanna). Let us now show that the biimplication of Theorem 3.3 fails when $s = s_1$, $a = \Uparrow(h, salt)$ and $\phi = sugar$. For the left-hand side, note that $s_1 \models B_h[\Uparrow(h, salt)^h]sugar$ iff $(\mathcal{M}_1, w) \models [(\mathcal{E}_{\Uparrow(h,salt)}, f)]sugar$, which trivially holds, since $w : sugar$ doesn't satisfy $pre(f) = salt$. For the right-hand side, we get $s_1 \models [\Uparrow(h, salt)]E_h sugar$ iff $s_1 \otimes \Uparrow(h, salt) \models E_h sugar$, which fails since in $s_1 \otimes \Uparrow(h, salt)$ the world most plausible to Hanna is vf where *salt* holds. Since the left-hand side is true and the right-hand side false, the biimplication fails. The right-hand side expresses: "after Rob announces *salt*, Hanna expects *sugar*", which is false, whereas the left-hand side expresses: "supposing that Hanna's current beliefs are true, if Rob would announce *salt*, then *sugar*", which is vacuously true since Rob cannot sincerely make such an announcement in any of Hanna's believed worlds. So in this case, only the formula involving the expectation operator is formally capturing the intended, since it correctly captures that Hanna does *not* expect that if Robs announces *salt*, it will be *sugar*, since she knows that she will trust the announcement and it will make her change her mind (this is encoded in the event model for the belief update).

Let us now analyze the situation when an action is applied in a state that is not standard. Here we can for instance take $s = s_0 \otimes opentoss$ of Figure 5, let a be a singleton 'skip' action, i.e., a has a single event e with $pre(e) = post(e) = \top$, and let $\phi = heads$. For the left-hand side of the biimplication, we then get $s \models B_h[a^h]heads$ iff $[a^h]heads$ is true in both worlds of s. The formula $[a^h]heads$ fails in the world wf of s where *tails* holds (since the action a is a skip action that changes nothing). Hence the left-hand side is false. For the right-hand side, we get $s \models [a]E_h heads$ iff $s \otimes a \models E_h heads$, which is true, since $s \otimes a$ is isomorphic to s, and in both worlds of s, Hanna expects heads (since $we <_h wf$). So here the left-hand side is false and the right-hand side is true. Again, the formula on the right-hand side is the only one capturing the intended: If we ask about Hanna's expectations about the situation

following from performing a skip action after an unobserved biased coin toss, then she expects the coin to show heads. These examples help to motivate the introduction of the expectation operator as a separate operator in the language, and help to see why we can't always express our expectation judgments by a belief operator followed by an (a priori perspective on an) action sequence.

4 False beliefs and false expectations

Bolander [20] defines false beliefs in a DEL-based setting. Here we additionally define false expectations, using our new expectation operator.

Definition 4.1. *Let i be an agent, ϕ be a formula, s a state, and a an action applicable in s. We define the following notions:*

1. *Agent i* falsely believes *ϕ in s if $s \models \neg\phi \wedge B_i\phi$.*
2. *Agent i* falsely expects *ϕ in s if $s \models \neg\phi \wedge E_i\phi$.*

Example 4.1. Let us apply these definitions to our running example. In s_1 of Figure 1, Hanna falsely believes that the dispenser contains sugar. In s_2 of Figure 1, she doesn't have this false belief, we are reasoning about a situation where she has already tasted. However, in s_2, she still falsely expects the dispenser to contain sugar, since a priori she finds the *sugar* world of s_2 more plausible than the *salt* world (she will only know that it's salt when arriving in the future state s_2).

Consider again the event models *pour* and *sip* introduced in Example 3.1 and illustrated in Figure 3. In Example 3.1, we concluded that Hanna initially believes that pouring and sipping will lead to delight or, equivalently, not to disgust: $s_1 \models B_h[pour; sip]\neg disgust$. Since $pour; sip$ has two designated, equiplausible events (cf. Example 3.1), we get $(pour; sip)^h = pour; sip$. Thus, Hanna has the correct a priori perspective on this action composition, cf. Definition 3.1. Intuitively, this means that she has no confusion about its possible outcomes. From $s_1 \models B_h[pour; sip]delight$ and $(pour; sip)^h = pour; sip$, we get $s_1 \models B_h[(pour; sip)^h]\neg disgust$. Since both applicability and subjective applicability of $pour; sip$ in s_1 trivially holds (the two events of $pour; sip$ are both designated), we can apply Theorem 3.3 to conclude that $s_1 \models [pour; sip]E_h\neg disgust$. In other words, Hanna expects $\neg disgust$ to follow from performing the actions. It might have been simpler to prove this fact directly using product updates and the E_h operator, but we wanted to provide another example of the connection between putting the belief operator in front of an action sequence and the expectation operator after.

Rob knows that Hanna's expectations are false, and that instead the action sequence *will* lead to disgust, $s_1 \models K_r[pour; sip]disgust$. Putting this together with what we concluded in the previous paragraph, we get:

$$s \models K_r[pour; sip](disgust \wedge E_h\neg disgust) \tag{1}$$

This expresses that Rob knows that Hanna has a false expection that pouring and sipping will not lead to disgust. Disgust is usually undesirable, and since Rob now knows that Hanna is falsely expecting the actions not to result in undesirability, it poses an opportunity for Rob to intervene. He could inform Hanna that the content of the dispenser is salt (the event model $\Uparrow(h, salt)$ of Figure 2). But *when* should he do

that? It does not make sense to announce "the dispenser contains salt" before Hanna even considers having coffee. It also doesn't make sense to announce it after she already has taken a sip, where it will be too late to avoid disgust. For the announcement to be *relevant*, it has to be announced at the *right* time. We explore this more formally in Section 5.

Suppose Rob decides to announce initially, in state s_1, that the dispenser contains salt. We can reason about how inserting this announcement would change the expected outcome of the action sequence from before:

$$s \models K_r[\Uparrow(h, salt); pour; sip](disgust \wedge E_h\, disgust) \tag{2}$$

This tells Rob that if he makes the announcement initially, then he can avoid Hanna having a false expectation that her actions will not result in disgust. Hence Rob can give Hanna information to avoid an undesirable outcome: When Hanna realizes that her planned action sequence leads to disgust, she gets the opportunity of choosing another line of action. We will explore more generally how to define and compute such 'relevant announcements' in the following section.

Let us conclude the example by briefly relating (2) to what we would have achieved by trying to capture Hanna's expectations using a combination of her belief operator, B_h, and her a priori perspective on the action sequence: $\Uparrow(h, salt)^h, pour^h, sip^h$. As shown in Example 3.3, $s_1 \not\models B_h\langle\Uparrow(h, salt)^h\rangle\top$, i.e., $\Uparrow(h, salt)$ is a surprise to Hanna in s_1. From this we conclude $s_1 \models B_h[\Uparrow(h, salt)^h]\bot$ (from $s_1 \not\models B_h\langle\Uparrow(h, salt)^h\rangle\top$ we get $(\mathcal{M}, w) \not\models \langle\Uparrow(h, salt)^h\rangle\top$ implying that $\Uparrow(h, salt)^h$ is not applicable in $(\mathcal{M}, w)$, and hence $(\mathcal{M}, w) \otimes \Uparrow(h, salt)^h \models \bot$, implying $(\mathcal{M}, w) \models [\Uparrow(h, salt)^h]\bot$ and hence $s_1 \models B_h[\Uparrow(h, salt)^h]\bot$). This means that we cannot even express the content of (2) using a combination of B_h and the a priori perspective on the action sequence, since it trivializes to having $s_1 \models B_h[\Uparrow(h, salt)^h; pour^h; sip^h]\bot$. This is because the belief update is a surprising action to Hanna, she would not consider it using her perspective on it, since she believes it to be inapplicable (as she is convinced that *sugar* is true). If we didn't have the expectation operator, we would have to find another way to be able to represent the content of (1) and (2), e.g. add a new type of 'subjective' product update operator, an alternative solution recently suggested by Pieper [11].

5 Undesirability and relevant announcements

5.1 Undesirability

In the kind of human-robot interaction we focus on here, the role of the robot (Rob) is to inform the human (Hanna) if she has false expectations about future outcomes. He does so by making soft announcements. In a domestic or industrial setting, Rob might intervene, announcing e.g. "don't touch that, it is still hot". The robot has to be told which states we try to avoid, i.e., which are *undesirable*, either by being dangerous (leading to injury) or unpleasant (leading to discomfort). In reality, desirability might be graded like in the work of Grosinger et al. [21], but here we are only interested in avoiding undesirable states, and everything not undesirable is then for simplicity just called *desirable*. Note that this gives a rather weak notion of desirability, as a state

being desirable only means that it is not one of the (usually few) states that the agent would necessarily want to avoid.

Definition 5.1. *Let i be an agent. An* undesirability formula *(or simply, an* undesirability*) for i is a formula $U \in \mathcal{L}$ representing the undesirable states for agent i. Suppose a fixed undesirability formula U for i is given. A state s is then called* undesirable *for i if $s \models U$, and otherwise* desirable*. Given a plausibility model $\mathcal{M}$, a world w of $\mathcal{M}$ is called* undesirable *for i if $(\mathcal{M}, w) \models U$, otherwise* desirable*.*

Undesirability formulas can be either propositional or epistemic formulas. Propositional undesirability formulas can be used to e.g. represent states of physical injury, discomfort or disgust. In the coffee scenario, we might for instance have $U = \mathit{disgust} \vee \mathit{burned}$ representing that Hanna neither wants to get disgusted (by drinking salty coffee), nor burn her tongue (by drinking too hot coffee). More generally, if we have a set of formulas $\{\phi_1, \ldots, \phi_n\}$ where each formula represents a particular type of undesirable outcome, we can let $U = \phi_1 \vee \cdots \vee \phi_n$. One could also consider desirability of epistemic formulas, e.g. agent i might consider it undesirable for agent j to get to know that ψ, and then we could let $K_j\psi$ be the undesirability formula for i. However, our main motivation for developing the formalism is to allow a robot to warn about physical dangers and discomfort, so we will mainly be occupied with propositional undesirability formulas.

5.2 Relevant announcements

Definition 5.2. *Let U be an undesirability formula for an agent i, and $a_1, \ldots, a_n$ an applicable action sequence in a state s. A formula ϕ is a* relevant announcement *(wrt. i, U, $a_1, \ldots, a_n$, and s) if the following holds:*

$$s \models \neg U \wedge [a_1; \cdots; a_n](U \wedge E_i \neg U) \wedge [\Uparrow(i, \phi); a_1; \cdots; a_n](U \to E_i U)$$

The first conjunct expresses that the current state is desirable. The second conjunct expresses that the outcome of the action sequence is falsely expected to be desirable. The third conjunct expresses that if ϕ is announced first, i will correct her false expectation (hence allowing her to reconsider her actions).

Example 5.1. Given the undesirability $U = \mathit{disgust}$ for Hanna, then *salt* is a relevant announcement wrt. the action sequence $\mathit{pour}, \mathit{sip}$ in the state s_1 of Figure 1 (this follows from (1) and (2) of Example 4.1). Note that we even have that Rob *knows* that ϕ is a relevant announcement, as Rob knows the formula expressing the relevance of ϕ. Hence it is a relevant announcement *for* Rob to announce to Hanna. When Rob has announced this to Hanna, she can reconsider her actions, for instance decide to drink the coffee plain, or look for sugar elsewhere. Note that we defined undesirability in terms of *disgust*, so the announced formula *salt* is not just a warning saying "you will get disgusted", but it's a formula whose announcement is sufficient to correct Hanna's false expectation of becoming delighted. Another relevant announcement would be $\neg \mathit{sugar}$, amounting to the same in this case.

Theorem 5.3. *Let $U \in \mathcal{L}_{prop}$ be an undesirability formula for an agent i, and $a_1, \ldots, a_n$ an applicable action sequence in a state s. If a relevant annnouncement*

exists wrt. these elements, then a relevant announcement $\phi \in \mathcal{L}_{prop}$ exists. Furthermore, deciding whether it exists, and computing it if it does, can both be done in time polynomial in $|s \otimes a_1 \otimes \cdots \otimes a_n| + |U|$.

Proof. Let $s = (\mathcal{M}, W_d)$ with $\mathcal{M} = (W, \sim_i, \leq_i, L)_{i \in Ag}$ and $a_1; \cdots ; a_n = (\mathcal{E}, E_d)$ with $\mathcal{E} = (E, \sim_i, \leq_i, pre, post)_{i \in Ag}$. Let $P' \subseteq P$ denote the set of atoms occurring in s, i.e., $P' = \bigcup_{w \in W} L(w)$. Then each world $w \in W$ has a characteristic formula $\delta(w) = \bigwedge_{p \in L(w)} p \wedge \bigwedge_{p \in P' \setminus L(w)} \neg p$, i.e., a formula $\delta(w)$ that is only true at worlds w' with $L(w') = L(w)$. To denote elements of W, we will exclusively be using the symbols $w, w', w_0, w_1, \ldots$, and to denote elements of E, we exclusively use $e, e', e_0, e_1, \ldots$ Hence we can abbreviate notation and e.g. write $we \models \phi$ for $(\mathcal{M} \otimes \mathcal{E}, we) \models \phi$. When we write $we \models \phi$, it is also assumed that $w \models pre(e)$ (otherwise the statement $we \models \phi$ is not even well-defined). Define

$$W_{avoid} = \{w \in W \mid \exists e \in E \text{ s.t. } we \models \neg U \text{ and } \forall e' <_i e \text{ with } w \models pre(e'), we' \models \neg U\}$$

W_{avoid} is intuitively the set of worlds we need to make least plausible to i by our announcement $\Uparrow(i, \phi)$ to ensure that when afterwards applying $\mathcal{E}$, they don't become most plausible worlds and make $E_i U$ false, cf. the last conjunct in Definition 5.2. Let $\phi = \bigwedge_{w \in W_{avoid}} \neg\delta(w)$. We want to show that if there exists a relevant announcement γ, then ϕ is also a relevant announcement. To denote events of $\Uparrow(i, \gamma)$, we will exclusively be using symbols $g, g', g_0, g_1, \ldots$, and for $\Uparrow(i, \phi)$ symbols $f, f', f_0, f_1, \ldots$, again to allow abbreviated notation. We use $g_{\neg\gamma}$ to refer to the event of $\Uparrow(i, \gamma)$ with precondition $\neg\gamma$ (the least plausible to i), and g_γ for the other. Similarly with $f_{\neg\phi}$ and f_ϕ for $\Uparrow(i, \phi)$.

Claim 1. Suppose

$$s \models [\Uparrow(i, \gamma); a_1; \cdots ; a_n] E_i U \tag{3}$$

Then

$$s \models [\Uparrow(i, \phi); a_1; \cdots ; a_n] E_i U \tag{4}$$

Proof of Claim 1. First note that (3) expresses the last conjunct of the condition for γ being a relevant announcement, and (4) similarly for the announcement of ϕ. To achieve a contradiction, suppose (3) is true and (4) is false. Since (4) is false, there exists a designated world $w_d f_d e_d$ of $s \otimes \Uparrow(i, \phi); a_1; \cdots ; a_n$ and $w_1 f e_1 \in \mathrm{Min}_{\leq_i}[w_d f_d e_d]_{\leqslant_i}$ s.t. $w_1 f e_1 \models \neg U$. Since $\Uparrow(i, \phi)$ has empty postconditions (all $post = \top$), $w_1 f e_1$ and $w_1 e_1$ satisfy the same propositional formulas. Hence $w_1 e_1 \models \neg U$. Suppose $e' <_i e_1$ with $w_1 \models pre(e')$. Then $w_1 f \models pre(e')$ (as $\Uparrow(i, \phi)$ has empty postconditions), and since $e' <_i e_1$, we get $w_1 f e' <_i w_1 f e_1$, contradicting minimality of $w_1 f e_1$. Hence no such e' exists, and since $w_1 e_1 \models \neg U$, we can conclude $w_1 \in W_{avoid}$. Since now $w_1 \models \delta(w_1)$ and $w_1 \in W_{avoid}$, by definition of ϕ, we get $w_1 \models \neg\phi$. Since $pre(f_\phi) = \phi$, $pre(f_{\neg\phi}) = \neg\phi$ and $w_1 f \in s \otimes \Uparrow(i, \phi)$, we can conclude $f = f_{\neg\phi}$. So we have $w_1 f_{\neg\phi} e_1 \in \mathrm{Min}_{\leq}[w_d f_d e_d]_{\leqslant_i}$ and $w_1 f_{\neg\phi} e_1 \models \neg U$. Since $\Uparrow(i, \phi)$ has no postconditions and $w_d f_d e_d$ is a designated world of $s \otimes \Uparrow(i, \phi); a_1; \ldots ; a_n$, there exists a g_d so that $w_d g_d e_d$ is a designated world of $s \otimes \Uparrow(i, \gamma); a_1; \ldots ; a_n$. Let $w_2 g e_2 \in \mathrm{Min}_{\leq_i}[w_d g_d e_d]_{\leqslant_i}$. From (3) we get $w_2 g e_2 \models U$. We now first prove by contradiction that $w_2 \models \neg\phi$. So suppose $w_2 \models \phi$. Then $w_2 \models pre(f_\phi)$. Since $w_2 g e_2$ is a world and g has no postconditions, we must have $w_2 \models pre(e_2)$, and since also $pre(f_\phi)$ has no postconditions, $w_2 f_\phi \models pre(e_2)$. Thus the world $w_2 f_\phi e_2$ exists. Since $f_\phi <_i f_{\neg\phi}$,

we get $w_2 f_\phi <_i w_1 f_{\neg\phi}$. We can now conclude that $e_1 <_i e_2$, since otherwise we could use $w_2 f_\phi <_i w_1 f_{\neg\phi}$ to conclude $w_2 f_\phi e_2 <_i w_1 f_{\neg\phi} e_1$, contradicting the minimality of $w_1 f_{\neg\phi} e_1$. Since $w_1 \models pre(e_1)$, also $w_1 g' \models pre(e_1)$ for some g' and hence also $w_1 g' e_1$ is a world. From $e_1 <_i e_2$ we can now conclude $w_1 g' e_1 <_i w_2 g e_2$, but this contradicts minimality of $w_2 g e_2$, and hence we have proved $w_2 \models \neg\phi$. Then by definition of ϕ we have $w_2 \models \delta(w_3)$ for some $w_3 \in W_{avoid}$. By definition of W_{avoid}, this implies the existence of an e_3 s.t. $w_3 e_3 \models \neg U$. Since $w_2 \models \delta(w_3)$, we have $L(w_2) = L(w_3)$ and hence from $w_3 e_3 \models \neg U$ we get $w_2 e_3 \models \neg U$, and hence also $w_2 g e_3 \models \neg U$. We either have $e_2 <_i e_3$ or $e_2 \geq_i e_3$. Suppose first that $e_2 <_i e_3$. Then by choice of e_3 and definition of W_{avoid}, we get $w_2 e_2 \models \neg U$ and hence $w_2 g e_2 \models \neg U$, contradicting that we also have $w_2 g e_2 \models U$. Suppose alternatively that $e_2 \geq_i e_3$. Then $w_2 g e_2 \geq_i w_2 g e_3$, implying that also $w_2 g e_3 \in \mathrm{Min}_{\leq_i}[w_d g_d e_d]_{\leqslant_i}$, but then we have $w_2 g e_3 \models U$ from (3), again a contradiction. This completes the proof of Claim 1.

Let $n = |s \otimes a_1 \otimes \cdots \otimes a_n| + |U_i|$. Suppose a relevant announcement γ exists. Then $s \models \neg U$, $s \models [a_1; \ldots; a_n](U \wedge E_i \neg U)$ and $s \models [\Uparrow(i,\gamma); a_1; \cdots; a_n](U \to E_i U)$. We want to prove that ϕ is also a relevant announcement, i.e., we want to prove that $s \models [\Uparrow(i,\phi); a_1; \cdots; a_n](U \to E_i U)$. As U is a propositional formula and $s \models [a_1; \ldots; a_n]U$, we also get $s \models [\Uparrow(i,\gamma); a_1; \ldots; a_n]U$. As $s \models [\Uparrow(i,\gamma); a_1; \cdots; a_n](U \to E_i U)$, we can then conclude $s \models [\Uparrow(i,\gamma); a_1; \cdots; a_n] E_i U$. Applying Claim 1, we now get $s \models [\Uparrow(i,\phi); a_1; \cdots; a_n] E_i U$, as wanted. Only left is to show that we can in polynomial time in n compute ϕ and decide whether $(\mathcal{M}, W_d) \models \neg U \wedge [\mathcal{E}, E_d](U \wedge E_i \neg U) \wedge [\Uparrow(i,\phi); (\mathcal{E}, e)](U \to E_i U)$. As model checking a propositional formula is polynomial in the length of the formula, and as there are $\leq n$ worlds in W_d, checking the first conjunct is polynomial in n. To check the second conjunct, we first compute $\mathcal{M} \otimes \mathcal{E}$, which can be done in polynomial time in n (see van de Pol et al. [22] and Bolander and Lequen [23] for details on the complexity of computing product updates). We now need to check whether $U \wedge E_i \neg U$ holds in all worlds $we \in W_d \times E_d$ of the computed product update. There are at most n worlds in the updated model, and since model checking of propositional formulas is polynomial, we can in polynomial time determine which of the worlds of the updated model satisfy U. Suppose we have already computed this, and marked the worlds satisfying U. For each $we \in W_d \times E_d$, we can then in polynomial time determine both whether $we \models U$ and whether $w'e' \models \neg U$ for all $w'e' \in \mathrm{Min}_{\leq_i}[we]_{\leqslant_i}$. Thus in polynomial time we also get to check the truth of the second conjunct. For the third conjunct, we need to compute W_{avoid}, which can now also be done in polynomial time (given the existing marking of which worlds of the updated model satisfy U). From W_{avoid}, we compute all the $\delta(w)$ with $w \in W_{avoid}$, which can clearly also be done in polynomial time, and hence we can compute ϕ in polynomial time. We now need to compute $\mathcal{M} \otimes \mathcal{E}_{\Uparrow(i,\phi)} \otimes \mathcal{E}$, which again can be done in polynomial time in the size of the resulting model (which is of the same size as $\mathcal{M} \otimes \mathcal{E}$). Finally, we need to check $U \to E_i U$ in this model, which by the same argument as before can be done in polynomial time. □

Example 5.2. For our running example with $U = \mathit{disgust}$, the set W_{avoid} computed in the proof of Theorem 5.3 is $\{w\}$, since the only world of s_1 in which Hanna can avoid getting disgusted is w. The formula ϕ computed in the proof is then $\neg\delta(w) =$

$\neg(sugar \wedge \neg salt)$, equivalent to the relevant announcements we previously computed by hand.

Note that Theorem 5.3 only covers the case of propositional undesirability formulas. In all our examples, undesirability formulas are propositional, which also as earlier mentioned is our primary focus. However, it is of course potentially interesting to still study the complexity of computing relevant announcements for arbitrary undesirability formulas. We leave this for future work. Note also that the complexity result is stated in terms of the size of $|s \otimes a_1 \otimes \cdots \otimes a_n|$, which is often not polynomial in $|s| + |a_1| + \cdots + |a_n|$ [23]. We define the *size* $|\phi|$ of a formula ϕ in the standard way, counting its number of symbols (see, for example, van Ditmarsch et al. [24]). We say that ϕ is *shorter* than ψ if $|\phi| < |\psi|$.

Definition 5.4. *A* minimal *relevant announcement is a shortest formula that is a relevant announcement (wrt.* $i, U, a_1 \cdots a_n$*, and* s*).*

The relevant announcement ϕ computed in the proof of Theorem 5.3 is not necessarily minimal in the sense of Definition 5.4. However, as finding the shortest propositional formula equivalent to another is NP-hard [25], we cannot in general preserve the polynomial bound if we also want to guarantee that the announced formula is minimal.

5.3 Relevant future announcements

There could be challenges in making a relevant announcement too early:

1. Uncertainty about the future. So far we have assumed a unique sequence of future actions $a_1, \ldots, a_n$ to be given. However, in practice, we cannot expect Rob to know exactly what actions will take place (see the elaborate example in Section 7). In a domestic or industrial setting, humans might often perform the same or similar action sequences. Hence, it might be reasonable to expect that Rob can learn a decently accurate probability distribution over action sequences of some length. However, the longer the action sequences, the less certain Rob will be. Given an unpredictable future, Rob might compute a relevant announcement with respect to *some* action sequence (of potentially many possible futures) that will actually not take place, and where the announcement might be a nuisance to the human (not relevant to the *actual* action sequence). For instance, if Hanna is going to the toilet, she might not care much about the content of the sugar dispenser. Prediction of future states could also be done with plan/goal recognition, see Section 8.

2. Taking into account the context. There may be reasons for making a relevant announcement later than immediately when it is inferred. The current state and wider context might play a role for when making the announcement is *best*. Consider Hanna being in the coffee room, just about to take her coffee. Rob infers that a relevant announcement is *salt* to avoid future *disgust*. Consider further that Hanna is on the phone. Should Rob interrupt her call to tell her about the dispenser containing salt? Certainly not. After she has hung up might be a better time (cf. annoying TUG hospital robots making relentless announcements [26]). Suppose Hanna is on the phone while Rob has a relevant announcement that she needs to take insulin. Then interrupting her might actually be good as it is a situation threatening her health. So, in

practice, there is a balance to strike between the urgency of Rob's announcement and how much it intrudes Hanna.

To avoid the problems of early announcements potentially being irrelevant or annoying, Rob should consider to instead make a later announcement that would still prevent undesirability. To this end we define *future relevant announcements*.

Definition 5.5. *Let U be an undesirability formula for an agent i, and let $a_1, \ldots, a_n$ be an applicable action sequence in state s. A formula ϕ is a* relevant announcement *at time t, $1 \leq t \leq n$, (wrt. U, $a_1, \ldots, a_n$, and s) if it is a relevant announcement in the state $s \otimes a_1 \otimes \cdots \otimes a_t$. The formula ϕ is a* relevant future announcement *if it is a relevant announcement at some time $t \geq 1$.*

Example 5.3. In our running example with $U = \mathit{disgust}$ and action sequence $a_1, a_2 = \mathit{pour}, \mathit{sip}$, we have

$$s_1 \models [\mathit{pour}]\big(\neg U \wedge [\mathit{sip}](U \wedge E_i \neg U) \wedge [\Uparrow(i, \mathit{salt}); \mathit{sip}](U \wedge E_i U)\big)$$

This means that announcing the true content of the dispenser is also a relevant announcement at time 1 (after pouring). There would then still be time to avoid the undesirability of tasting salty coffee. However, since we also have $s_1 \models [\mathit{pour}; \mathit{sip}](U \wedge E_i U)$, announcing at time 2 is not relevant (it's too late to avoid the undesirability of disgust). Note that whether a formula is a relevant future announcement depends on the time at which the announcement is made. At time 1, a relevant announcement of Rob regarding U is $\neg \mathit{sweet}$, but this is clearly not relevant at time 0, in s_1, where the relevant announcement regarding U is salt.

Given Definition 5.5, one can compute the latest time points at which it is possible to make relevant announcements regarding a specific undesirability U for each considered future action sequence $a_1, \ldots, a_n$ (hence avoiding the mentioned problems of early announcements). However, announcing as late as possible is not always the best strategy. In the example above, if Rob plans to announce *salt* at time 1, then it is sufficiently early to avoid disgust, but too late to avoid salty coffee. In this particular example, one could simply add $\neg \mathit{sweet}$ as a disjunction to the undesirability formula, but *degrees* of undesirability might give a more general solution, allowing for a more fine-grained notion of relevant announcement (see Section 9). Another approach could be goal recognition: Rob might recognize that Hanna's most likely current goal is to have a cup of sweetened coffee (for instance if she usually has that at this time of day). He could then compute the cost of the revised plan that Hanna would need to execute depending on when he makes his announcement. He could infer that announcing *salt* immediately is best, as otherwise Hanna would have to make a new cup of coffee after having poured salt into her cup. This discussion suggests many open questions to be investigated regarding the "windows of opportunity" for relevant announcements.

There could also be challenges in making a relevant announcement too late. There could in principle be scenarios having a "point of no return", a future point in time from which all action sequences of a certain length will lead to an undesirable outcome, and hence announcing at that point would be too late.

6 Axiomatization

In our logic, not all modalities have a standardly defined semantics where $\Box\phi$ means that ϕ is true in all worlds accessible from the current world by some relation R. This makes axiomatization non-trivial, as we cannot trivially rely on the standard canonical model constructions [27]. However, we can introduce new modalities that make the logic stronger and more straightforward to axiomatize. This path was also explored by Baltag and Smets [9, Theorem 2.5] for their logic $K\Box$. Their logic has a knowledge modality K_i and a "safe belief" modality $\Box_i$, for each agent i. The semantics of their safe belief modality is standardly defined in terms of $\geq_i$ (where $\geq_i$ is the inverse of $\leq_i$): $\mathcal{M}, w \models \Box_i\phi$ iff for all v with $w \geq_i v$, $\mathcal{M}, v \models \phi$. The semantics of their K_i modality is defined in terms of an epistemic equivalence relation $\sim_i$, however, it is not generally the same relation as in our logic, since we don't require $\sim_i = \leqslant_i$. They use their knowledge operator to axiomatize local connectedness of $\leq_i$. As we don't have the same operator, we need to extend our language in order to be able to do the same. More precisely, we replace our original language $\mathcal{L}$ by the language $\mathcal{L}'$ defined by:

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi \wedge \phi \mid K_i\phi \mid \vec{K}_i\phi \mid E_i\phi \mid \Box_i\phi \mid [a]\phi \qquad (\mathcal{L}')$$

We read $\vec{K}_i\phi$ as "agent i a priori knows ϕ". The arrow on the $\vec{K}_i$ operator signifies that it is an a priori modality that "looks into the future", similarly to the expectation operator. We still read $K_i\phi$ as "agent i knows ϕ"; or, alternatively, "agent i a posteriori knows (or will know) ϕ" when we want to make the distinction between the a priori and a posteriori knowledge operators more clear. We read $\Box_i\phi$ as "agent i (a priori) safely believes ϕ". The dual operator $\Diamond_i\phi$ is defined by abbreviation in the standard way: $\Diamond_i\phi := \neg\Box_i\neg\phi$.

Note that there is no longer a belief modality, B_i, in the language $\mathcal{L}'$. The belief modality was introduced in order to be able to explain and defend the expectation operator (Section 3), but the belief modality plays no role in our framework for modeling undesirability and relevant announcements (Section 5). By omitting it from the language, axiomatization becomes simpler.[7] The semantics of $\mathcal{L}'$ is exactly as for the semantics of $\mathcal{L}$, with the following clauses added for the new modalities:

$$(\mathcal{M}, w) \models \Box_i\phi \text{ iff for all } v \leq_i w, (\mathcal{M}, v) \models \phi$$
$$(\mathcal{M}, w) \models \vec{K}_i\phi \text{ iff for all } v \leqslant_i w, (\mathcal{M}, v) \models \phi$$

The following result mimics a result by Baltag and Smets [9], but as our setting differs slightly, we still provide the details.

Proposition 6.1. *For all formulas ϕ of $\mathcal{L}'$, the formula $E_i\phi \leftrightarrow \Diamond_i\Box_i\phi$ is valid.*

[7] Actually, axiomatization becomes significantly simpler without the belief modality. Note that the semantics of the belief modality is defined in terms of both $\leq_i$ and $\sim_i$. It is possible to define the belief modality in terms of a new a posteriori safe belief modality, but the semantics of this new modality is then defined in terms of $\sim_i \cap \leq_i$, which requires us to axiomatize the intersection of two relations. We can use a similar trick as for axiomatizing distributed knowledge [28], however, to ensure that $\leq_i$ becomes a union of mutually disjoint well-preorder, the canonical model we build has to be finite. For standard canonical models, this is simple to achieve using filtrations, but for the path-based models of the cited paper, it is significantly less obvious how to achieve it (while still preserving the relevant properties).

Proof. Left to right: Suppose $(\mathcal{M}, w) \models E_i\phi$, that is, suppose $(\mathcal{M}, v) \models \phi$ for all $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$. Choose some $v' \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$. For any $v \leq_i v'$ we then have $v \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$. Thus, for any $v \leq_i v'$, we have $(\mathcal{M}, v) \models \phi$, and hence $(\mathcal{M}, v') \models \Box_i\phi$. From $v' \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$ we get $v' \leq_i w$, and thus $(\mathcal{M}, w) \models \Diamond_i\Box_i\phi$, as required. *Right to left:* Suppose $(\mathcal{M}, w) \models \Diamond_i\Box_i\phi$. Then for some $v' \leq_i w$, $(\mathcal{M}, v') \models \Box_i\phi$. Now choose any $v'' \in \mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$. Since v'' is a $\leq_i$-least element in $[w]_{\leqslant_i}$ and since $v' \in [w]_{\leqslant_i}$, we get $v'' \leq_i v'$. Since $(\mathcal{M}, v') \models \Box_i\phi$, this implies $(\mathcal{M}, v'') \models \phi$. As v'' was chosen arbitrarily in $\mathrm{Min}_{\leq_i}[w]_{\leqslant_i}$, we can conclude that $(\mathcal{M}, w) \models E_i\phi$. □

Theorem 6.1. *Let Λ be the smallest normal logic of $\mathcal{L}'$ containing:*

1. K_i *axioms for both* K_i *and* $\vec{K}_i$
2. $S4$ *axioms for* $\Box_i$
3. $\vec{K}_i\phi \to K_i\phi$
4. $\vec{K}_i\phi \to \Box_i\phi$
5. $\Diamond_i\phi \wedge \Box_i\psi \to \vec{K}_i(\Diamond_i\phi \vee \Box_i\psi)$[8]
6. $E_i\phi \leftrightarrow \Diamond_i\Box_i\phi$
7. $[(\mathcal{E}, E_d)]\phi \leftrightarrow \bigwedge_{e\in E_d}[(\mathcal{E}, e)]\phi$
8. $[(\mathcal{E}, e)]p \leftrightarrow (pre(e) \to p)$ *if* $p \notin post(e)$
9. $[(\mathcal{E}, e)]p \leftrightarrow pre(e)$ *if* $p \in post(e)$
10. $[(\mathcal{E}, e)]\neg\phi \leftrightarrow (pre(e) \to \neg[(\mathcal{E}, e)]\phi)$
11. $[(\mathcal{E}, e)](\phi \wedge \psi) \leftrightarrow (pre(e) \to ([(\mathcal{E}, e)]\phi \wedge [(\mathcal{E}, e)]\psi))$
12. $[(\mathcal{E}, e)]K_i\phi \leftrightarrow \big(pre(e) \to \bigwedge_{f\sim_i e} K_i[(\mathcal{E}, f)]\phi\big)$
13. $[(\mathcal{E}, e)]\vec{K}_i\phi \leftrightarrow \big(pre(e) \to \bigwedge_{f\leqslant_i e} \vec{K}_i[(\mathcal{E}, f)]\phi\big)$
14. $[(\mathcal{E}, e)]\Box_i\phi \leftrightarrow \big(pre(e) \to \bigwedge_{f<_i e} \vec{K}_i[(\mathcal{E}, f)]\phi \wedge \bigwedge_{f\simeq_i e} \Box_i[(\mathcal{E}, f)]\phi\big)$

Λ is sound and complete for the class of finite plausibility models.

Proof. Soundness amounts to proving validity of all our axioms (noting that our class of models is defined exclusively in terms of properties of the underlying frames). First note that any plausibility model $\mathcal{M} = (W, \sim_i, \leq_i, L)_{i\in Ag}$ satisfies the following (see Definition 2.1 and the discussion that follows): $\sim_i$ and $\leqslant_i$ are equivalence relations, $\leq_i$ is a preorder (reflexive and transitive), and $\sim_i \subseteq \leqslant_i$. The S5 axioms (item 1 above) hold since $\sim_i$ and $\leqslant_i$ are equivalence relations. The S4 axioms (item 2 above) hold since $\leq_i$ is a preorder. The third axiom holds since $\sim_i \subseteq \leqslant_i$. The fourth axiom holds since $\leq_i \subseteq \leqslant_i$. To prove validity of the fifth axiom, suppose $(\mathcal{M}, w) \models \Diamond_i\phi \wedge \Box_i\psi$ and let $v \leqslant_i w$. We need to prove $(\mathcal{M}, v) \models \Diamond_i\phi \vee \Box_i\psi$. From $v \leqslant_i w$, we get that either $w \leq_i v$ or $v \leq_i w$. If $w \leq_i v$, then since $(\mathcal{M}, w) \models \Diamond_i\phi$, we get $(\mathcal{M}, v) \models \Diamond_i\Diamond_i\phi$, and hence $(\mathcal{M}, v) \models \Diamond_i\phi$ from the S4 axioms for $\Box_i$. If $v \leq_i w$, then since $(\mathcal{M}, w) \models \Box_i\psi$ implies $(\mathcal{M}, w) \models \Box_i\Box_i\psi$ (by the S4 axioms again), we get $(\mathcal{M}, v) \models \Box_i\phi$. This shows that $(\mathcal{M}, v) \models \Diamond_i\phi \vee \Box_i\psi$, as required. Axiom 6 holds by Proposition 6.1. Axiom 7 holds by Lemma 2.1, item 3. Axioms 8–12 are the standard for dynamic epistemic logic with postconditions [14] (except for the syntactic differences arising from us representing postconditions as conjunctions of literals instead of mappings from atoms to formulas). Axiom 13 corresponds to axiom 12, but for the a priori knowledge modality. Axiom

[8] This is a slightly simpler axiom than the corresponding axiom for connectedness in Baltag and Smets [9]. The simplification used here was proposed by Alexandru Baltag in personal communication.

14 directly corresponds to the reduction axiom in [9] for the interaction between the dynamic modality and the safe belief modality. The conjunction on the right-hand side of the axiom encodes the action priority-update rule, cf. Definition 2.4.

For completeness, we first define $\mathcal{L}'^{-}$ as the language $\mathcal{L}'$ with the modalities E_i and $[a]$ removed, that is, the language with only the modalities K_i, $\vec{K}_i$ and $\Box_i$. We then define Λ as the smallest normal logic of $\mathcal{L}'^{-}$ containining only the axioms 1-5 above. We only need to show completeness of Λ since, using the other axioms, any formula of $\mathcal{L}'$ can be rewritten into an equivalent formula in $\mathcal{L}'^{-}$. To show completeness of Λ, take any Λ-consistent formula ϕ_0. We need to show that there exists a plausibility models $\mathcal{M}$ and a world w such that $(\mathcal{M}, w) \models \phi_0$. We construct $\mathcal{M}$ in two steps. First we build the canonical model $\mathcal{M}^\Lambda$, and next we create a finite filtration $\mathcal{M}^f$ of it. Only the filtrated model will satisfy all the requirements of being a plausibility model. The canonical model is $\mathcal{M}^\Lambda = (W^\Lambda, \sim_i^\Lambda, \geq_i^\Lambda, \approx_i, L^\Lambda)_{i \in Ag}$ where W^Λ is the set of all maximally Λ-consistent set of formulas, $\sim_i^\Lambda = \{(w, v) \mid \forall \phi \in \mathcal{L}'^{-} : K_i\phi \in w \Rightarrow \phi \in v\}$, $\geq_i^\Lambda = \{(w, v) \mid \forall \phi \in \mathcal{L}'^{-} : \Box_i\phi \in w \Rightarrow \phi \in v\}$, $\approx_i^\Lambda = \{(w, v) \mid \forall \phi \in \mathcal{L}'^{-} : \vec{K}_i\phi \in w \Rightarrow \phi \in v\}$, and $L^\Lambda(w) = \{p \mid p \in w\}$ [27, Def 4.18]. The semantics of the modalities are standardly defined in the canonical model. More precisely, let $\succ$ denote any of the symbols $\sim_i$, $\geq_i$ or $\approx_i$, where $i \in Ag$, and let $\Box_\succ$ denote the corresponding modality: When $\succ = \sim_i$ then $\Box_\succ = K_i$; when $\succ = \geq_i$ then $\Box_\succ = \Box_i$; and when $\succ = \approx_i$ then $\Box_\succ = \vec{K}_i$. Then for each choice of $\succ$, the semantic condition for the corresponding modality is: $(\mathcal{M}^\Lambda, w) \models \Box_\succ\phi$ iff for all v with $w \succ^\Lambda v$, $(\mathcal{M}, v) \models \phi$. Note that $\mathcal{M}^\Lambda$ has extra accessibility relations $\approx_i$ that plausibility models do not: these are the relations over which the $\vec{K}_i$ operators are interpreted in $\mathcal{M}^\Lambda$. We are going to rely on the following two properties of the canonical model $\mathcal{M}^\Lambda$, proved in [27]:

A. $\sim_i^\Lambda$ and $\approx_i^\Lambda$ are equivalence relations, and $\geq_i^\Lambda$ is a preorder (this is due to the canonicity of the S4 and S5 axioms).
B. For any formula ϕ and any world w of $\mathcal{M}^\Lambda$, $(\mathcal{M}^\Lambda, w) \models \phi$ iff $\phi \in w$ (this is the Truth Lemma [27, Lemma 4.21]).

We now show the following further properties:

C. $\geq_i^\Lambda \subseteq \approx_i^\Lambda$ and $\sim_i^\Lambda \subseteq \approx_i^\Lambda$.

For the first of these, suppose $w \geq_i^\Lambda v$ and $\vec{K}_i\phi \in w$. We need to prove that $\phi \in v$. From our fourth axiom (and the fact that w is a maximally consistent set), we get $\Box_i\phi \in w$. Since $w \geq_i^\Lambda v$, we then get immediately get $\phi \in v$, using the definition of $\geq_i^\Lambda$. This proves $\geq_i^\Lambda \subseteq \approx_i^\Lambda$. The other inclusion is proved symmetrically, using instead our third axiom.

Let now Σ be the minimal set of formulas satisfying the following conditions:

1. $\phi_0 \in \Sigma$.
2. *Closure under subformulas*: If $\phi \in \Sigma$ and ψ is a subformula of ϕ, then $\psi \in \Sigma$.
3. *Closure under single negations*: If $\phi \in \Sigma$ and ϕ is not on the form $\neg\psi$, then $\neg\phi \in \Sigma$.
4. If $\neg\Box_i\phi \in \Sigma$ and $\Box_i\psi \in \Sigma$, then $\vec{K}_i(\neg\Box_i\phi \vee \Box_i\psi) \in \Sigma$.

We now prove that Σ is finite. Let Σ' be the minimal set closed under conditions 1–3 only. Then Σ' is finite, since any formula in Σ' is either a subformula of ϕ_0

or the negation of such a subformula. When adding condition 4, for each choice of $\neg\Box_i\phi, \Box_i\psi \in \Sigma'$, we can at most produce the following new formulas that were not already in Σ': $\vec{K}_i(\neg\Box_i\phi \vee \Box_i\psi)$, $\neg\vec{K}_i(\neg\Box_i\phi \vee \Box_i\psi)$, $\neg\Box_i\phi \vee \Box_i\psi$ and $\neg(\neg\Box_i\phi \vee \Box_i\psi)$. However, none of these have the form required to make new applications of condition 4, and hence Σ must be finite.

We now define the filtrated model. For each $w \in W^\Lambda$, define $|w| = \{v \in W^\Lambda \mid \forall\phi \in \Sigma : \phi \in v \Leftrightarrow \phi \in w\}$. Now define a model $\mathcal{M}^f = (W^f, \sim_i^f, \geq_i^f, \approx_i^f, L^f)_{i\in Ag}$ by letting $W^f = \{|w| \mid w \in W^\Lambda\}$, $L^f(w) = \{p \in \Sigma \mid p \in w\}$, and

1. $|w| \approx_i^f |v|$ iff for all $\vec{K}_i\phi \in \Sigma$, $(\mathcal{M}^\Lambda, w) \models \vec{K}_i\phi \Leftrightarrow (\mathcal{M}^\Lambda, v) \models \vec{K}_i\phi$.
2. $|w| \sim_i^f |v|$ iff a) for all $K_i\phi \in \Sigma$, $(\mathcal{M}^\Lambda, w) \models K_i\phi \Leftrightarrow (\mathcal{M}^\Lambda, v) \models K_i\phi$; and b) for all $\vec{K}_i\phi \in \Sigma$, $(\mathcal{M}^\Lambda, w) \models \vec{K}_i\phi \Leftrightarrow (\mathcal{M}^\Lambda, v) \models \vec{K}_i\phi$.
3. $|w| \geq_i^f |v|$ iff a) for all $\Box_i\phi \in \Sigma$, $(\mathcal{M}^\Lambda, w) \models \Box_i\phi \Rightarrow (\mathcal{M}^\Lambda, v) \models \Box_i\phi$; and b) for all $\vec{K}_i\phi \in \Sigma$, $(\mathcal{M}^\Lambda, w) \models \vec{K}_i\phi \Leftrightarrow (\mathcal{M}^\Lambda, v) \models \vec{K}_i\phi$.

Clearly, by construction, $\sim_i^f$ and $\approx_i^f$ are equivalence relations, and $\geq_i^f$ is a preorder. We now want to prove that $\mathcal{M}^f$ is a filtration of $\mathcal{M}^\Lambda$ through Σ [27, Def 2.36]. This amounts to proving the following, where again $\succ$ is any of the symbols $\sim_i$, $\geq_i$ or $\approx_i$, and $\Box_\succ$ denotes the corresponding modality:

i. If $w \succ^\Lambda v$ then $|w| \succ^f |v|$.
ii. If $|w| \succ^f |v|$ then for all $\Box_\succ\phi \in \Sigma$, $(\mathcal{M}^\Lambda, w) \models \Box_\succ\phi$ implies $(\mathcal{M}^\Lambda, v) \models \phi$.

Case $\succ = \approx_i$. To prove condition i, let $w \approx_i^\Lambda v$. As we already showed $\approx_i^\Lambda$ to be an equivalence relation (property A above), w and v have the same $\approx_i^\Lambda$-successors. Hence for any formula ϕ, $(\mathcal{M}^\Lambda, w) \models \vec{K}_i\phi$ iff $(\mathcal{M}^\Lambda, v) \models \vec{K}_i\phi$, which proves $|w| \approx_i^f |v|$, as required. To prove condition ii, assume $|w| \approx_i^f |v|$, $\vec{K}_i\phi \in \Sigma$ and $(\mathcal{M}^\Lambda, w) \models \vec{K}_i\phi$. The definition of $\approx_i^f$ then gives us $(\mathcal{M}^\Lambda, v) \models \vec{K}_i\phi$. By reflexivity of $\approx_i^\Lambda$ (property A), we get $\mathcal{M}^\Lambda, v \models \phi$, as required.

Case $\succ = \sim_i^f$. To prove condition i, let $w \sim_i^\Lambda v$. We need to prove $|w| \sim_i^f |v|$, which amounts to proving a) and b) of item 2 above. We get a proof of a) by replacing $\approx_i$ by $\sim_i$ and $\vec{K}_i$ by K_i in the proof of condition i for $\approx_i$. For b), property C gives us $w \approx_i^\Lambda v$, and then we can directly reuse the proof of condition i for $\approx_i$. To prove condition ii, we only need to replace $\approx_i$ by $\sim_i$ and $\vec{K}_i$ by K_i in the proof of condition ii of $\approx_i$.

Case $\succ = \geq_i$. To prove condition i, let $w \geq_i^\Lambda v$. We need to prove $|w| \geq_i^f |v|$, which amounts to proving a) and b) of item 3 above. For a), suppose $\Box_i\phi \in \Sigma$ and $(\mathcal{M}^\Lambda, w) \models \Box_i\phi$. Since $w \geq_i^\Lambda v$ and $\geq_i^\Lambda$ is transitive (property A), any $\geq_i$-successor of v is also a $\geq_i$-successor of w and hence also $(\mathcal{M}^\Lambda, v) \models \Box_i\phi$, as required. For b), property C gives us $w \approx_i v$, and then we can again reuse the proof of condition i for $\approx_i$. For condition ii, we only need to replace $\approx_i$ by $\geq_i$ and $\vec{K}_i$ by $\Box_i$ in the proof of condition ii of $\approx_i$.

We have now proved that $\mathcal{M}^f$ is a filtration of $\mathcal{M}^\Lambda$ through Σ. By the Filtration Theorem [27, Theorem 2.39], we then have for all formulas $\phi \in \Sigma$ and all worlds $w \in W^\Lambda$, $(\mathcal{M}^\Lambda, w) \models \phi$ iff $(\mathcal{M}^f, |w|) \models \phi$. It is also clear that $\mathcal{M}^f$ is finite, since Σ is. We now prove the following additional properties of $\mathcal{M}^f$:

D. $\leqslant_i^f = \approx_i^f$.

E. $\leq_i^f$ is a union of mutually disjoint well-preorders.

Proof of D. Since the condition defining $\approx_i^f$ (item 1 above) is condition b) in the definition of $\geq_i^f$ (item 3 above), we get $\geq_i^f \subseteq \approx_i^f$. Since $\approx_i^f$ is symmetric, we also get $\leq_i^f \subseteq \approx_i^f$, and hence $\leqslant_i^f \subseteq \approx_i^f$. To prove that also $\approx_i^f \subseteq \leqslant_i^f$, suppose to achieve a contradiction that $|w| \approx_i^f |v|$, $|w| \not\geqslant_i^f |v|$ and $|v| \not\geqslant_i^f |w|$. From $|w| \approx_i^f |v|$ we get that condition b) of the definition of both $|w| \geq_i^f |v|$ and $|v| \geq_i^f |w|$ is satisfied. Hence, it must be condition a) that fails. Thus there exists formulas $\Box_i\phi, \Box_i\psi \in \Sigma$ such that $(\mathcal{M}^f, |v|) \models \Box_i\phi$, $(\mathcal{M}^f, |w|) \models \neg\Box_i\phi$, $(\mathcal{M}^f, |w|) \models \Box_i\psi$, and $(\mathcal{M}^f, |v|) \models \neg\Box_i\psi$. Using the Filtration Theorem, we get $(\mathcal{M}^\Lambda, w) \models \neg\Box_i\phi$ and $(\mathcal{M}^\Lambda, w) \models \Box_i\psi$, and hence $(\mathcal{M}^\Lambda, w) \models \neg\Box_i\phi \wedge \Box_i\psi$. By our fifth axiom, we then get $(\mathcal{M}^\Lambda, w) \models \vec{K}_i(\neg\Box_i\phi \vee \Box_i\psi)$ and hence $(\mathcal{M}^f, |w|) \models \vec{K}_i(\neg\Box_i\phi \vee \Box_i\psi)$, using the Filtration Theorem and the fact that $\vec{K}_i(\neg\Box_i\phi \vee \Box_i\psi) \in \Sigma$ by closure condition 4 in the definition of Σ. Since $|w| \approx_i^f |v|$, we then get $(\mathcal{M}^f, |v|) \models \neg\Box_i\phi \vee \Box\psi$. This contradicts that $(\mathcal{M}^f, |v|) \models \Box_i\phi$ and $(\mathcal{M}^f, |v|) \models \neg\Box_i\psi$, and we are done.

Proof of E. We already know that $\geq_i^f$, and hence $\leq_i^f$, is a preorder. The restriction of $\leq_i^f$ to any subset of W^f is hence also a preorder. We now prove that the restriction of $\leq_i^f$ to each $\approx_i^f$-equivalence class is a well-preorder. We need to prove that if X is a subset of an $\approx_i^f$-equivalence class, then X has smallest elements with respect to $\leq_i^f$. As $\mathcal{M}^f$ is a finite model, X is also finite, and hence, since $<_i^f$ is transitive and irreflexive, there must exist an element $x \in X$ such that there exists no $y \in X$ with $y <_i^f x$ (i.e., $<_i^f$ is well-founded). Since, by D, $\leqslant_i^f = \approx_i^f$, any two elements of X are comparable by $\leq_i$, and we must then have $x \leq_i^f y$ for all $y \in X$. This proves that x is a least element of X, and we are done. We now know that the restriction of $\leq_i^f$ to each $\approx_i^f$-equivalence class is a well-preorder. These well-preorders are also disjoint, since by D, any two elements not in the same $\approx_i^f$-equivalence class are also not comparable by $\leq_i$.

The final step is to turn $\mathcal{M}^f$ into a plausibility model $\mathcal{M}$. We simply define $\mathcal{M}$ by $\mathcal{M} = (W^f, \sim_i^f, \leq_i^f, L^f)_{i \in Ag}$, that is, we reverse the $\geq_i^f$ relation of $\mathcal{M}^f$ to be consistent with the direction used in the definition of plausibility models, and we omit the $\approx_i$ relation. We now prove that for all $w \in W^f$ and all formulas $\phi \in \mathcal{L}'^-$, $(\mathcal{M}, w) \models \phi$ iff $(\mathcal{M}^f, w) \models \phi$. This is a trivial induction proof for all cases except when ϕ is of the form $\vec{K}_i\psi$, as for all the other modalities, we have the same relations in the two models. For the case of $\phi = \vec{K}_i\psi$, note that in $\mathcal{M}^f$ the semantics of $\vec{K}_i$ is defined in terms of $\approx_i$, whereas in $\mathcal{M}$ it is defined in terms of $\leqslant_i^f$ (we originally defined the semantics of $\vec{K}_i\phi$ in plausibility models $\mathcal{M} = (W, \sim_i, \leq_i, L)_{i \in Ag}$ by letting $(\mathcal{M}, w) \models \vec{K}_i\phi$ whenever for all $v \leqslant_i w$, $(\mathcal{M}, v) \models \phi$). However, by property D, we have $\leqslant_i^f = \approx_i^f$, and hence this case also goes trivially through.

To prove that $\mathcal{M}$ is a plausibility model, we need to prove that $\leq_i^f$ is reflexive, which we know, and that it is a union of mutually disjoint well-preorders, which we also know, as this is property E. We also need to prove that $\sim_i^f$ is an equivalence relation, something we already proved, and that $\sim_i^f \subseteq \leqslant_i^f$. For the latter, note that as condition b) of the definition of $\sim_i^f$ (item 3 above) is identical to the condition defining $\approx_i^f$ (item 1 above), we get $\sim_i^f \subseteq \approx_i^f$. Then by property D, we get $\sim_i^f \subseteq \leqslant_i^f$, as required.

These properties together show that $\mathcal{M}$ is a plausibility model (Definition 2.1). It only remains to show that ϕ_0 is true in a world of $\mathcal{M}$. Since ϕ_0 was chosen as a Λ-consistent formula, there exists $w_0 \in W^\Lambda$ with $\phi_0 \in w$, from which we can conclude $(\mathcal{M}^\Lambda, w_0) \models \phi_0$ (Truth Lemma) and hence $(\mathcal{M}^f, |w_0|) \models \phi_0$ (Filtration Theorem). By the argument of the previous paragraph, we then also have $(\mathcal{M}, |w_0|) \models \phi_0$, and the proof is finally complete. □

Note the role played by the filtration in the proof above: The finiteness of the filtrated model was required to ensure well-foundedness of $<_i$, which in turn guaranteed that $\leq_i$ became a union of mutually disjoint well-preorders.

7 A Comprehensive Example

To illustrate the strengths of the logical framework introduced in this work, we present a comprehensive example. In Figure 6 (page 42) we show the state development of the example as a graph. The setting is as follows. Hanna is at home and will in the upcoming future either go hiking or stay at home and read her tablet. Her robot Rob can observe the current state and knows that the future will be either *hiking* or *home*. In the initial state, s_0, the tablet is charged and the raincoat is in the backpack. Hanna has a false belief that the backpack does not have a hole. It is undesirable for Hanna to become wet in the rain, that is, $U_{hike} = \mathrm{Rain} \wedge \neg\mathrm{Wear}(h, \mathrm{Raic}) \wedge \mathrm{On}(h, \mathrm{Mount})$, which, for convenience, also makes the Wet atom true. However, we have that $s_{40} \models U_{hike}$. Informing Hanna about the backpack's hole is a relevant announcement because of her expectation of future desirability: she will not have the raincoat on the mountain and will be wet, $s_0 \models \neg U_{hike} \wedge [\mathrm{Drop}(\mathrm{Raic}, \mathrm{Bckp}); \mathrm{Take}(\mathrm{Bckp}); \mathrm{GoTo}(\mathrm{Mount}); \mathrm{Try\_PutOn}(\mathrm{Raic})](\mathrm{Wet} \wedge E_h \neg \mathrm{Wet}) \wedge [\Uparrow(h, \mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})); \mathrm{Drop}(\mathrm{Raic}, \mathrm{Bckp}); \mathrm{Take}(\mathrm{Bckp}); \mathrm{GoTo}(\mathrm{Mount}); \mathrm{Try\_PutOn}(\mathrm{Raic})](\mathrm{Wet} \to E_h \mathrm{Wet})$. The announcement $\mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})$ in the expression above is only relevant for one possible future: *hiking*—but not the other future: *home*—hence, it is not *timely* to announce it. The state is advanced by product update, $s_1 = s_0 \otimes a_1$, the raincoat drops out of the backpack through its hole which Hanna does not observe. She now has a false belief that the raincoat is in the backpack. For the minimal relevant announcement we have again $\mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})$, now, in s_1, after the action sequence $\mathrm{Take}(\mathrm{Bckp}); \mathrm{GoTo}(\mathrm{Mount}); \mathrm{Try\_PutOn}(\mathrm{Raic})$.

Note that multiple minimal announcements can exist. A hypothetical example is $\mathrm{In}(\mathrm{Raic}, \mathrm{Hom})$, the raincoat is in the home, if it would be part of the model. The negative of the first alternative, $\neg\,\mathrm{In}(\mathrm{Raic}, \mathrm{Bckp})$, is relevant as well, but it is not minimal, the formula is longer[9]. Announcing $\mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})$ is relevant and minimal only for one possible future (*hiking*) but not the other (*home*), hence, it is not timely. Let's suppose a state development such that $s_{20} = s_1 \otimes a_{20}$: Hanna takes her backpack. She still has her false beliefs about the backpack not having a hole and the raincoat being in the backpack, such that, in s_{20} it holds that her expectation can be corrected by the relevant announcement $\mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})$ given the action sequence $\mathrm{GoTo}(\mathrm{Mount}); \mathrm{Try\_PutOn}(\mathrm{Raic})$. The announcement seems to be the most timely in s_{20}, as we now know Hanna will go hiking, hence, the

[9] Said explicitly, Hanna can infer from $\mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})$ that $\neg\,\mathrm{In}(\mathrm{Raic}, \mathrm{Bckp})$ resp. $\mathrm{In}(\mathrm{Raic}, \mathrm{Hom})$ holds.

announcement is relevant for all possible futures, and the extra effort for Hanna is low: from $\mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})$ she infers $\neg\,\mathrm{In}(\mathrm{Raic}, \mathrm{Bckp})$ which is remedied by repairing the backpack's hole and putting the raincoat back inside. Suppose instead we make the announcement in state s_{30}: she is on the mountain after having traveled there. We have $s_{30} \models \neg\mathrm{Wet} \wedge [\mathrm{Try\_PutOn}(h, \mathrm{Raic})](\mathrm{Wet} \wedge E_h\neg\mathrm{Wet}) \wedge [\Uparrow(h, \mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})); \mathrm{Try\_PutOn}(h, \mathrm{Raic})](\mathrm{Wet} \rightarrow E_h\mathrm{Wet})$. The announcement is the same as in s_{20} and relevant for all futures (*hiking*). However, the extra effort now is higher: Hanna needs to do all the steps as when announcing in s_{20}, but additionally she first needs to travel home from the mountain. In state $s_{40} = s_{30} \otimes a_{40}$[10], Hanna tries to put the raincoat on, which fails: there is no raincoat in her backpack. In s_{40} there is no relevant announcement: we do not have to change Hanna's expectation because she already truly believes (knows) Wet. As $s_{40} \models B_h\mathrm{Wet} \wedge E_h\mathrm{Wet}$, belief and expectation coincide. Because we no longer can make a relevant announcement, we can consider s_{40} being *too late*, thus, not timely to make an announcement. Note that depending on *when* we make the relevant announcement, its content might change, for example, in s_0, the only possible minimal relevant announcement is $\mathrm{Has}(\mathrm{Hol}, \mathrm{Bckp})$, whereas in s_1, we have the alternative minimal relevant announcement $\mathrm{In}(\mathrm{Raic}, \mathrm{Hom})$.

Now consider that Hanna stays at home instead of going hiking, i.e., Hanna goes to the sofa, $s_{21} = s_1 \otimes a_{21}$. Here, we have the undesirability formula $U_{home} = \mathrm{At}(h, \mathrm{Sofa}) \wedge \neg\mathrm{Charg}(\mathrm{Tab})$. However, we have that $s_{31} \models U_{home}$. Announcing that the backpack has a hole, is not relevant here. Although the future will be undesirable (in s_{31} her tablet battery is depleted), we cannot compute a relevant announcement to change her expectation in s_{21}. Why? If we try to compute a relevant announcement by $s_{21} \models \neg U \wedge [\mathrm{Read}(\mathrm{Tab})]\big((U \wedge E_h\neg U) \wedge [\Uparrow(h, \neg\mathrm{Charg}(\mathrm{Tab})); \mathrm{Read}(\mathrm{Tab})]((U \rightarrow E_hU) \wedge B_hU)\big)$, we can see that we can't do anything any more in s_{21} to avoid future undesirability, that is, s_{31} is undesirable and Hanna's expectation and belief (even knowledge) coincide. Thus, we would like to make an announcement to correct her expectation *directly*, i.e., saying what will be wrong with her future expectation. This, however, is unexplored in this work (see discussion in Section 9). Lastly, in s_{31} the tablet is uncharged. Announcing this is not relevant as the state is already undesirable.

8 Background and Related Work

Theory of Mind (ToM), ascribing mental states (such as beliefs or expectations as we do here) to oneself and others, has earlier been investigated in psychology [29, 30], and later in AI and logic [31, 32]. A way to model ToM is using the seminal work on *Epistemic Logic (EL)* by Hintikka [33]. EL is laying the ground for its dynamic version, *Dynamic Epistemic Logic (DEL)* [1] which in turn is the basis of the *plausibility (event) models* [3] used in this work. Reasoning about false beliefs using DEL, as we do here, has been formalized by Bolander [20]. There are previous approaches that address modeling a posteriori and a priori beliefs, however as explained in Section 3, none of them seems to suffice for our purposes. Baltag and Smets [9] acknowledge the need of addressing the problem, and present an operator of *appearance* to model how

[10] For conciseness, we assume that it is starting to rain at the same time.

an action appears to some agent *while* it is happening, defined in terms of the $\sim_i$ relation, capturing a posteriori indistinguishability. Importantly, in our work we address preventing undesirable outcomes, so we want to be able to reason about expected future actions and states *before* they happen, captured by the $\lesssim_i$ relation. Pieper [11] also recently introduced an operator for modeling the a priori perspectives on actions in the setting of plausibility models. The approach there is slightly different from ours. Instead of defining the perspective on an action by pointing out the most plausible events of that action, Pieper [11] defines a completely new *plausible product update* by $s \otimes_i a = (\mathcal{M} \otimes \mathcal{E}, \{(w, e) \in W_d \times E_d \mid e \in \text{Min}_{\leq_i}\{f \in E_d \mid (\mathcal{M}, w) \models pre(f)\}\})$. However, their approach is not defining a perspective on the actions, since the perspective is only taken into account in the product update and it depends on the state in which it is applied. Concepts parallel to ours can also be found in epistemic planning: Andersen et al. [10] present a framework based on plausibility models to compute plans and distinguish between *runtime* (a posteriori) and *plantime* (a priori) knowledge/belief, but only in a single-agent setting. In our work, we model state transitions from current to future states by plausibility models and a sequence of product updates of plausibility event models (action sequence). An alternative to model state transitions for epistemic reasoning may be Li and Wang [34] who use *Epistemic Transition Systems* $\mathcal{M}$ based on which an *Execution Tree* $\mathcal{T}$ is defined. However, this seems to be a more coarse approach with atomic states and actions, which would not suffice in our work.

Our framework is aimed to be applied in human-AI systems such that the artificial agent can *proactively* intervene to revise a false belief of the human and in that way help to avoid undesirable future outcomes. Humans themselves are known to have proactive behavior [35]. Therefore it can naturally facilitate the collaboration with humans if AI systems are proactive too [36–38]. Recent work has been presented to address AI system proactivity based on reasoning about humans' false beliefs. Favier et al. [6] enable a robot to proactively intervene when the human has a false belief which will jeopardize achieving the task goal. They use a simpler method to do this than DEL, with state variables modeling beliefs. However, DEL is more expressive and, unlike their formalism, allows to express agents' uncertainty. Proactive robot assistance in Shvo et al. [7] is done by reasoning about robot and human beliefs encoded in modal KD45 logic and by using an epistemic planner [39]. They compute discrepancies between the human's and the optimal plan to reach a goal and resolve them by making relevant announcements. As opposed to Shvo et al. [7], in our approach we choose not to do plan and goal recognition because it's not our foremost aim to help the human achieve their goal—and also since plan and goal recognition comes with its own set of challenges. Instead, we focus on preventing humans from ending up in undesirable states. This can mean keeping the human in desirable states from a more "overall" point of view instead of supporting her achieving her goal. For example, Hanna's goal is to eat a whole cake, but she is diabetic so overall this is undesirable and the robot should not help with achieving this goal[11]. Spinning this further we can take into account many different preference functions instead of only one single human's: Hanna's preference of eating the cake, her physician's preference of keeping her blood sugar levels right, her husband's preference of keeping Hanna happy but also healthy;

[11] For a discussion on disobedient robots, see for example [40].

etc. This allows us to be more flexible, and makes us, in principle, able to consider multiple sources of preferences instead of only one. Zhang and Williams [8] extend Baltag and Smets [3]'s work by using knowledge bases. They introduce plausibility models with knowledge bases and conditional doxastic logic [9] for knowledge bases. As a result, they can model false beliefs and revise them, referring directly to the plan space instead of states. To remedy human false beliefs about feasible plans, the robot can adapt to the human's actions or make announcements.

9 Discussion and Future Directions

We have presented a new theory that can enable agents to proactively intervene by reasoning on the human's false beliefs and expectations about future states. We use plausibility models and introduce a new logic including our new *expectation* operator, based on which we define relevant announcements to prevent undesirable outcomes. Instead of *hard* announcements which 'cut links to worlds', we have *soft* announcements which make previously most plausible worlds less plausible. This is suitable for expectations which are a priori: what is most plausible can still change and is not definite. We axiomatize our logic and prove soundness and completeness. We formally show that deciding whether a relevant announcement exists and computing one can be done in polynomial time in the size of the updated model. By an elaborate example we demonstrate that our formalism has the potential to be applicable to real human-robot systems. Next steps include a deeper investigation of our introduced logic (bisimulation, complexity, etc.), as well as further exploring non-propositional undesirability formulas and non-propositional relevant announcements. We also plan to implement the framework in one of our existing robotic settings, for example the one used in work by Dissing and co-authors [41, 42] that already implement a DEL-based reasoner in a robot for human-robot interaction applications.

We have assumed the future action sequence given. In future work we might employ learning or plan and goal recognition techniques [43] to predict future action sequences. We might relate relevance and goals and achieve stronger notions of relevant announcements. However, goal recognition is not always realistic to do and complications are foreseeable (too many possible goals, multiple goals pursued at once, dynamic or conflicting goals, etc.). It is conceivable to have multiple arguments taken into account of *when* to make a relevant announcement. Although Hanna has a false expectation in s_{21} about a future state s_{31} being desirable, we cannot compute a relevant announcement. Although undesirability is still in the future, not in the present state, it is too late to avoid undesirability. Thus, announcing in s_{21} that the action of reading the tablet leads to a state where the tablet is uncharged, hence, where E_hU is true but also B_hU and K_hU are true, hence, it does not help Hanna avoiding undesirability. We want to investigate in future work another type of relevant announcement that can catch such cases. One might be to make the announcement "as early as possible", as soon as we find an announcement that is relevant for one possible future. Another option is to wait until the announcement is relevant for all possible futures. Yet another factor to take into account is whether the human still has a "repair plan" to prevent undesirable outcomes, or even better, we might want to find the time to make the relevant

announcement so that this repair plan is optimal according to some measure. Finally, one might want to accept a less optimal repair plan in favor of considering that there might be contexts where the human should not be disturbed.

Definition 5.1 allows us to say that an undesirability formula U can be true and hence be the source of why a state is undesirable. In their work on proactive robots, Grosinger et al. [21] show that reasoning about fuzzy as opposed to binary desirability allows for a more realistic world model. Similarly, we could attach degrees to a formula U, creating priorities on relevant announcements.

References

[1] Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements and common knowledge and private suspicions. In: Gilboa, I. (ed.) Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-98), pp. 43–56. Morgan Kaufmann, USA (1998)

[2] Ditmarsch, H., Hoek, W., Kooi, B.: Dynamic Epistemic Logic vol. 337. Springer, Netherlands (2007)

[3] Baltag, A., Smets, S.: Dynamic belief revision over multi-agent plausibility models. In: G. Bonanno, W. van der Hoek, M. Woolridge (eds.), Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision (LOFT 2006), pp. 11–24 (2006)

[4] Grosinger, J.: On proactive human-AI systems. In: AIC Workshop on AI and Cognition (2022)

[5] Grant, A.M., Ashford, S.J.: The dynamics of proactivity at work. Research in Organizational Behavior **28**, 3–34 (2008)

[6] Favier, A., Shekhar, S., Alami, R.: Anticipating false beliefs and planning pertinent reactions in human-aware task planning with models of theory of mind. In: International Conference on Automated Planning and Scheduling (ICAPS) 2023 (2023)

[7] Shvo, M., Hari, R., O'Reilly, Z., Abolore, S., Wang, S.-Y.N., McIlraith, S.A.: Proactive robotic assistance via theory of mind. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 9148–9155 (2022). https://doi.org/10.1109/IROS47612.2022.9981627

[8] Zhang, Y., Williams, B.: Adaptation and communication in human-robot teaming to handle discrepancies in agents' beliefs about plans. Proceedings of the International Conference on Automated Planning and Scheduling **33**(1), 462–471 (2023) https://doi.org/10.1609/icaps.v33i1.27226

[9] Baltag, A., Smets, S.: A qualitative theory of dynamic interactive belief revision. In: Bonanno, G., Hoek, W., Wooldridge, M. (eds.) Logic and the Foundations

of Game and Decision Theory (LOFT7). Texts in Logic and Games, vol. 3, pp. 13–60. Amsterdam University Press, Netherlands (2008)

[10] Andersen, M.B., Bolander, T., Jensen, M.H.: Don't plan for the unexpected: Planning based on plausibility models. Logique et Analyse **58(230)**, 145–176 (2015) https://doi.org/10.2143/LEA.230.0.3141807

[11] Pieper, J.: Plausibility planning for simplified implicit coordination. Master's thesis, University of Freiburg (2023). https://doi.org/10.6094/UNIFR/258786

[12] Bacchus, F., Petrick, R.: Modeling an agent's incomplete knowledge during planning and during execution. In: Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR 1998), pp. 432–443 (1998)

[13] Bolander, T., Andersen, M.B.: Epistemic planning for single- and multi-agent systems. Journal of Applied Non-Classical Logics **21**, 9–34 (2011) https://doi.org/10.3166/jancl.21.9-34

[14] Ditmarsch, H., Kooi, B.: Semantic results for ontic and epistemic change. In: Bonanno, G., Hoek, W., Wooldridge, M. (eds.) Logic and the Foundation of Game and Decision Theory (LOFT 7). Texts in Logic and Games 3, pp. 87–117. Amsterdam University Press, Netherlands (2008)

[15] Engesser, T., Bolander, T., Mattmüller, R., Nebel, B.: Cooperative epistemic multi-agent planning for implicit coordination. In: Proceedings of Methods for Modalities. Electronic Proceedings in Theoretical Computer Science, pp. 75–90 (2017)

[16] Li, Y., Yu, Q., Wang, Y.: More for free: a dynamic epistemic framework for conformant planning over transition systems. Journal of Logic and Computation **27**(8), 2383–2410 (2017)

[17] Bolander, T., Engesser, T., Herzig, A., Mattmüller, R., Nebel, B.: The dynamic logic of policies and contingent planning. In: European Conference on Logics in Artificial Intelligence (JELIA). Lecture Notes in Computer Science, vol. 11468, pp. 659–674 (2019). Springer

[18] Bolander, T.: A gentle introduction to epistemic planning: The DEL approach. Electronic Proceedings in Theoretical Computer Science **243**, 1–22 (2017)

[19] Benthem, J., Smets, S.: Chapter 7. In: Ditmarsch, H., Halpern, J.Y., Hoek, W., Kooi, B. (eds.) Dynamic logics of belief change, pp. 313–393. College Publications, UK (2015)

[20] Bolander, T.: Seeing Is Believing: Formalising False-Belief Tasks in Dynamic Epistemic Logic, pp. 207–236. Springer, Cham (2018). https://doi.org/10.1007/

978-3-319-62864-6_8

[21] Grosinger, J., Pecora, F., Saffiotti, A.: Robots that maintain equilibrium: Proactivity by reasoning about user intentions and preferences. Pattern Recognition Letters **118**, 85–93 (2019) https://doi.org/10.1016/j.patrec.2018.05.014

[22] Pol, I., Rooij, I., Szymanik, J.: Parameterized complexity of Theory of Mind reasoning in dynamic epistemic logic. Journal of Logic, Language, and Information **27**, 255–294 (2018) https://doi.org/10.1007/s10849-018-9268-4

[23] Bolander, T., Lequen, A.: Parameterized complexity of dynamic belief updates: A complete map. Journal of Logic and Computation **33**(6), 1270–1300 (2023)

[24] van Ditmarsch, H., van der Hoek, W., Halpern, J., Kooi, B. (eds.): Handbook of Epistemic Logic. College Publications, UK (2015)

[25] Umans, C.: The minimum equivalent dnf problem and shortest implicants. Journal of Computer and System Sciences **63**(4), 597–611 (2001)

[26] Barras, C.: Useful, lovable and unbelievably annoying. New Scientist **204**(2738), 22–23 (2009)

[27] Blackburn, P., Rijke, M., Venema, Y.: Modal Logic. Cambridge Tracts in Theoretical Computer Science, vol. 53. Cambridge University Press, Cambridge, UK (2001). https://doi.org/10.1017/CBO9781107050884

[28] Wáng, Y.N., Ågotnes, T.: Simpler completeness proofs for modal logics with intersection. In: International Workshop on Dynamic Logic, pp. 259–276 (2020). Springer

[29] Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1**(4), 515–526 (1978)

[30] Baron-Cohen, S., Leslie, A.M., Frith, U.: Does the autistic child have a "theory of mind" ? Cognition **21**(1), 37–46 (1985) https://doi.org/10.1016/0010-0277(85)90022-8

[31] Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.: Reasoning About Knowledge. MIT press, USA (1995)

[32] Ditmarsch, H., Labuschagne, W.: My beliefs about your beliefs: a case study in theory of mind and epistemic logic. Synthese **155**, 191–209 (2007)

[33] Hintikka, K.J.J.: Knowledge and Belief: An Introduction to the Logic of the Two Notions. Cornell University Press, Ithaca, NY, USA (1962)

[34] Li, Y., Wang, Y.: Knowing how to plan about planning: Higher-order and meta-level epistemic planning. Artificial Intelligence **337**, 104233 (2024) https://doi.

org/10.1016/j.artint.2024.104233

[35] Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences **28**, 675–735 (2005)

[36] Harman, H., Simoens, P.: Action graphs for proactive robot assistance in smart environments. Journal of Ambient Intelligence and Smart Environments **12**(2), 1–21 (2020)

[37] Baraglia, J., Cakmak, M., Nagai, Y., Rao, R.P., Asada, M.: Efficient human-robot collaboration: when should a robot take initiative? International Journal of Robotics Research **36**(5-7), 563–579 (2017)

[38] Kraus, M., Schiller, M., Behnke, G., Bercher, P., Dorna, M., Dambier, M., Glimm, B., Biundo, S., Minker, W.: "was that successful?" on integrating proactive meta-dialogue in a DIY-assistant using multimodal cues. In: Proceedings of the 2020 International Conference on Multimodal Interaction. ICMI '20, pp. 585–594. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3382507.3418818

[39] Muise, C., Belle, V., Felli, P., McIlraith, S., Miller, T., Pearce, A.R., Sonenberg, L.: Efficient multi-agent epistemic planning: Teaching planners about nested belief. Artificial Intelligence **302**, 103605 (2022) https://doi.org/10.1016/j.artint.2021.103605

[40] Arnold, T., Briggs, G., Scheutz, M.: Only those who can obey can disobey: The intentional implications of artificial agent disobedience. In: International Conference on Autonomous Agents and Multiagent Systems, pp. 130–143 (2022). Springer

[41] Dissing, L., Bolander, T.: Implementing theory of mind on a robot using dynamic epistemic logic. In: IJCAI, pp. 1615–1621 (2020)

[42] Bolander, T., Dissing, L., Herrmann, N.: DEL-based epistemic planning for human-robot collaboration: Theory and implementation. In: Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning (KR 2021) (2021)

[43] Mirsky, R., Keren, S., Geib, C.: Introduction to symbolic plan and goal recognition. Synthesis Lectures on Artificial Intelligence and Machine Learning **16**(1), 1–190 (2021)

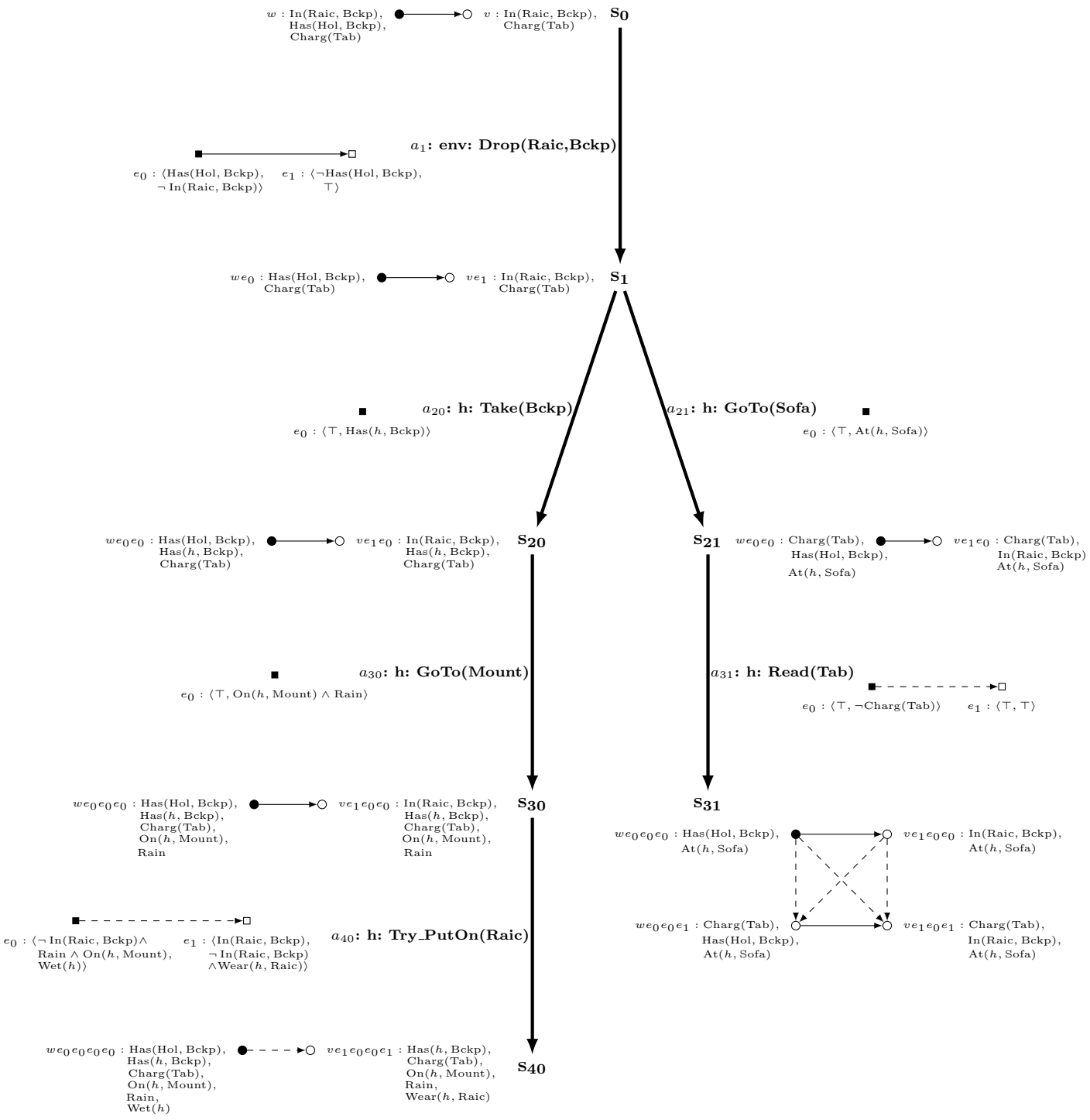

Fig. 6 Comprehensive example (Section 7): state development