Bisimulation for single-agent plausibility models

Mikkel Birkegaard Andersen[1], Thomas Bolander[1], Hans van Ditmarsch[2] and Martin Holm Jensen[1]

[1] DTU Compute, Technical University of Denmark
{mibi,tobo,mhje}@dtu.dk
[2] LORIA, CNRS, Université de Lorraine
hans.van-ditmarsch@loria.fr

Abstract. Epistemic plausibility models are Kripke models agents use to reason about the knowledge and beliefs of themselves and each other. Restricting ourselves to the single-agent case, we determine when such models are indistinguishable in the logical language containing conditional belief, i.e., we define a proper notion of bisimulation, and prove that bisimulation corresponds to logical equivalence on image-finite models. We relate our results to other epistemic notions, such as safe belief and degrees of belief. Our results imply that there are only finitely many non-bisimilar single-agent epistemic plausibility models on a finite set of propositions. This gives decidability for single-agent epistemic plausibility planning.

1 Introduction

A typical approach in belief revision involves preferential orders to express degrees of belief and knowledge [10, 13]. This goes back to the 'systems of spheres' in [11, 9]. Dynamic doxastic logic was proposed and investigated in [14] in order to provide a link between the (non-modal logical) belief revision and modal logics with explicit knowledge and belief operators. A similar approach was pursued in belief revision in dynamic epistemic logic [3, 19, 17, 5, 20], that continues to develop strongly [7, 18]. We focus on the proper notion of structural equivalence on (static) models encoding knowledge and belief simultaneously. A prior investigation into that is [8], which we relate our results to at the end of the paper. Our motivation is to find suitable structural notions to reduce the complexity of planning problems. Such plans are sequences of actions, such as iterated belief revision. It is the dynamics of knowledge and belief that, after all, motivates our research.

The semantics of belief depend on the structural properties of models. To relate the structural properties of models to a logical language we need a notion of structural similarity, known as bisimulation. A bisimulation relation relates a modal operator to an accessibility relation. Epistemic plausibility models do not have an accessibility relation as such but a plausibility relation. This induces a set of accessibility relations: the *most plausible* states are the *accessible* states for the modal belief operator; and the *plausible* states are the *accessible* states for

the modal knowledge operator. But it contains much more information: to each modal operator of conditional belief (or of degree of belief) one can associate a possibly distinct accessibility relation. This begs the question how one should represent the bisimulation conditions succinctly. Can this be done by reference to the plausibility relation directly, instead of by reference to these, possibly many, induced accessibility relations? It is now rather interesting to observe that relative to the modal operations of knowledge and belief the plausibility relation is already in some way too rich.

Example 1. The (single-agent) epistemic plausibility model on the left in Figure 1 consists of three worlds w_1, w_2, and w_3. p is only false in w_2, and $w_1 < w_2 < w_3$[1]: the agent finds it most plausible that p is true, less plausible that p is false, and even less plausible that p is true. As p is true in the most plausible world, the agent believes p. If we go to slightly less plausible, the agent is already uncertain about the value of p, she only knows trivialities such as $p \vee \neg p$. The world w_3 does not make the agent even more uncertain. We therefore can discard that other world where p is true. This is the model in the middle in Figure 1. It is bisimilar to the model on the left! Therefore, and that is the important observation: *having one world more or less plausible than another world in a plausibility model does not mean that in any model with the same logical content we should find a matching pair of worlds.* This is evidenced in the figure: on the left w_3 is less plausible than w_2, but in the middle no world is less plausible than v_2; there is no match.

Now consider retaining w_3 and making it as plausible state as w_1. This gives the plausibility model on the right in Figure 1, where u_1 and u_3 are equiplausible (equally plausible), written $u_1 \simeq u_3$. This model is bisimilar to both the left and the middle model. But the right and middle one share the property that more or less plausible in one, is more or less plausible in the other: now there is a match. This makes for another important observation: *we can reshuffle the plausibilities such that models with the same logical content preserve the plausibility order.*

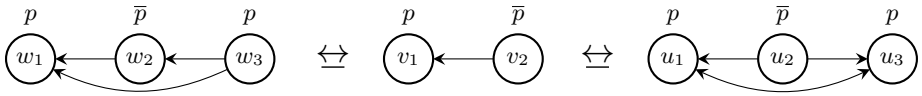

Fig. 1: All three models are bisimilar. The models in the middle and on the right are normal, the model on the left is not normal. An arrow $w_1 \leftarrow w_2$ corresponds to $w_1 \leq w_2$. Reflexive edges are omitted. $\overline{p}$ means that p does not hold.

In Section 2 we define the epistemic doxastic logic, the epistemic plausibility models on which it is interpreted, the suitable notion of bisimulation, and demonstrate the adequacy of this notion via a correspondence between modal equivalence and bisimilarity. The final sections 3, 4, and 5 respectively translate our results to degrees of belief and safe belief, discuss the problematic generalization to the multi-agent case, and demonstrate the relevance of our results for epistemic planning.

[1] If $s < t$, we have $s \leq t$ and $t \not\leq s$.

2 Single-agent plausibility models and bisimulation

2.1 Language, structures, and semantics

Definition 1 (Epistemic doxastic language). *For any countable set of propositional symbols P, we define the epistemic-doxastic language $\mathcal{L}_P$ by:*

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

where $p \in P$, K is the epistemic modality (knowledge) and B^{φ} the conditional doxastic modality (conditional belief). We use the usual abbreviations for the other boolean connectives as well as for $\top$ and $\bot$, and the abbreviation B for $B^{\top}$. The dual of K is denoted $\widehat{K}$, and the dual of B^{φ} is denoted $\widehat{B}^{\varphi}$.

We consider epistemic plausibility models as in [5]. A *well-preorder* on a set S is a reflexive and transitive relation $\leq$ on S such that every non-empty subset has minimal elements. The set of *minimal elements* of a subset T of S is given by:

$$Min_{\leq}T = \{s \in T \mid s \leq s' \text{ for all } s' \in T\}.$$

This is a non-standard notion of minimality, taken from [5]. Usually a minimal element of a set is an element that is not greater than any other element. On total preorders the two notions of minimality coincide. In fact, using the definition of minimality above, any well-preorder is total: For any pair of worlds s, t, $Min_{\leq}\{s, t\}$ is non-empty, and therefore $s \leq t$ or $t \leq s$.[2] These well-preorders are the *plausibility relations* (or *plausibility orderings*), expressing that a world is considered at least as plausible as another. This encodes the doxastic content of a model.

We can define such epistemic plausibility models with the plausibility relation as a primitive and with the epistemic relation as a derived notion. Alternatively, we can assume both as primitive relations, but require that more plausible means (epistemically) possible. We chose the latter.

Definition 2 (Epistemic plausibility model). *An* epistemic plausibility model *(or simply* plausibility model*) on a set of propositional symbols P is a tuple $\mathcal{M} = (W, \leq, \sim, V)$, where*

- *W is a set of* worlds*, called the* domain*.*
- *$\leq$ is a well-preorder on W, called the* plausibility relation*.*
- *$\sim$ is an equivalence relation on W called the* epistemic relation*. We require, for all $w, v \in W$, that $w \leq v$ implies $w \sim v$.*
- *$V : W \to 2^P$ is a* valuation*.*

For $w \in W$ we name $(\mathcal{M}, w)$ a pointed epistemic plausibility model*, and refer to w as the* actual world *of $(\mathcal{M}, w)$.*

[2] A well-preorder is not the same as a well-founded preorder; e.g., $y \leq x$, $z \leq x$ is a well-founded preorder, but not a well-preorder, as z and y are incomparable. Well-founded preorders are not necessarily total.

As we require that $\leq$-comparable worlds are indistinguishable, totality of $\leq$ gives that $\sim$ is the universal relation $W \times W$.

Definition 3 (Satisfaction Relation). *Let $\mathcal{M} = (W, \leq, \sim, V)$ be a plausibility model on P. The satisfaction relation is given by, for $w \in W$, $p \in P$, $\varphi, \varphi' \in \mathcal{L}_P$,*

$$\begin{array}{ll} \mathcal{M}, w \models p & \textit{iff } p \in V(w) \\ \mathcal{M}, w \models \neg\varphi & \textit{iff not } \mathcal{M}, w \models \varphi \\ \mathcal{M}, w \models \varphi \wedge \varphi' & \textit{iff } \mathcal{M}, w \models \varphi \textit{ and } \mathcal{M}, w \models \varphi' \\ \mathcal{M}, w \models K\varphi & \textit{iff } \mathcal{M}, v \models \varphi \textit{ for all } v \sim w \\ \mathcal{M}, w \models B^{\psi}\varphi & \textit{iff } \mathcal{M}, v \models \varphi \textit{ for all } v \in \mathit{Min}_{\leq}[\![\psi]\!]_{\mathcal{M}}, \end{array}$$

where $[\![\psi]\!]_{\mathcal{M}} := \{w \in W \mid \mathcal{M}, w \models \psi\}$. We write $\mathcal{M} \models \varphi$ to mean $\mathcal{M}, w \models \varphi$ for all $w \in W$. Further, $\models \varphi$ (φ is valid) means that $\mathcal{M} \models \varphi$ for all models $\mathcal{M}$, and $\Phi \models \varphi$ (φ is a logical consequence of the set of formulas Φ) stands for: for all $\mathcal{M}$ and $w \in \mathcal{M}$, if $\mathcal{M}, w \models \psi$ for all $\psi \in \Phi$, then $\mathcal{M}, w \models \varphi$.[3]

Example 2. Consider again the the models in Figure 1. The model on the left is of the form $\mathcal{M} = (W, \leq, \sim, V)$ with $W = \{w_1, w_2, w_3\}$ and $\leq$ defined by: $w_1 \leq w_2$, $w_2 \leq w_3$, $w_1 \leq w_3$ (plus the reflexive edges). The valuation V of the model on the left maps w_1 into $\{p\}$, w_2 into $\emptyset$ and w_3 into $\{p\}$. In all three models of the figure, the formula $Bp \wedge \neg Kp$ holds, that is, p is believed but not known.

2.2 Normal epistemic plausibility models and bisimulation

The examples and proposal of Section 1 are captured by the definition of bisimulation that follows after these preliminaries. First, given a plausibility model $\mathcal{M} = (W, \sim, \leq, V)$ consider an equivalence relation on worlds defined as follows:

$$w \approx w' \quad \text{iff} \quad V(w) = V(w').$$

The $\approx$-equivalence class of a world is defined as usual as $[w]_{\approx} = \{w' \in W \mid w' \approx w\}$. Next, the ordering $\leq$ on worlds in W can be lifted to an ordering between sets of worlds $W', W'' \subseteq W$ in the following way:

$$W' \leq W'' \quad \text{iff} \quad w' \leq w'' \text{ for all } (w', w'') \in W' \times W''.$$

Finally, the lifted ordering leads us to a formalization of normal models of Example 1.

Definition 4 (Normal Plausibility Relation). *Given a plausibility model $\mathcal{M} = (W, \leq, \sim, V)$, the* normal plausibility relation *on $\mathcal{M}$ is the relation on W defined by:*

$$w \preceq w' \quad \textit{iff} \quad \mathit{Min}_{\leq}[w]_{\approx} \leq \mathit{Min}_{\leq}[w']_{\approx}.$$

$\mathcal{M}$ is called normal *if $\preceq = \leq$. The* normalisation *of $\mathcal{M} = (W, \leq, \sim, V)$ is $\mathcal{M}' = (W, \preceq, \sim, V)$. As for $<$, we write $w \prec w'$ for $w \preceq w'$ and $w' \not\preceq w$.*

[3] For an axiomatization of this logic see e.g. [16].

Note that if $u, v \in Min_{\leq} W'$ for some set W' then, by definition of $Min_{\leq}$, both $u \leq v$ and $v \leq u$. Hence, the condition $Min_{\leq}[w]_{\approx} \leq Min_{\leq}[w']_{\approx}$ above is equivalent to the existence *some* minimal element of $[w]_{\approx}$ being $\leq$-smaller than *some* minimal element of $[w']_{\approx}$.

Lemma 1. *Let w and w' be two worlds in the normal model $\mathcal{M} = (W, \preceq, \sim, V)$. If w and w' have the same valuation, they are equiplausible.*

Proof. As $w \approx w'$, we have $[w]_{\approx} = [w']_{\approx}$, and thus $Min_{\preceq}[w]_{\approx} = Min_{\preceq}[w']_{\approx}$. By Definition 4 we $w \preceq w'$ and $w' \preceq w$, which is equivalent to $w \simeq w'$.

Example 3. Take another look at the models of Figure 1 (for reference, we name them $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$). We want models $\mathcal{M}_1$ and $\mathcal{M}_2$ to be bisimilar via the relation $\mathfrak{R}$ given by $\mathfrak{R} = \{(w_1, v_1), (w_3, v_1), (w_2, v_2)\}$ (see Section 1). Usually, in a bisimulation, every modal operator has corresponding back and forth requirements. For our logic of conditional belief there is an infinity of modal operators, as there is an infinity of of conditional formulas. (Having *only* unconditional belief $B\varphi$ defined as $B^{\top}\varphi$ is not enough, see Example 4.) Instead, we define our bisimulation indirectly by way of the plausibility relation. Example 1 showed that we cannot match 'more plausible' in $\mathcal{M}_1$ with 'more plausible' in $\mathcal{M}_2$ using simply $\leq$. With $\leq$ as seen in $\mathcal{M}_3$ (the normalization of $\mathcal{M}_1$) where $\leq = \preceq$, we can.

Definition 5 (Bisimulation). *Let plausibility models $\mathcal{M} = (W, \leq, \sim, V)$ and $\mathcal{M}' = (W', \leq', \sim', V')$ be given. Let $\preceq, \preceq'$ be the respective derived normal plausibility relations. A non-empty relation $\mathfrak{R} \subseteq W \times W'$ is a* bisimulation *between $\mathcal{M}$ and $\mathcal{M}'$ if for all $(w, w') \in \mathfrak{R}$:*

[**atoms**] $V(w) = V'(w')$.
[**forth**$_{\preceq}$] *If $v \in W$ and $v \preceq w$, there is a $v' \in W'$ s.t. $v' \preceq' w'$ and $(v, v') \in \mathfrak{R}$.*
[**back**$_{\preceq}$] *If $v' \in W'$ and $v' \preceq' w'$, there is a $v \in W$ s.t. $v \preceq w$ and $(v, v') \in \mathfrak{R}$.*
[**forth**$_{\sim}$] *If $v \in W$ and $w \sim v$, there is a $v' \in W'$ s.t. $w' \sim' v'$ and $(v, v') \in \mathfrak{R}$.*
[**back**$_{\sim}$] *If $v' \in W'$ and $w' \sim' v'$, there is a $v \in W$ s.t. $w \sim v$ and $(v, v') \in \mathfrak{R}$.*

A total bisimulation *between $\mathcal{M}$ and $\mathcal{M}'$ is a bisimulation with domain W and codomain W'. For a bisimulation between pointed models $(\mathcal{M}, w)$ and $(\mathcal{M}', w')$ it is required that $(w, w') \in \mathfrak{R}$. If a bisimulation between $(\mathcal{M}, w)$ and $(\mathcal{M}', w')$ exists, the two models are called* bisimilar *and we write $(\mathcal{M}, w) \leftrightarrows (\mathcal{M}', w')$. Two worlds w, w' of a model $\mathcal{M}$ are called* bisimilar *if there exists a bisimulation $\mathfrak{R}$ between $\mathcal{M}$ and itself with $(w, w') \in \mathfrak{R}$.*

This definition gives us the bisimulation put forth in Example 3. As $\sim$ is the universal relation on W, [forth$_{\sim}$] and [back$_{\sim}$] enforce that all bisimulations are total.

If $\sim$ was not a primitive, we could instead have conditions [up-forth$_{\preceq}$] and [up-back$_{\preceq}$] (that consider less plausible v and v'), in place of [forth$_{\sim}$] and [back$_{\sim}$]. This would define the same bisimulations.

2.3 Correspondence between bisimilarity and modal equivalence

In the following we prove that bisimilarity implies modal equivalence and vice versa. This shows that our notion of bisimulation is proper for the language and models at hand. First we define modal equivalence.

Definition 6 (Modal equivalence). *Given are models $\mathcal{M} = (W, \leq, \sim, V)$ and $\mathcal{M}' = (W', \leq', \sim', V')$ on P with $w \in W$ and $w' \in W'$. We say that $(\mathcal{M}, w)$ and $(\mathcal{M}', w')$ are* modally equivalent *iff for all $\varphi \in \mathcal{L}_P$, $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}', w' \models \varphi$. In this case we write $(\mathcal{M}, w) \equiv (\mathcal{M}', w')$.*

Lemma 2. *If two worlds of a model are $\approx$-equivalent, they are bisimilar.*

Proof. Assume worlds w and w' of a model $\mathcal{M} = (W, \leq, \sim, V)$ have the same valuation. Let $\mathfrak{R}$ be the relation that relates each world of $\mathcal{M}$ to itself and additionally relates w to w'. We want to show that $\mathfrak{R}$ is a bisimulation. This amounts to showing [atoms], [forth$_\preceq$], [back$_\preceq$], [forth$_\sim$] and [back$_\sim$] for the pair $(w, w') \in \mathfrak{R}$. [atoms] holds trivially since $w \approx w'$. [forth$_\sim$] and [back$_\sim$] also hold trivially, by choice of $\mathfrak{R}$. For [forth$_\preceq$], assume $v \in W$ and $v \preceq w$. We need to find a $v' \in W$ such that $v' \preceq w'$ and $(v, v') \in \mathfrak{R}$. Letting $v' = v$, it suffices to prove $v \preceq w'$. Since $w \approx w'$ this is immediate: $v \preceq w$ iff $Min_\leq[v]_\approx \leq Min_\leq[w]_\approx$ iff (because $w \approx w'$) $Min_\leq[v]_\approx \leq Min_\leq[w']_\approx$ iff $v \preceq w'$. [back$_\preceq$] is proved similarly.

Proposition 1. *Bisimilarity implies modal equivalence.*

Proof. We will prove that for all formulas $\varphi \in \mathcal{L}_P$, if $\mathfrak{R}$ is a bisimulation between pointed models $(\mathcal{M}, w)$ and $(\mathcal{M}', w')$ then $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}', w' \models \varphi$. The proof is by induction on the structure of φ. The base case is when φ is propositional. Then the required follows immediately from [atoms], using that $(w, w') \in \mathfrak{R}$. For the induction step, we have the following cases of φ: $\neg\psi, \psi \wedge \gamma, K\psi, B^\gamma\psi$. We skip the first three, fairly standard cases and show only $B^\gamma\psi$.

Let $\mathfrak{R}$ be a bisimulation between $(\mathcal{M}, w)$ and $(\mathcal{M}', w')$ with $\mathcal{M} = (W, \leq, \sim, V)$ and $\mathcal{M} = (W', \leq', \sim', V')$. We only prove $\mathcal{M}, w \models B^\gamma\psi \Rightarrow \mathcal{M}', w' \models B^\gamma\psi$, the other direction being proved symmetrically. So assume $\mathcal{M}, w \models B^\gamma\psi$, that is, $\mathcal{M}, v \models \psi$ for all $v \in Min_\leq[\![\gamma]\!]_\mathcal{M}$. We need to prove $\mathcal{M}', v' \models \psi$ for all $v' \in Min_{\leq'}[\![\gamma]\!]_{\mathcal{M}'}$. So let $v' \in Min_{\leq'}[\![\gamma]\!]_{\mathcal{M}'}$. Choose $x \in Min_\leq\{u \in W \mid u \approx z \text{ and } (z, v') \in \mathfrak{R} \text{ for some } z \in W\}$. Let $y \in [\![\gamma]\!]_\mathcal{M}$ be chosen arbitrarily, and choose y' with $(y, y') \in \mathfrak{R}$ (recall that any bisimulation is total). The induction hypothesis implies $\mathcal{M}', y' \models \gamma$. Let $y'' \approx y'$ be chosen arbitrarily. Lemma 2 implies the existence of a bisimulation $\mathfrak{R}'$ between $(\mathcal{M}', y'')$ and $(\mathcal{M}', y')$. Since $\mathcal{M}', y' \models \gamma$, the induction hypothesis gives us $\mathcal{M}', y'' \models \gamma$, that is, $y'' \in [\![\gamma]\!]_{\mathcal{M}'}$. Since v' was chosen $\leq'$-minimal in $[\![\gamma]\!]_{\mathcal{M}'}$, we must have $v' \leq' y''$. Since y'' was chosen arbitrarily with $y'' \approx y'$, we get $v' \leq' Min_{\leq'}[y']_\approx$. We can now conclude $Min_{\leq'}[v']_\approx \leq' v' \leq' Min_{\leq'}[y']_\approx$, and hence $v' \preceq y'$.

By [back$_\preceq$] there is a v such that $(v, v') \in \mathfrak{R}$ and $v \preceq y$. By choice of x, $x \leq Min_\leq[v]_\approx$. Since $v \preceq y$ we now get: $x \leq Min_\leq[v]_\approx \leq Min_\leq[y]_\approx \leq y$. Since y was chosen arbitrarily in $[\![\gamma]\!]_\mathcal{M}$, we can conclude:

$$x \leq u \text{ for all } u \in [\![\gamma]\!]_\mathcal{M}. \tag{1}$$

By choice of x, there is a $z \approx x$ with $(z, v') \in \mathfrak{R}$. From $z \approx x$, Lemma 2 implies the existence of a bisimulation $\mathfrak{R}''$ between $(\mathcal{M}, x)$ and $(\mathcal{M}, z)$. Since $\mathfrak{R}''$ is a bisimulation between $(\mathcal{M}, x)$ and $(\mathcal{M}, z)$, and $\mathfrak{R}$ is a bisimulation between $(\mathcal{M}, z)$ and $(\mathcal{M}', v')$, the composition $\mathfrak{R}'' \circ \mathfrak{R}$ must be a bisimulation between $(\mathcal{M}, x)$ and $(\mathcal{M}', v')$. Applying the induction hypothesis to the bisimulation $\mathfrak{R}'' \circ \mathfrak{R}$, we can from $v' \in [\![\gamma]\!]_{\mathcal{M}'}$ conclude $x \in [\![\gamma]\!]_{\mathcal{M}}$. Combining this with (1), we get $x \in \mathit{Min}_{\leq}[\![\gamma]\!]_{\mathcal{M}}$. By original assumption this implies $\mathcal{M}, x \models \psi$. Applying again the induction hypothesis to the bisimulation $\mathfrak{R}'' \circ \mathfrak{R}$, this gives us $\mathcal{M}, v' \models \psi$, as required, thereby concluding the proof.

We proceed now to the converse, that modal equivalence with regard to $\mathcal{L}_P$ implies bisimulation, though first taking a short detour motivating the need for conditional belief.

Example 4. The normal plausibility models $(\mathcal{M}_1, w_1)$ and $(\mathcal{M}_2, v_1)$ of Figure 2 are modally equivalent for the language with only unconditional belief. We can show this by first demonstrating that $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same modal description Φ (a modal description Φ of a model $\mathcal{M}$ is a set of formulas such that $\Phi \models \psi$ iff $\mathcal{M} \models \psi$). We observe that the description of both models is

$$B(p_1 \wedge \neg p_2 \wedge \neg p_3) \wedge K((p_1 \wedge \neg p_2 \wedge \neg p_3) \vee (\neg p_1 \wedge p_2 \wedge \neg p_3) \vee (\neg p_1 \wedge \neg p_2 \wedge p_3))$$

To see why, note that w_1 and v_1 are both the only minimal worlds in their respective models, so belief in (description of the valuation) $p_1 \wedge \neg p_2 \wedge \neg p_3$ will be the same. Further, in both models all three constituent worlds are epistemically possible, so K cannot distinguish either between the models (the disjunction sums up the three different valuations). We then note that, as both w_1 and v_1 satisfy $p_1 \wedge \neg p_2 \wedge \neg p_3$, $(\mathcal{M}_1, w_1)$ and $(\mathcal{M}_2, v_1)$ of Figure 2 must be modally equivalent: any boolean formula must be a consequence of $p_1 \wedge \neg p_2 \wedge \neg p_3$, whereas any belief or knowledge formula evaluated in the points of these models must be a model validity that is a consequence from the model description Φ.

On the other hand, $(\mathcal{M}_1, w_1)$ and $(\mathcal{M}_2, v_1)$ are not bisimilar. Pairs in the bisimulation must have matching valuations, so the only option is the relation $\{(w_1, v_1), (w_2, v_2), (w_3, v_3)\}$. But this does neither satisfy [forth$_{\leq}$] nor [back$_{\leq}$].

We do not want that these models are modally equivalent in, for example, a *dynamic* epistemic language. Consider an agent learning $\neg p_1$ from a public announcement. This deletes w_1 and v_1 from their respective models. After this announcement in $\mathcal{M}_1$, the agent believes p_2. In $\mathcal{M}_2$ this is not the case. Here the agent will believe p_3. With conditional belief we can capture this distinction already in the static language ($\mathcal{M}_1 \models B^{\neg p_1} p_2$, while $\mathcal{M}_2 \not\models B^{\neg p_1} p_2$).

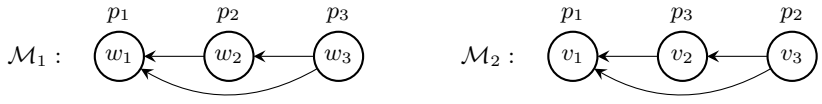

Fig. 2: The models $\mathcal{M}_1$ and $\mathcal{M}_2$ of Example 4. For visual clarity, we leave out false propositional variables.

Definition 7 (Δ). *Let two worlds w, w' of a model $\mathcal{M} = (W, \leq, \sim, V)$ on P be given where $V(w) \neq V(w')$. If there is a $p \in V(w) - V(w')$, then let $\delta_{w,w'}$ be such a p; otherwise, let $\delta_{w,w'} = \neg q$ for some $q \in V(w') - V(w)$. Any such choice of $\delta_{w,w'}$ for a given pair w, w' is called a* propositional difference *between w and w'. Further, let $\Delta_w = \bigwedge_{w' \prec w} \delta_{w,w'}$ be the conjunction of some propositional difference between w and each world that is strictly more $\preceq$-plausible than w (the empty conjunction when no such world exist).*

Continuing Example 4, we can choose $\Delta_{w_2} = \neg p_1$. We then have that $\widehat{B}^{\Delta_{w_2}} p_2$ distinguishes $\mathcal{M}_1$ and $\mathcal{M}_2$ by evaluating belief on worlds no more plausible than w_2 and v_2 respectively. However, choosing $\Delta_{w_2} = p_2$ would not distinguish, so we add an additional disjunct for w_3. Regardless of which propositional differences are used in Δ_{w_2} and Δ_{w_3}, $\widehat{B}^{\Delta_{w_2} \vee \Delta_{w_3}} p_2$ distinguishes the models. This is, of course, not sufficient for constructing distinguishing formulas in the general case, but for our purposes of proving Proposition 2 it is enough.

Lemma 3. *Let w and w' be worlds of the model $\mathcal{M} = (W, \leq, \sim, V)$ and φ a formula of $\mathcal{L}_P$, s.t. $w' \preceq w$ and $\mathcal{M}, w' \models \varphi$. Then $\mathcal{M}, w \models \widehat{B}^{\Delta_w \vee \Delta_{w'}} \varphi$.*

Proof. In the following we abbreviate $\Delta_w \vee \Delta_{w'}$ by $\Delta_{w,w'}$. We need to show that $\exists u \in Min_{\leq} [\![\Delta_{w,w'}]\!]_{\mathcal{M}}$, s.t. $\mathcal{M}, u \models \varphi$. By construction of $\Delta_{w,w'}$, we have that for all $s \in [\![\Delta_{w,w'}]\!]_{\mathcal{M}}$, either $s \approx w$, $s \approx w'$ or ($w \preceq s$ and $w' \preceq s$). By choice of w and w', we have $w' \preceq w$, meaning that $\exists w'' \in Min_{\leq} [\![\Delta_{w,w'}]\!]_{\mathcal{M}}$ such that $w' \approx w''$. Lemma 2 then says that w' and w'' are bisimilar, and Proposition 1 that they are modally equivalent. Thus $\mathcal{M}, w'' \models \varphi$. This is the u we are looking for, giving $\mathcal{M}, w \models \widehat{B}^{\Delta_{w,w'}} \varphi$.

Proposition 2. *On the class of image-finite models, modal equivalence implies bisimilarity.*

Proof. Let $\mathcal{M} = (W, \leq, \sim, V)$ and $\mathcal{M}' = (W', \leq', \sim', V')$ be two image-finite, plausibility models on P, and define $\mathfrak{R} \subseteq W \times W'$, such that $(w, w') \in \mathfrak{R}$ iff $(\mathcal{M}, w) \equiv (\mathcal{M}', w')$. We show that $\mathfrak{R}$ is in fact a bisimulation of the kind defined in Definition 5. Showing that $\mathfrak{R}$ satisfies [atoms] is trivial. We skip the, less trivial, [forth$_\sim$], and [back$_\sim$] and show the considerably more complicated case of [forth$_\preceq$] ([back$_\preceq$] is similar) as follows: Assume $(\mathcal{M}, w) \equiv (\mathcal{M}', w')$, $v \in W$ and $v \preceq w$ and show that assuming that for all $v' \in W'$, $v' \preceq w'$ implies $(\mathcal{M}, v) \not\equiv (\mathcal{M}', v')$, leads to a contradiction. This then gives $(\mathcal{M}, v) \equiv (\mathcal{M}', v')$ and therefore $(v, v') \in \mathfrak{R}$.

Let $S' = \{v' \mid v' \preceq w'\} = \{v'_1, \ldots v'_n\}$ be the successors of w'. This set is finite, due to image-finiteness of the model. If v and no successor of w' is modally equivalent, there exists formulae $\varphi^{v'_i}$, such that $\mathcal{M}, v \models \varphi^{v'_i}$ and $\mathcal{M}', v'_i \not\models \varphi^{v'_i}$. Therefore, $\mathcal{M}, v \models \varphi^{v'_1} \wedge \cdots \wedge \varphi^{v'_n}$. For notational ease, let $\Phi = \varphi^{v'_1} \wedge \cdots \wedge \varphi^{v'_n}$.

With $\mathcal{M}, v \models \Phi$, Lemma 3 gives $\mathcal{M}, w \models \widehat{B}^{\Delta_{w,v}} \Phi$ ($\Delta_{w,v}$ is finite due to image-finiteness of the models). Now, $\mathcal{M}', w' \models \widehat{B}^{\Delta_{w,v}} \Phi$ (which we must have due to modal equivalence) iff there exists a $u' \in Min_{\leq} [\![\Delta_{w,v}]\!]_{\mathcal{M}'}$ such that $\mathcal{M}', u' \models \Phi$. By construction of Φ, no world v'_i exists such that $v'_i \preceq w'$ and $\mathcal{M}', v'_i \models \Phi$, so

we must have $w' \prec u'$. There are two cases for (the weakest requirements for) this u' to be minimal. Either (i) $u' \leq w'$ or (ii) $w' < u'$ and $w' \notin [\![\Delta_{w,v}]\!]_{\mathcal{M}'}$. If (i) is the case, we must have a world w'', with $w'' \approx w'$ and $w'' < u'$, or we couldn't have $w' \prec u'$. But $w'' < u'$ means that u' cannot be minimal unless $w' \notin [\![\Delta_{w,v}]\!]_{\mathcal{M}'}$, because otherwise $w'' \in [\![\Delta_{w,v}]\!]_{\mathcal{M}'}$. So, for (i) and (ii) both, we must have $w' \notin [\![\Delta_{w,v}]\!]_{\mathcal{M}'}$. This yields $\mathcal{M}', w' \models \neg\Delta_{w,v}$. But as $\mathcal{M}, w \models \Delta_{w,v}$, we get the sought after contradiction of $(\mathcal{M}, w) \equiv (\mathcal{M}', w')$.

3 Degrees of belief and safe belief

In this section we sketch some further results that can be obtained for our single-agent setting of the logic of knowledge and conditional belief. Apart from *conditional belief*, other familiar epistemic notions in the philosophical logical and artificial intelligence community are *safe belief* [16] and *degrees of belief* [10, 15]. Our results generalize fairly straightforwardly to such other notions. An agent has *safe belief* in formula φ iff it will continue to believe φ no matter what *true* information conditions its belief.[4]

Definition 8 (Safe belief). *We extend the inductive language definition with a clause $\Box\varphi$ for safe belief in φ. The semantics are $\mathcal{M}, w \models \Box\varphi$ iff ($\mathcal{M}, w \models B^{\psi}\varphi$ for all ψ such that $\mathcal{M}, w \models \psi$).*

Degrees of belief are a quantitative alternative to conditional belief. The zeroth degree of belief $B^0\varphi$ is defeasible belief $B\varphi$ as already defined. For $\mathcal{M}, w \models B^1\varphi$ to hold φ should be true in (i) all minimal worlds accessible from w; but additionally, (ii) if you take away those from the equivalence class, in all worlds that are now minimal. If we do this with the normal plausibility relation we get what we want (otherwise, we run into the same problems as before — our treatment is not compatible with e.g. Spohn's approach [15], that allows 'gaps' (layers without worlds) in between different degrees of belief).

$$
\begin{array}{lll}
Min^0_{\preceq}[w]_{\sim} & := Min_{\preceq}([w]_{\sim}) & \\
Min^{n+1}_{\preceq}[w]_{\sim} & := Min^n_{\preceq}[w]_{\sim} & \text{if } Min^n_{\preceq}([w]_{\sim}) = [w]_{\sim} \\
Min^{n+1}_{\preceq}[w]_{\sim} & := Min^n_{\preceq}[w]_{\sim} \cup Min_{\preceq}([w]_{\sim} \setminus Min^n_{\preceq}[w]_{\sim}) & \text{otherwise}
\end{array}
$$

We now can define the logic of knowledge and degrees of belief.

Definition 9 (Degrees of belief). *We replace the clause for conditional belief in the inductive language definition by a clause $B^n\varphi$ for belief in φ to degree n, for $n \in \mathbb{N}$. The semantics are*

$$\mathcal{M}, w \models B^n\varphi \text{ iff for all } v \in Min^n_{\preceq}([w]_{\sim}) : \mathcal{M}, v \models \varphi$$

In an extended version of this paper we are confident that we will prove that the logics of conditional belief and knowledge, of degrees of belief and knowledge, and both with the addition of safe belief are all expressively equivalent.

[4] This definition is conditional to modally definable subsets, unlike [5, 16] where it is on any subset. In that case safe belief is not bisimulation invariant and increases the expressivity of the logic.

4 Multi-agent epistemic doxastic logic

For a finite set A of agents and a set of propositional symbols P the *multi-agent epistemic-doxastic language* $\mathcal{L}_{P,A}$ is

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

where $p \in P$ and $a \in A$. *Epistemic plausibility models* are generalized similarly, we now have plausibility relations $\leq_a$ and epistemic relations $\sim_a$ for each agent a. For each agent the domain is partitioned into (possibly) various equivalence classes, such that each class is a well-preorder. The single-agent results do not simply transfer to the multi-agent stage. We give an example.

Example 5. Consider Figure 3. The solid arrows represent the plausibilities for agent a and the dashed arrow for agent b. In our example, the partition for a is $\{w_0\}, \{w_1, w_2, w_3\}$, whereas the partition for b is $\{w_0, w_1\}, \{w_2\}, \{w_3\}$. Unlike before, the two p-states are not bisimilar, because in the state w_1 agent b is uncertain about the value of p but defeasibly believes p (there is a less plausible alternative w_0, whereas in state w_3 agent b knows (and believes) that p. In both worlds, of course, agent a still believes that p, but a distinguishing formula between the two is now, for example, $\neg K_b p \wedge B_a p$, true in w_1 but false in w_3.

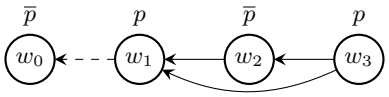

Fig. 3: A plausibility model wherein the two p worlds are not bisimilar, because they have different higher-order belief properties.

It will be clear from Example 5 that we cannot, for each agent, derive a normal plausibility relation $\preceq_a$ from a given plausibility relation $\leq_a$ by identifying worlds with the same valuation: $w \approx_a w'$ iff $V(w) = V(w')$ and $w \sim_a w'$ does not work (worlds w_1 and w_3 in Example 3 satisfy different formulas). Some strengthening guarantees that bisimilarity still implies modal equivalence. An example is, using the above $\approx_a$:

$$\begin{array}{ll} w \approx w' & \text{iff } \text{ (for all agents } a : w \approx_a w') \\ w \preceq_a w' & \text{iff } \text{ } (Min_{\leq_a}[w]_{\approx} \leq_a Min_{\leq_a}[w']_{\approx}) \end{array}$$

Unfortunately we do not get that modal equivalence then implies bisimilarity. The strongest possible approach is of course to require that [$w \approx w'$ iff (w, w') is a pair in the bisimulation relation]. This works, but it is is rather self-defeating. In due time we hope to find a proper generalisation in between these two extremes.

5 Planning

In planning an agent is tasked with finding a course of action (i.e. a plan) that achieves a given goal. A planning problem implicitly represents a state-transition

system, where transitions are induced by actions. Exploring this state-space is a common method for reasoning about and synthesising plans. A growing community investigates planning in dynamic epistemic logic [6, 12, 4, 1], and using the framework presented here we can in similar fashion consider planning with doxastic attitudes. To this end we identify states with plausibility models, and the goal with a formula of the epistemic doxastic language. Further we can describe the dynamics of actions by using e.g. hard announcements or soft announcements [17], or yet more expressive notions such as event models [5].

With the state-space consisting of plausibility models, model theoretic results become pivotal to the development of planning algorithms. In general, we cannot require even single-agent plausibility models (even on a finite set of propositional symbols) to be finite. Also, normal plausibility models need not be finite — obvious, as the 'normalising' procedure in which we replace $\leq$ by $\preceq$ does not change the domain. Our definition of bisimulation has a crucial property in this regard: By Lemma 2 the bisimulation contraction of a model will contain no two worlds with the same valuation, hence any bisimulation minimal model on a finite set of propositions is finite. Moreover, two bisimulation minimal models are bisimilar exactly when they are isomorphic, and it follows that are only finitely many distinct bisimulation minimal epistemic plausibility models. With the reasonable assumption that actions preserve bisimilarity (this is the case for the types of actions mentioned above), our investigations on the proper notion of bisimulation therefore allow us to employ a smaller class of models in planning. This is a chief motivation for our work here, and an immediate consequence is that determining whether there exists a plan for a plausibility planning problem is *decidable* (see [2]).

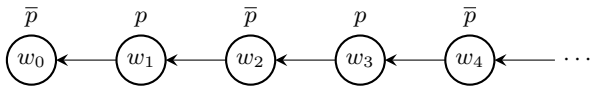

Fig. 4: Uncontractable chain of p and $\neg p$-worlds.

It is remarkable that the approach of [8] to defining bisimulation for epistemic plausibility models does not yield decidability of planning problems, not even for single-agent models defined on a single proposition. It has, for instance, that the model in Figure 4 consisting of an infinite 'directed chain' of alternating p and $\neg p$ worlds (a copy of the natural numbers axis) is bisimulation minimal. In our approach the bisimulation minimal model would be the middle one of Figure 1, regardless of the number of worlds. Though [8] also shows that bisimilarity implies modal equivalence and vice versa (for image finite models), this is not inconsistent with our results here. Another difference between our approach and [8] lies in the semantics of safe belief. There, safe belief is relative to any subset (see also Footnote 4). For a 'directed chain' model, the safe belief semantics of [8] permits counting the number of p and $\neg p$ worlds. Such more expressive semantics naturally come at a cost, namely having no finite bound on the size of minimal single-agent models.

Acknowledgements

We thank the reviewers of the AI 13 conference for their comments. Hans van Ditmarsch is also affiliated to IMSc (Institute of Mathematical Sciences), Chennai, as research associate. He acknowledges support from European Research Council grant EPS 313360.

References

1. M.B. Andersen, T. Bolander, and M.H. Jensen. Conditional epistemic planning. In *Proc. of 13th JELIA*, LNCS 7519, pages 94–106. Springer, 2012.
2. M.B. Andersen, T. Bolander, and M.H. Jensen. Don't plan for the unexpected: Planning based on plausibility models. *Logique et Analyse*, 2014, To Appear.
3. G. Aucher. A combined system for update logic and belief revision. In *Proc. of 7th PRIMA*, pages 1–17. Springer, 2005. LNAI 3371.
4. G. Aucher. DEL-sequents for regression and epistemic planning. *Journal of Applied Non-Classical Logics*, 22(4):337–367, 2012.
5. A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In *Proc. of 7th LOFT*, Texts in Logic and Games 3, pages 13–60. Amsterdam University Press, 2008.
6. T. Bolander and M.B. Andersen. Epistemic planning for single and multi-agent systems. *Journal of Applied Non-classical Logics*, 21(1):9–34, 2011.
7. K. Britz and I. Varzinczak. Defeasible modalities. In *Proc. of the 14th TARK*, 2013.
8. L. Demey. Some remarks on the model theory of epistemic plausibility models. *Journal of Applied Non-Classical Logics*, 21(3-4):375–395, 2011.
9. A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.
10. S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.
11. D.K. Lewis. *Counterfactuals*. Harvard University Press, Cambridge (MA), 1973.
12. B. Löwe, E. Pacuit, and A. Witzel. DEL planning and some tractable cases. In *Proc. of LORI 3*, pages 179–192. Springer, 2011.
13. T.A. Meyer, W.A. Labuschagne, and J. Heidema. Refined epistemic entrenchment. *Journal of Logic, Language, and Information*, 9:237–259, 2000.
14. K. Segerberg. Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic*, 39(3):287–306, 1998.
15. W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134, 1988.
16. R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.
17. J. van Benthem. Dynamic logic of belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.
18. J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011.
19. H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action)*, 147:229–275, 2005.
20. H. van Ditmarsch and W.A. Labuschagne. My beliefs about your beliefs – a case study in theory of mind and epistemic logic. *Synthese*, 155:191–209, 2007.