Seeing is Believing: Formalising False-Belief Tasks in Dynamic Epistemic Logic

Thomas Bolander

Abstract In this paper we show how to formalise false-belief tasks like the Sally-Anne task and the second-order chocolate task in *Dynamic Epistemic Logic* (DEL). False-belief tasks are used to test the strength of the *Theory of Mind* (ToM) of humans, that is, a human's ability to attribute mental states to other agents. Having a ToM is known to be essential to human social intelligence, and hence likely to be essential to social intelligence of artificial agents as well. It is therefore important to find ways of implementing a ToM in artificial agents, and to show that such agents can then solve false-belief tasks. In this paper, the approach is to use DEL as a formal framework for representing ToM, and use reasoning in DEL to solve false-belief tasks. In addition to formalising several false-belief tasks in DEL, the paper introduces some extensions of DEL itself: *edge-conditioned event models* and *observability propositions*. These extensions are introduced to provide better formalisations of the false-belief tasks, but expected to have independent future interest.

1 Introduction

Social intelligence is the ability to understand others and the social context effectively and thus to interact with other agents successfully. Research has suggested that *Theory of Mind* (ToM) may play an important role in explaining social intelligence. ToM is the ability to attribute mental states—beliefs, intentions, etc.—to oneself and others and to understand that others might have mental states that are different from one's own (Premack and Woodruff 1978). The strength of a human child's ToM is often tested with a *false-belief task* such as the *Sally-Anne task* (Wimmer and Perner 1983).

Thomas Bolander
Technical University of Denmark; Richard Petersens Plads, Building 324; DK-2800 Lyngby
e-mail: tobo@dtu.dk

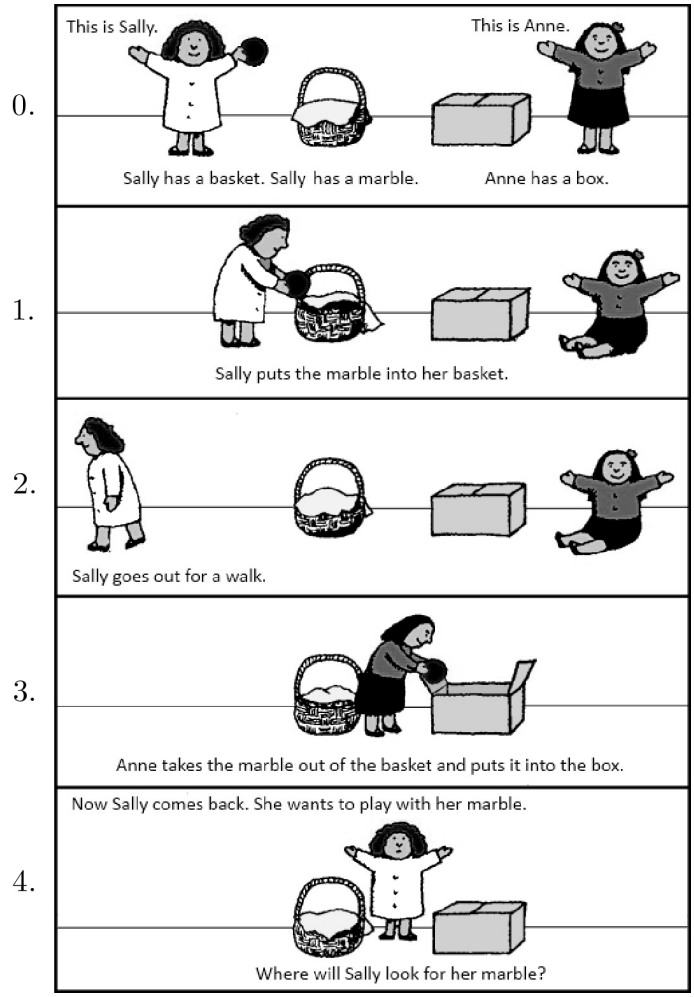


Fig. 1 An illustration of the Sally-Anne false belief task. The illustration is by Axel Scheffler and is borrowed from the autism book by Frith (1989).

Example 1 (The Sally-Anne task). The Sally-Anne task is illustrated in Figure 1. It is based on a story with two agents, Sally and Anne, that has the following 5 steps, corresponding to the 5 pictures of Figure 1:

0. Sally and Anne are in a room. Sally is holding a marble. There is a basket and a box in the room.
1. Sally puts the marble into the basket.
2. Sally leaves the room.
3. Anne transfers the marble to the box.
4. Sally re-enters.

When used as a cognitive test for children, the child is told or shown the story in the figure. At the end, the child is asked "where does Sally believe the marble to be?" Passing the test means answering "in the basket", since Sally didn't see Anne transfer the marble from the basket to the box, and hence Sally has the *false belief* that it is still in the basket. If the child answers "in the box", where in fact the marble is, the child has *failed* the test.[1] Children under the age of 4, and autistic children in general, are generally unable to pass the Sally-Anne test (Wimmer and Perner 1983; Baron-Cohen et al. 1985).

To create AI agents with social intelligence, it seems relevant to consider the possibility of equipping such agents with a ToM, and to test them using false-belief tasks. The idea here is that for an AI agent, e.g. a robot, to be considered truly 'socially intelligent', it should at least be able to pass these false-belief tasks. Hence it becomes important to find ways of *formalising* ToM and false-belief tasks in a way that will allow computers to do the required reasoning.

The goal of the present paper is to present one such possible formalisation, using the framework of *Dynamic Epistemic Logic* (DEL). We will now explain why DEL is a fairly natural choice here. First of all, we need a formalism that can represent the beliefs of other agents, e.g. the beliefs of Sally, Sally's beliefs about Anne, etc. This naturally leads one to consider an *epistemic logic* (or, more precisely, a *doxastic logic*, but we will here still refer to it as *epistemic*). Basic epistemic logic is however only sufficient to model static state of affairs, like "Sally believes the marble to be in the basket." In the false-belief tasks we also need to be able to model the *dynamics* of beliefs, e.g. "After Anne has moved the marble to the box, Sally still believes it to be in the basket." This is where DEL comes into the picture: it has a natural way to deal with static states of beliefs (the *epistemic models* of DEL), a natural way to describe actions with epistemic and/or world changing effects (the *event models* of DEL), and a simple way of calculating the result of executing an action in a state (the *product update* of DEL).

Hintikka, transmissibility and autistic agents

DEL is a dynamic version of epistemic logic, where actions and their effects can directly be described in the logic. The seminal treatise on (non-dynamic) epistemic logic is the book 'Knowledge and Belief' by Hintikka (1962). Hintikka carefully introduces and discusses the required semantic properties of the knowledge and belief modalities. He settles for a knowledge modality in which the accessibility relation has to satisfy reflexivity, seriality and transitivity, and where the only difference between the knowledge and belief modalities is that (the accessibility relation of)

[1] One might argue that if Sally is capable of doing *intention recognition*, that is, predict the goals of Anne, she might actually suspect that Anne has been transferring the marble while she was away, because she perhaps knows that Anne wants the marble for herself. However, it is implicit in the Sally-Anne task that intention recognition is not taken into account, and we will not do it here either.

the belief modality does not have to satisfy reflexivity. This essentially means that the only difference between knowledge and belief is that beliefs can be false (since reflexivity of the knowledge modality is equivalent to the property that everything known is true). False beliefs are clearly essential to this paper, since they are indeed the most essential ingredient of the false-belief tasks. False-belief tasks are actually about a certain type of false beliefs, where the false belief ϕ is ascribed to another agent, that is, the situation can be formalised by a formula of the form $B_a\neg\phi \wedge B_aB_b\phi$: agent a believes $\neg\phi$ and at the same time believes that b wrongly believes ϕ (see Section 3 for details on the epistemic language). In the case of the Sally-Anne task, a formula describing the state of mind of a child having successfully passed the test would be $B_{child}\neg basket \wedge B_{child}B_{Sally}basket$: the child believes, indeed knows, that the marble is is not in the basket, but at the same time believes that Sally believes it to be in the basket.

Hintikka (1962) discusses the principle of transmissibility: "If I know that you know that p is true, I virtually know myself that p is true." The idea is that if an agent a knows that another agent b knows some fact ϕ, then agent a should itself be allowed to claim to know ϕ. The principle is formalised by $K_aK_b\phi \rightarrow K_a\phi$. It is different with beliefs. Agent a might believe agent b to believe ϕ without agent a then starting to believe p. "Beliefs are not transmissible", as Hintikka says, that is, $B_aB_b\phi \rightarrow B_a\phi$ is not valid. In fact, transmissibility of belief directly contradicts the possibility of attributing a false belief to another agent, the essence of the false belief tasks. This is so since the combination of the false belief attribution formula $B_a\neg\phi \wedge B_aB_b\phi$ and transmissibility directly leads to the inconsistent belief $B_a\neg\phi \wedge B_a\phi$, which is a non-satisfiable formula when the accessibility relation of belief is serial. The fact that knowledge is transmissible and belief is not can be explained in terms of the trivial difference between reflexivity and non-reflexivity of the underlying accessibility relations, as Hintikka notes.

Van Ditmarsch and Labuschagne (2007) consider an "autistic" agent type described as an "agent a such that the ToM of agent a imputes to every agent b a state of mind identical to a's own". A possible way to formalise this (different from the one considered by Van Ditmarsch and Labuschagne (2007)) is the converse of Hintikka's transmissibility principle: $B_a\phi \rightarrow B_aB_b\phi$ (whatever agent a believes, agent a also believes agent b to believe). The connection to autism is that autistic children are known to have a defective ToM, and "one interpretation of this failure of mentalising is to regard autistic children as possessing a rudimentary ToM in which the belief of other agents are assumed to be identical to those of the imputing agent" (Van Ditmarsch and Labuschagne 2007). The converse transmissibility principle clearly also makes it impossible to ascribe a false belief to another agent. From a false-belief attribution $B_a\neg\phi \wedge B_aB_b\phi$ and the converse transmissibility principle $B_a\phi \rightarrow B_aB_b\phi$ we get $B_aB_b\neg\phi \wedge B_aB_b\phi$, which is also non-satisfiable when the accessibility relation is serial. Hence an agent a that satisfies either the transmissibility or converse transmissibility principle will have no possibility of passing a false-belief task.[2]

[2] Except if the tested agent itself ends up forming a false belief. For instance, in the case of the Sally-Anne task, an agent without a ToM could in principle pass the test by being fooled into itself

In this paper, the accessibility relation of belief is in all models going to satisfy seriality and transitivity. This is not as obvious and trivial as one might at first think. In general, seriality is not preserved under the update of epistemic states in dynamic epistemic logic (see, e.g., Aucher (2012)). However, as we will see, the false-belief tasks we are going to consider only involve action types that preserve both transitivity and seriality, the conditions of Hintikka's belief modality.

Structure of the paper

Below we will first, in Section 2, briefly present the qualities we aim for in our false-belief task formalisations. Next, in Section 3, we introduce the required parts of DEL, and then apply it to formalise the Sally-Anne task in Section 4. The formalisation turns out not to be entirely satisfactory, and hence we will, in Section 5, introduce an extension of DEL that gives more appropriate formalisations. The improved formalisations are in Section 6.

2 Robustness and Faithfulness

Above we claim that DEL is a fairly natural choice for the formalisation of false-belief tasks. This of course doesn't imply that it is the *only* natural choice. Indeed, there are several existing formalisations of false-belief tasks in the literature, using different formal frameworks. Figure 2 gives a brief overview of the full formalisations and implemented systems we know of. In addition to these we should mention Stenning and Van Lambalgen (2008), who gives a detailed logical analysis of several false-belief tasks, though no full formalisations. The Sally-Anne task is usually referred to as a *first-order* false-belief task since it only involves *first-order belief attribution*: the child has to attribute beliefs to Sally, but not, say, to Sally's beliefs about Anne's beliefs (which would be second-order belief attribution). Most of the existing formalisations can only deal with first-order or at most second-order false-belief tasks. We wish to be more general, and at the same time have formalisations that stay as close as possible to the informal versions of the tasks, and so propose the following two criteria:

Robustness. *The formalism should not only be able to deal with one or two selected false-belief tasks, but with as many as possible, with no strict limit on the order of belief attribution.*

Faithfulness. *Each action of the false-belief story should correspond to an action in the formalism in a natural way, and it should be fairly straightforward, not requiring*

believing that the marble is in the basket. When asked about where Sally believes the marble to be, the agent would consult its own beliefs, and answer "in the basket". But often in the Sally-Anne task, two questions are asked: "where is the marble?" and "where does Sally believe the marble to be". To pass the test, the answers to the two questions must be distinct.

system/reference	year	formalism/platform	h-o reas.	other features
CRIBB (Wahl and Spada 2000)	2000	Prolog	≤ 2	goal recognition, plan recognition
Edd Hifeng (Arkoudas and Bringsjord 2008)	2008	event calculus	≤ 1	Second Life avatar
Leonardo (Breazeal et al. 2011)	2011	C5 agent architecture	≤ 1	goal recognition, learning
(Sindlar 2011)	2011	extension of PDL, implemented in 2APL	≤ 1	goal recognition
ACT-R agent (Arslan et al. 2013)	2013	ACT-R cognitive architecture	∞	learning
(Braüner 2013)	2013	hybrid logic	∞	temporal reasoning

Fig. 2 Existing full formalisations/implementations of false-belief tasks, ordered chronologically. The numbers in the 'h-o reas.' column refer to the highest level of belief attribution the formalism/system allows (∞ if there is no upper bound).

ingenuity, to find out what that action of the formalism is. The formalisation of the false-belief story should only consist of these formalised actions.

The idea behind the faithfulness criterion is that the ultimate aim is to have an autonomous agent who can formalise the false-belief story only by being told the informal, natural language variant of it. This agent should not be required ingenuity in translating the steps of the informal story into their formal counterparts, and it should not be necessary to provide the agent with information that goes beyond the story itself (that is, it is not allowed to "cheat" by providing the agent with additional information which is not explicitly part of the informal version of the story).

Of the formalisations listed in Figure 2, the first four only allow belief attribution to a fixed order (first- or second-order), and hence do not satisfy our robustness criterion. In principle, all of them could be generalised to handle any *fixed* level of higher-order belief attribution, but a fixed level is still not satisfying our criterion. The last two formalisations have the generality in terms of higher-order belief-attribution that we are after. However, in the hybrid logic approach, there is no explicit representation of actions, which goes against our chosen faithfulness criterion. The closest to our approach of modelling the false-belief tasks in DEL is probably the ACT-R agent listed second to last in Figure 2. However, in the ACT-R formalisation, it is explicitly mentioned as part of the formalised story who observes who at which points of time during the story (Arslan et al. 2013). As this is usually not explicitly mentioned as part of the false-belief stories (see in particular the second-order chocolate task formalised in Section 6), it does not fully satisfy our faithfulness criterion.

One can distinguish approaches to formalising false-belief tasks that seek to: 1) provide formal models of human reasoning; 2) provide the basis for a reasoning engine of autonomous agents. These two are of course not always disjoint aims, as discussed by Verbrugge (2009) (and further explored in the context of strategic reasoning by Ghosh et al. (2014)). In this paper, however, we are exclusively con-

cerned with the second aim. The ultimate aim of this line of work is to construct autonomous planning agents with ToM capabilities using dynamic epistemic logic (see Bolander and Andersen (2011) for further details).

3 Dynamic Epistemic Logic

In this section we will introduce the required basics of dynamic epistemic logic (DEL). The less technically inclined, or interested, reader can browse very quickly through the definitions and instead focus on the examples that illustrate the workings of the formalism in relation to the Sally-Anne task. Basic familiarity with epistemic logic, but not necessarily DEL, is expected. All definitions in this section are well-known and standard in DEL. The particular variant presented here is adopted from van Ditmarsch and Kooi (2008).

Epistemic Models

Throughout this article, P is an infinite, countable set of atomic propositions (propositional symbols), and $\mathcal{A}$ is a non-empty finite set of agents. We will most often use lower case letters $p,q,r,\ldots$ for atomic propositions and capital letters $A,B,C,\ldots$ for agents. Variables ranging over agents will be denoted $i,j,k,\ldots$. The epistemic language $\mathcal{L}(P,\mathcal{A})$ is generated by the following BNF:

$$\phi \ ::= \ p \ \mid \ \neg\phi \ \mid \ \phi\wedge\phi \ \mid \ B_i\phi \ \mid \ C_{\mathcal{B}}\phi$$

where $p\in P$, $i\in\mathcal{A}$, and $\mathcal{B}\subseteq\mathcal{A}$. We read $B_i\phi$ as "agent i believes ϕ", and $C_{\mathcal{B}}\phi$ as "it is common belief among the agents in $\mathcal{B}$ that ϕ". The formula $\phi\vee\psi$ is an abbreviations of $\neg(\neg\phi\wedge\neg\psi)$, and we define $\top$ as an abbreviation for $p\vee\neg p$ and $\bot$ as an abbreviation for $p\wedge\neg p$ for some arbitrarily chosen $p\in P$. Furthermore, we use $E\phi$ as abbreviation for $\bigwedge_{i\in\mathcal{A}}B_i\phi$. We read $E\phi$ as "everybody believes ϕ." The semantics of $\mathcal{L}(P,\mathcal{A})$ is defined through *epistemic models*.

Definition 1 (Epistemic models and states). An *epistemic model of* $\mathcal{L}(P,\mathcal{A})$ is $\mathcal{M}=(W,R,V)$, where

- W, the *domain*, is a set of *worlds*;
- $R:\mathcal{A}\to 2^{W\times W}$ assigns an *accessibility relation* $R(i)$ to each agent $i\in\mathcal{A}$;
- $V:P\to 2^W$, the *valuation*, assigns a set of worlds to each atomic proposition.

The relation $R(i)$ is usually abbreviated R_i, and we write wR_iv when $(w,v)\in R_i$. For $w\in W$, the pair $(\mathcal{M},w)$ is called a *state* of $\mathcal{L}(P,\mathcal{A})$, and w is referred to as the *actual world*. An epistemic model $\mathcal{M}=(W,R,V)$ or state $(\mathcal{M},w)$ is called *serial* if each relation R_i is serial, that is, if for all $w\in W$ and all $i\in\mathcal{A}$, there exists a $v\in W$ with wR_iv.

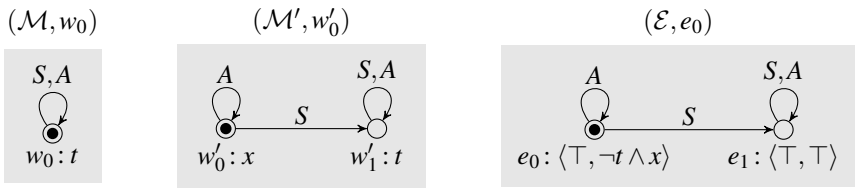


Fig. 3 Two states, $(\mathcal{M}, w_0)$ and $(\mathcal{M}', w_0')$, and an action, $(\mathcal{E}, e_0)$.

The truth conditions—that is, the definition of $(\mathcal{M}, w) \models \phi$ for models $\mathcal{M} = (W, R, V)$, worlds $w \in W$ and formulas $\phi \in \mathcal{L}(P, \mathcal{A})$—are standard:

$$
\begin{array}{lll}
(\mathcal{M}, w) \models p & \text{iff} & w \in V(p) \\
(\mathcal{M}, w) \models \neg\phi & \text{iff} & \mathcal{M}, w \not\models \phi \\
(\mathcal{M}, w) \models \phi \wedge \psi & \text{iff} & \mathcal{M}, w \models \phi \text{ and } \mathcal{M}, w \models \psi \\
(\mathcal{M}, w) \models B_i\phi & \text{iff} & \text{for all } v \in W, \text{ if } wR_iv \text{ then } \mathcal{M}, v \models \phi \\
(\mathcal{M}, w) \models C_{\mathcal{B}}\phi & \text{iff} & \text{for all } v \in W, \text{ if } (w, v) \in (\bigcup_{i \in \mathcal{B}} R_i)^* \text{ then } \mathcal{M}, v \models \phi
\end{array}
$$

In this paper, all considered epistemic models and states are going to be serial. Seriality is necessary to preserve consistent beliefs. If a model $\mathcal{M} = (W, V, R)$ is *not* serial, it means there is a world $w \in W$ and an agent $i \in \mathcal{A}$ such that there is no $v \in W$ with wR_iv. According to the truth conditions above this implies that $(\mathcal{M}, w) \models B_i\bot$. Since we are going to formalise false-belief tasks, consistency can not always be expected to be preserved. If Sally gets a false belief that the marble is in the basket, and Anne then announces: "the marble is in the box", then Sally will get inconsistent beliefs (if we treat the announcement as a standard truthful announcement). However, for the false-belief tasks considered in this paper, we are going to see that seriality can be preserved, and inconsistent beliefs hence be avoided (essentially because the considered false-belief tasks do not involve announcements).

Example 2. We will now illustrate the notion of a state relative to the Sally-Anne task of Example 1. The example states are $(\mathcal{M}, w_0)$ and $(\mathcal{M}', w_0')$ of Figure 3. Here we have two atomic propositions, x and t, where x is intended to mean "the marble is in the box", and t means "the marble is in the basket". We use the agent symbols S and A for Sally and Anne, respectively.

In $(\mathcal{M}, w_0)$ and $(\mathcal{M}', w_0')$, and states in general, each world is marked by its name followed by a list of the atomic propositions true at that world (which may be empty if none holds true). Sometimes we will drop names on worlds and just label them by the list of true propositions. Edges are labelled with the name of the relevant accessibility relations (agents). We use the symbol ◉ to mark the actual world.

Consider $(\mathcal{M}, w_0)$. The actual world is w_0, that is, the marble is in the basket (t holds). The loop at w_0 for S and A means that Sally and Anne consider the actual world w_0 possible, and the absence of other edges means that they *only* consider w_0 possible. Hence we have $(\mathcal{M}, w_0) \models C_{S,A}t$: it is common belief among Sally and

Anne that the marble is in the basket. The state $(\mathcal{M}, w_0)$ corresponds to the situation before Anne has transferred the marble to the box.

Consider now $(\mathcal{M}', w_0')$. This corresponds to the situation after Anne has transferred the marble in Sally's absence. The actual world now satisfies x. In the actual world, w_0, Anne only considers w_0 possible (signified by the loop labelled A at w_0): she *knows* the marble to be in the box. However, Sally doesn't have such a loop at w_0, rather she has an edge going to w_1 where t holds. This means that, in the actual world w_0, Sally only considers it possible that the actual world is in fact w_1. Hence she has a *false belief* that the marble is in the basket (a false belief that t holds). Formally,

$$(\mathcal{M}', w_0') \models x \wedge B_A x \wedge B_S t.$$

We have now seen how we can use states to model the beliefs of Sally and Anne before and after the marble is moved. But we also need a way to model the act of moving the marble. This is done using DEL event models, presented next.

Event Models

DEL introduces the concept of *event model* (or *action model*) for modeling the changes to states brought about by the execution of actions (Baltag et al. 1998; Baltag and Moss 2004). We here use a variant that includes postconditions (van Ditmarsch et al. 2005; van Benthem et al. 2006; Bolander and Andersen 2011), which means that actions can have both epistemic effects (changing the beliefs of agents) and ontic effects (changing the physical facts of the world).

Definition 2 (Event models and actions). An *event model of* $\mathcal{L}(P, \mathcal{A})$ is $\mathcal{E} = (E, Q, pre, post)$, where

- E, the *domain*, is a finite non-empty set of *events*;
- $Q : \mathcal{A} \to 2^{E \times E}$ assigns an *accessibility relation* $Q(i)$ to each agent $i \in \mathcal{A}$;
- $pre : E \to \mathcal{L}(P, \mathcal{A})$ assigns to each event a *precondition*, which can be any formula in $\mathcal{L}(P, \mathcal{A})$.
- $post : E \to \mathcal{L}(P, \mathcal{A})$ assigns to each event a *postcondition*. Postconditions are conjunctions of propositional literals, *i.e.*, conjunctions of atomic propositions and their negations (including $\top$ and $\bot$).

The relation $Q(i)$ is generally abbreviated Q_i. For $e \in E$, $(\mathcal{E}, e)$ is called an *action* (or *pointed event model*) of $\mathcal{L}(P, \mathcal{A})$, and e is referred to as the *actual event*. An event model $\mathcal{E} = (E, Q, pre, post)$ or action $(\mathcal{E}, e)$ is called *serial* if each relation Q_i is serial, that is, if for all $e \in E$ and all $i \in \mathcal{A}$, there exists an $f \in E$ with $eR_i f$.

Example 3. Consider the action $(\mathcal{E}, e_0)$ of Figure 3. Labeling events by the pair $\langle \phi_1, \phi_2 \rangle$ means that the event has precondition ϕ_1 and postcondition ϕ_2. Hence the actual event, e_0, corresponds to the action of making t false and x true, that is, it is the act of transferring the marble from the basket to the box. The event e_1 has trivial

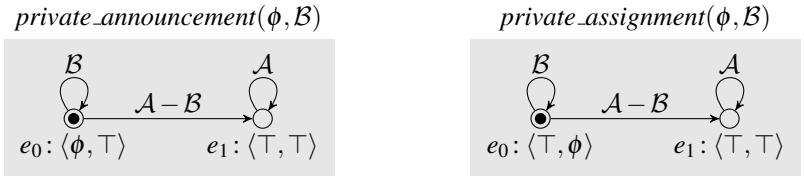


Fig. 4 Left: Private announcement of ϕ to the agents in $\mathcal{B} \subseteq \mathcal{A}$. Right: Private assignment of ϕ (ϕ becomes true) to the agents in $\mathcal{B}$.

pre- and post-conditions meaning that it is a 'skip' action representing that nothing happens. Looking at the edges of the action, we see that Anne only considers it possible that the marble is transferred (the loop at e_0), whereas Sally only considers it possible that nothing happens (she only has en edge from the actual event to the 'skip' event e_1). Hence the model encodes an action where the marble is *actually* transferred from the basket to the box, Anne is aware of this, but Sally thinks that nothing happens. It hence encodes step 3 of the Sally-Anne task, cf. Example 1.

The action $(\mathcal{E}, e_0)$ has the same form as a private announcement (Baltag et al. 1998), except it is an ontic action, so it should probably rather be called a *private assignment*. More generally, a *private announcement* of ϕ to a group of agents $\mathcal{B} \subseteq \mathcal{A}$ can be represented as the event model *private_announcement*$(\phi, \mathcal{B})$ of Figure 4, and the corresponding *private assignment* of ϕ to group $\mathcal{B}$ as *private_assignment*$(\phi, \mathcal{B})$ of the same figure. Note that the two event models only differ by ϕ being a precondition in the announcement and a postcondition in the assignment. In both event models, the agents in $\mathcal{B}$ observe that the event e_0 takes place (the $\mathcal{B}$-loop at e_0), whereas the agents *not* in $\mathcal{B}$ thinks that nothing happens (the $\mathcal{A}-\mathcal{B}$-edge leading to the 'skip' event e_1). We note that the action $(\mathcal{E}, e_0)$ of transferring the marble in Sally's absence is *private_assignment*$(\neg t \wedge x, \{A\})$.

Product Update

Assume given a state $(\mathcal{M}, w_0)$ and an action $(\mathcal{E}, e_0)$. The product update yields a new state $(\mathcal{M}, w_0) \otimes (\mathcal{E}, e_0)$ representing the situation after the action $(\mathcal{E}, e_0)$ has been executed in the state $(\mathcal{M}, w_0)$.

Definition 3 (Product update). Let $(\mathcal{M}, w_0)$ be a state and $(\mathcal{E}, e_0)$ an action, where $\mathcal{M} = (W, R, V)$, $\mathcal{E} = (E, Q, pre, post)$, and $\mathcal{M}, w_0 \models pre(e_0)$. The *product update* of $(\mathcal{M}, w_0)$ with $(\mathcal{E}, e_0)$ is defined as the state $(\mathcal{M}, w_0) \otimes (\mathcal{E}, e_0) = ((W', R', V'), (w_0, e_0))$, where

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \models pre(e)\}$
- $R'_i = \{((w, e), (v, f)) \in W' \times W' \mid wR_iv \text{ and } eQ_if\}$
- $(w, e) \in V'(p)$ iff $post(e) \models p$ or $(\mathcal{M}, w \models p$ and $post(e) \not\models \neg p)$.

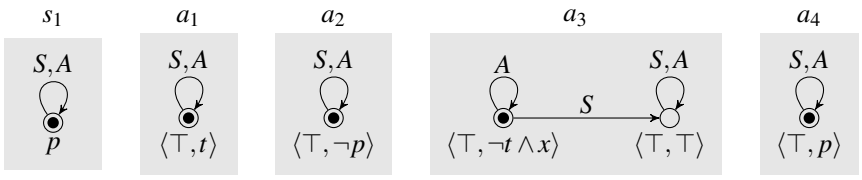


Fig. 5 The states and actions in the DEL formalisation of Sally-Anne.

Example 4. Referring again to Figure 3, we can calculate the product update of $(\mathcal{M}, w_0)$ with $(\mathcal{E}, e_0)$. Intuitively, the calculation works like this. For each event in $\mathcal{E}$, we first find the worlds in $\mathcal{M}$ that satisfies the precondition of the event. Each such matching world-event pair will become a world in the resulting model. Since both e_0 and e_1 have the trivial precondition $\top$, both have their precondition satisfied in the world w_0. This gives us two matching world-event pairs (w_0, e_0) and (w_0, e_1) that will become the worlds of the new model. Now we have to use the postconditions of the events in order to figure out what the labels of these new worlds will be. In (w_0, e_0) we have paired w_0 with e_0. This means that we should take the existing label of w_0 and then update it according to the postcondition of e_0. The label of w_0 is t and the postcondition of e_0 is $\neg t \wedge x$. The postcondition $\neg t \wedge x$ will force t to become false and x to become true, so the label of (w_0, e_0) will be x. The label of (w_0, e_1) is the same as of w_0, since e_0 has the trivial postcondition $\top$. So the updated model $(\mathcal{M}, w_0) \otimes (\mathcal{E}, e_0)$ will have the two worlds $(w_0, e_0){:}x$ and $(w_0, e_1){:}t$. Now we only need to find the edges connecting these two worlds. There will be an A-loop at (w_0, e_0), since there is both an A-loop at w_0 in $\mathcal{M}$ and an A-loop at e_0 in $\mathcal{E}$. Similarly there will be an $\{S, A\}$-loop at (w_0, e_1). Finally, we need to check the edges between (w_0, e_0) and (w_0, e_1). Since there is an S-loop at w_0 and an S-edge from e_0 to e_1, we get an S-edge from (w_0, e_0) to (w_0, e_1). In total, the product update becomes:

$$(\mathcal{M}, w_0) \otimes (\mathcal{E}, e_0) = \quad \underset{(w_0, e_0){:}\,x}{A} \quad \overset{S}{\longrightarrow} \quad \underset{(w_0, e_1){:}\,t}{S, A}$$

Note that the resulting model is isomorphic to $(\mathcal{M}', w_0')$ of Figure 3. Since $(\mathcal{M}, w_0)$ represents the situation before Anne transfers the marble, and $(\mathcal{M}', w_0')$ represents the situation afterwards (cf. Example 2), $(\mathcal{E}, e_0)$ correctly captures the action of transferring the marble in Sally's absence.

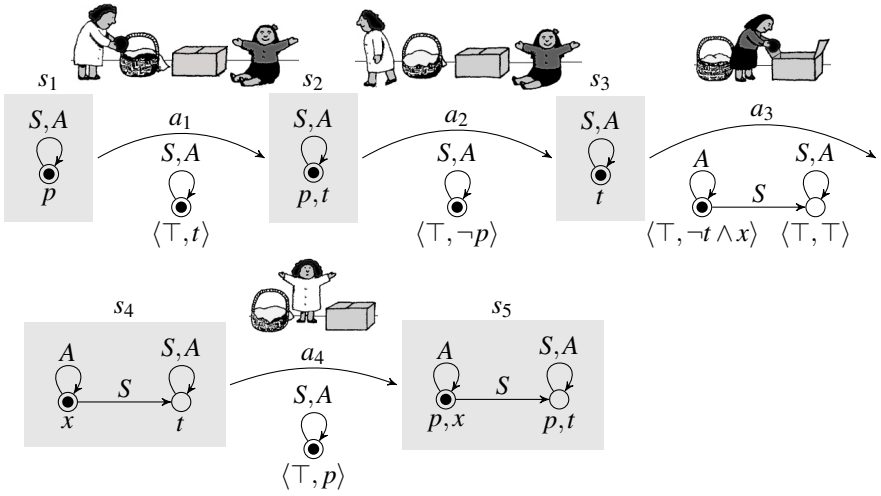


Fig. 6 The DEL-formalisation of the Sally-Anne task.

4 Formalising the Sally-Anne Task in DEL

We now have all the necessary ingredients for our first formalisation of the Sally-Anne task. Consider again the 5 steps of the Sally-Anne story presented in Example 1. The first step, step 0, describes the initial state, whereas the rest, 1–4, describes a sequence of actions. We will now show how to represent step 0 as a state and steps 1–4 as actions. We use the same symbols as in the previous examples, except we add a new atomic proposition p meaning "Sally is *p*resent in the room with Anne". The following 5 step list, corresponding to the list of Example 1, shows the relevant states and actions:

0. Sally is in the room, holding the marble: state s_1 of Figure 5.
1. Sally puts the marble into the basket: action a_1 of Figure 5.
2. Sally leaves the room: action a_2 of Figure 5.
3. Anne transfers the marble to the box: action a_3 of Figure 5.
4. Sally re-enters: action a_4 of Figure 5.

Figure 6 calculates the result of executing the action sequence $a_1,\ldots,a_4$ in s_1, that is, $s_{i+1} = s_i \otimes a_i$ for all $i = 1,\ldots,4$, and hence $s_5 = s_1 \otimes a_1 \otimes \cdots \otimes a_4$. The first two actions, a_1 and a_2, are very simple. As seen from Figure 6, executing a_1 in the initial state s_1 simply adds the proposition t to the actual world (in s_2), signifying that now the marble is in the basket. Executing a_2 in the resulting state s_2 amounts to deleting p from the actual world: in s_3, Sally is no longer present in the room. The action a_3, the most complex one, has already been discussed in Example 3, and in Example 4 we carefully checked that $s_4 = s_3 \otimes a_3$. The final action, a_4, simply adds p to every

world of the model, corresponding to the fact the Sally returns to the room, and this is observed by both agents.

What is important is now of course to check what holds in s_5, the model resulting from executing $a_1, \ldots, a_4$ in s_1. From Figure 6 we can see that

$$s_5 \models \neg t \wedge B_S t,$$

that is, Sally mistakenly believes the marble to be in the basket. Assume an agent presented with steps 0–4 of the original informal story is able to formalise the steps as $s_1, a_1, \ldots, a_4$, and is afterwards asked "where does Sally believe the marble to be". Then that agent can first calculate the final state $s_5 = s_1 \otimes a_1 \otimes \cdots \otimes a_4$ and conclude that $s_5 \models B_S t$ holds. From this the agent can answer "in the basket", hence passing the Sally-Anne test!

5 Extending the DEL formalism

So far so good, or at least it seems that way. But a closer look shows that there are two problems with the DEL-formalisation that need to be addressed. The first is: where do the event models come from? How is an agent supposed to get from the informal steps of the story to the formalisations $s_1, a_1, \ldots, a_4$? It seems to require ingenuity to come up with the right event models to formalise the informal action descriptions, in particular action a_3. Hence the proposed solution does not yet really satisfy the *faithfulness* criterion presented in Section 2.

The second problem with the formalisation can be illustrated by considering a shortened version of the Sally-Anne task where Sally does not leave the room, that is, it only includes the steps 0, 1 and 3 of Example 1. These steps ought to have the same formalisations as before, that is, s_1, a_1 and a_3, respectively. Hence the situation after the shortened Sally-Anne story should correspond to $s_1 \otimes a_1 \otimes a_3$. However, consulting Figure 6 it can be checked that $s_1 \otimes a_1 \otimes a_3 = s_5$ (since a_2 only makes p false, and a_4 makes it true again). Hence, an agent presented with the shortened Sally-Anne story would conclude that

$$s_1 \otimes a_1 \otimes a_3 \models B_S t,$$

implying that Sally ends up believing the marble to be in the basket. This is clearly not correct, since in this version she never left the room!

In the following we will propose an improved formalisation that solves both of these problems. We start out by analysing the source of the second problem, which is in the formalisation of a_3 (see Figure 5). As explained in Example 3, a_3 "encodes an action where the marble is *actually* transferred from the basket to the box, Anne is aware of this, but Sally thinks that nothing happens". All this is clearly not part of step 3 of the story, which simply states "Anne transfers the marble to the box". The problem with a_3, and private announcements and assignments in general, is that it is hardcoded into the event model who observes the action taking place. For some

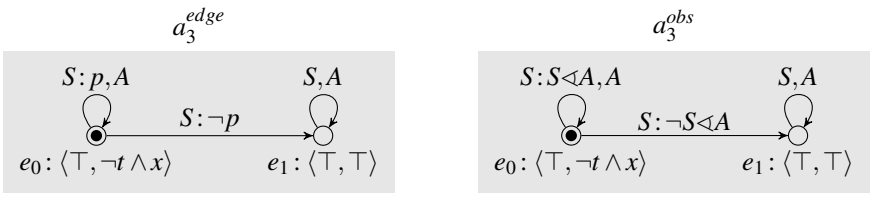


Fig. 7 Two generalised variants of the action a_3 of the Sally-Anne example.

modelling purposes this is sufficient, but in most real-life cases when modelling actions, who observes an action taking place is a feature of the state in which the action is applied, not a feature of the action description itself. This is also the case in the Sally-Anne story: whether Sally observes the action "the marble is transferred" depends on whether she is in the room or not, which is a feature of the state in which the action is applied, not a feature of the action description "the marble is transferred".

Hence the edges of the event model for action a_3 ought to depend on whether Sally is present, that is, whether p holds or not. This leads us to a more general type of event model like a_3^{edge} of Figure 7. Here $S:p$ at the loop of e_0 means that there is an edge here for agent S if p is true: Sally observes the event e_0 if she is present in the room. The other label A at the loop of e_0 simply as usual means that A has an edge here (Anne unconditionally observes the event e_0). Similarly, the label $S:\neg p$ on the edge from e_0 to e_1 means that if Sally is not in the room ($\neg p$) then she thinks that nothing (e_1) happens. This is a new type of event model, called an *edge-conditioned event model*, to be defined formally in the next subsection.

With edge-conditioned event models we can solve the second problem mentioned above. We now have an event model that will behave correctly both if applied in a state where Sally is present (p holds) and in a state where Sally is not present (p doesn't hold). If a_3^{edge} is applied in a state where p holds, from e_0 Sally will only consider e_0 possible (have a loop at e_0), but if p does not hold, from e_0 she will only consider e_1 possible (have an edge from e_0 to e_1). Hence, if p holds she observes the event e_0, otherwise she does not. Using edge-conditioned event models also brings us a step closer to satisfying the first criterion of *faithfulness*. In almost all existing false-belief tasks, all ontic actions have the same structure as a_3^{edge}, and we can hence define a *generic event model* for all such ontic actions (which we will do in Section 5). However, it is still not quite satisfactory to use ad hoc symbols like p to state that a certain agent is present. This leads us to our next new idea.

In addition to our existing set P of propositional symbols, we add to the language a new set of propositional symbols $i\triangleleft j$ (i sees j) for each pair of agents i, j. The intended meaning of $i\triangleleft j$ is that agent i observes the actions of agent j. Using such symbols we can replace the event model a_3^{edge} by a_3^{obs}, see Figure 7. The meaning of the label $S:S\triangleleft A$ at the loop of e_0 is that agent S observes the event e_0 if S

currently sees A ($S \triangleleft A$ is the case). We will now define these new technical constructs formally, and afterwards apply them to give an improved formalisation of the Sally-Anne task.

Edge-Conditioned Event Models

Definition 4 (Edge-conditioned event models). An *edge-conditioned event model* of $\mathcal{L}(P,\mathcal{A})$ is $\mathcal{E} = (E, Q, pre, post)$, where E, *pre* and *post* are defined as for standard event models (Definition 2), and $Q : \mathcal{A} \to (E \times E \to \mathcal{L}(P,\mathcal{A}))$ assigns to each agent i a mapping $Q(i)$ from pairs of events into formulas of $\mathcal{L}(P,\mathcal{A})$. The mapping $Q(i)$ is generally abbreviated Q_i. For $e \in E$, $(\mathcal{E}, e)$ is called an *edge-conditioned action* of $\mathcal{L}(P,\mathcal{A})$.

For standard event models (Definition 2), eQ_if means that event f is accessible from event e by agent i, and we include i in the label of the edge from e to f in the graph of the event model. In edge-conditioned event models, accessibility has become conditioned by a formula: $Q_i(e,f) = \phi$ means that f is accessible from e by i *under condition* ϕ. When $Q_i(e,f) = \phi$, we include $i : \phi$ in the label of the edge from e to f in the graph of the event model. There are two exceptions to this: when $Q_i(e,f) = \bot$ we do not include i in the label of (e,f), and when $Q_i(e,f) = \top$ we simply put i in the label of (e,f) instead of $i : \top$. We already saw an example of such an edge-conditioned event model: a_3^{edge} of Figure 7. We also have to generalise the notion of product update:

Definition 5 (Edge-conditioned product update). Let a state $(\mathcal{M}, w_0)$ and an edge-conditioned action $(\mathcal{E}, e_0)$ be given with $\mathcal{M} = (W, R, V)$, $\mathcal{E} = (E, Q, pre, post)$, and $\mathcal{M}, w_0 \models pre(e_0)$. The *product update* of $(\mathcal{M}, w_0)$ with $(\mathcal{E}, e_0)$ is defined as the state $(\mathcal{M}, w_0) \otimes (\mathcal{E}, e_0) = ((W', R', V'), (w_0, e_0))$, where W' and V' are defined as in the standard product update (Definition 3) and $R_i' = \{((w,e),(v,f)) \in W' \times W' \mid wR_iv$ and $\mathcal{M}, w \models Q_i(e,f)\}$.

The only difference to the standard product update is that the R_i' relations have become parametrised by the $Q_i(e,f)$ formulas. There is an i-edge from a world-event pair (w,e) to a world-event pair (v,f) iff there is an i-edge from w to v in the epistemic model, and the condition $Q_i(e,f)$ for having an edge from e to f in the event model is true in w.

Note that edge-conditioned event models naturally generalise standard event models: Any standard event model $\mathcal{E} = (E, Q, pre, post)$ can be equivalently represented as an edge-conditioned event model $\mathcal{E}' = (E, Q', pre, post)$ by simply letting $Q_i'(e,f) = \top$ for all $(e,f) \in Q_i$ and $Q_i'(e,f) = \bot$ otherwise. It is easy to verify that we then for any state $(\mathcal{M}, w_0)$ have

$$(\mathcal{M}, w_0) \otimes (\mathcal{E}', e_0) = (\mathcal{M}, w_0) \otimes (\mathcal{E}, e_0).$$

It can be shown that, conversely, any edge-conditioned event model induces a standard event model in a canonical way, but the induced standard event model might be exponentially bigger. In technical terms, it can be shown that edge-conditioned event models are exponentially more succinct than standard event models (we will prove this and other interesting properties of edge-conditioned event models in a future paper). In particular, our generic event models for ontic actions and observability change (to be presented in Section 5) are going to consist of 2 events each, whereas the same actions using only standard event models would contain $2^{n-1}+1$ events, where n is the number of agents!

Observability Propositions

We now define a new language $\mathcal{L}^{obs}(P,\mathcal{A})$ extending $\mathcal{L}(P,\mathcal{A})$ by the addition of *observability propositions* on the form $i \lhd j$:

$$\phi ::= p \mid i \lhd j \mid \neg\phi \mid \phi \wedge \phi \mid B_i\phi,$$

where $p \in P$ and $i,j \in \mathcal{A}$. As noted above, the intended meaning of $i \lhd j$ is that "agent i observes all actions performed by agent j". We have also included the reflexive propositions $i \lhd i$, so we can represent a situation in which an agent i is not even observing its own actions (a "drunk agent") by $\neg i \lhd i$. However, in this paper we will generally assume our models to be *normal*, which we define to mean that $i \lhd i$ holds in all worlds of the model for all agents. For simplicity, we will not include $i \lhd i$ in the label of all worlds, so the reader has to remember that these formulas are always implicitly taken to be true everywhere. In the expression $i \lhd j$ we call i the *observer*. Given a formula ϕ, we use $obs(\phi)$ to denote the set of agents occurring (positively or negatively) as observers in ϕ, that is, $obs(\phi) = \{i \mid i \lhd j \text{ is a subformula of } \phi \text{ for some } j\}$. For instance we have $obs(i \lhd j \wedge \neg k \lhd l) = \{i,k\}$ (note that k is in the set even though the formula $k \lhd l$ occurs negated).

The idea of introducing observability propositions in the context of DEL was first introduced in van Ditmarsch et al. (2013). They, however, only use a simpler type of proposition h_i with the intended meaning "agent i observes *all* actions" (agent i is in a state of paying attention to everything that happens). Here we need something more fine-grained, in particular for our later formalisation of the chocolate task (Section 6) where we need to be able to represent that an agent i is observing the actions of an agent j without j observing the actions of i.

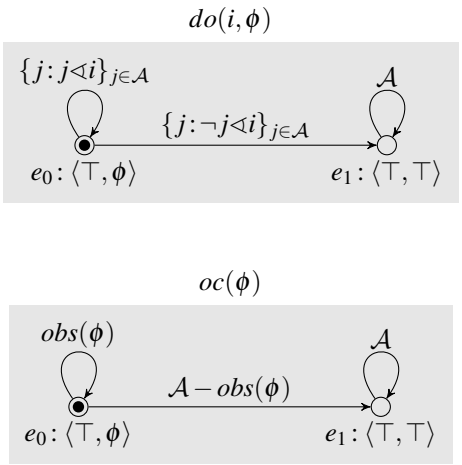


Fig. 8 The edge-conditioned actions $do(i,\phi)$ and $oc(\phi)$.

Ontic Actions and Observability Change

The previous definitions of edge-conditioned event models and product update extend to the language $\mathcal{L}^{obs}(P,\mathcal{A})$ in the obvious way (after all, we only added some additional atomic propositions). We can now finally define two generic types of edge-conditioned actions that are sufficient to formalise a number of different false-belief tasks of varying belief-attribution order. The first action type is an ontic action $do(i,\phi)$: agent i makes ϕ true. Step 1 of the Sally-Anne task is for instance going to be formalised by $do(S,t)$: Sally makes t true. The second is an *o*bservability *c*hanging action $oc(\phi)$ for changing who observes who. For instance step 2 of the Sally-Anne task where Sally leaves the room is going to be formalised by $oc(\neg S\lhd A \wedge \neg A\lhd S)$: Sally stops observing Anne ($\neg S\lhd A$), and Anne stops observing Sally ($\neg A\lhd S$).

Definition 6. We define the following edge-conditioned actions on $\mathcal{L}^{obs}(P,\mathcal{A})$.

- $do(i,\phi)$: for each agent i and each conjunction of propositional literals ϕ, this is the *ontic action* shown at the top of Figure 8.
- $oc(\phi)$: for each conjunction of observability literals (observability propositions and their negations), this is the *observability changing action* shown at the bottom of Figure 8.

These new actions need a little explanation. Consider first $do(i,\phi)$. As mentioned, this is an action where agent i makes ϕ true (since the actual event e_0 has postcondition ϕ). From the label at the loop of e_0 we can see that the agents who observe the action taking place, and hence come to believe ϕ, are all the agents who is currently observing agent i (all the agents j for which $j\lhd i$ is true). The agents who are not observing i will think that nothing happens (the label $\{j : \neg j\lhd i\}_{j\in\mathcal{A}}$ on the edge to

e_1). This also explains the title of the paper, "Seeing is believing": If agent j *sees* agent i, $j \triangleleft i$, then j comes to *believe* any formula ϕ that i brings about.

The action $oc(\phi)$ follows the same principle (note that the two event models only differ in their edge labels). Looking at the label of the loop at e_0, we can see that the agents observing the observability change are those whose observer status is affected by the action. This is not the only reasonable way to define an observability changing action. An alternative could be to say that those who observe the action are those in $obs(\phi)$ whose observer status is affected *and* anyone observing at least one of the agents in $obs(\phi)$. That is, we could make the label of the loop at e_0 be $\{j : \bigvee_{k \in obs(\phi)} j \triangleleft k\}_{j \in \mathcal{A}}$ instead. The intuition here would be that if i is currently observing j, and j either starts or stops to observe k, then i will also observe this change. This would be the natural way of formalising things if we think of the action "j stops observing k" as an action that j performs, since if i is currently observing j, then i is supposed to observe *any* action performed by j. However, one could conversely argue that even if an agent i observes all actions of an agent j, it might not necessarily imply that agent i can observe it whenever there is a change in what j pays attention to. If you are in the same room as your spouse, and you are paying attention to him, you will notice all his ontic (world-changing) actions, but not necessarily notice when he starts and stops paying attention to you. For the purposes of this paper, either way of formalising $oc(\phi)$ will work, and for simplicity we have chosen the one with the simpler edge-conditions.

Joint Attention

Note that *joint attention* (Tomasello 1995; Lorini et al. 2005; Bolander et al. 2015) in a group of agents $\mathcal{B} \subseteq \mathcal{A}$ can be achieved by the action $oc(\bigwedge_{i,j \in \mathcal{B}} i \triangleleft j)$, which we will abbreviate $jointAtt(\mathcal{B})$. Executing this action will create a situation after which any action performed by any of the agents in $\mathcal{B}$ will be jointly observed by all agents in $\mathcal{B}$ and hence lead to common belief in $\mathcal{B}$ of the action effects. More precisely, by consulting the event models of Figure 8 it is easy to show that for any state $(\mathcal{M}, w_0)$ and any agent $i \in \mathcal{B}$ we have

$$(\mathcal{M}, w_0) \otimes jointAtt(\mathcal{B}) \otimes do(i, \phi) \models C_{\mathcal{B}} \phi.$$

Seriality and announcements

In general, seriality is not preserved under product update, that is, the product update of a serial epistemic model with a serial event model might still produce a non-serial resulting model (see e.g. Aucher (2012)). However, since both $do(i, \phi)$ and $oc(\phi)$ only have trivial preconditions (all preconditions being $\top$), any sequence of updates of a serial epistemic model with such actions is going to be serial (Aucher 2012). This would no longer hold if we chose to include a standard *announcement action* in our framework, as previously noted. We chose not to include announcements for the

following reasons: 1) to save space; 2) since announcements are not part of the false-belief tasks we are interested in studying in this paper; 3) to ensure the preservation of seriality. It would be very simple to add announcements, though: simply take the event model for $do(i,\phi)$ and put ϕ in the precondition instead of the postcondition of e_0, similar to the distinction between private assignments and announcements in Figure 4. Note that $do(i,\phi)$ is indeed a straightforward generalisation of the private assignment of Figure 4, where we have simply replaced the outgoing fixed-label edges of e_0 by edge-conditioned labels.

Agency

In standard DEL there is no explicit notion of agency, that is, an action simply happens without any need to say who did it. But in our *do* action we need to include the agent performing it as a parameter, since what will be observed by the other agents depends on it.

6 New Formalisations of False-Belief Tasks

Example 5 (Formalising the Sally-Anne task). Given the generic actions from the previous section, it is now quite straightforward to provide a new formalisation of the Sally-Anne task using these actions:

0. Sally is in the room with Anne, holding the marble: state $s_1 =$ 

1. Sally puts the marble into the basket: $a_1 = do(S,t)$.
2. Sally leaves the room: $a_2 = oc(\neg S \lhd A \wedge \neg A \lhd S)$.
3. Anne transfers the marble to the box: $a_3 = do(A, \neg t \wedge x)$.
4. Sally re-enters: $a_4 = oc(S \lhd A \wedge A \lhd S)$.

Note that we no longer use the atomic proposition p, as we now have a more generic way to deal with observability through our observability propositions. Note also that in step 0 we could have chosen to start with an initial state satisfying no propositions, and then have created joint attention by first executing $jointAtt(\{S,A\})$ in this state. This would generate the state s_1 above (recall that we are omitting the reflexive observability propositions $i \lhd i$ in figures).

Similar to the previous formalisation in Section 4, it can now be checked that

$$s_1 \otimes a_1 \otimes \cdots \otimes a_4 \models B_S t,$$

hence again the formalisation gives the right answer to the Sally-Anne test. We should also note that now we have

$$s_1 \otimes a_1 \otimes a_3 \models B_S x,$$

so if Sally does not leave the room, she will not get a false belief. Thus we have successfully solved the problem of the shortened Sally-Anne task that was discussed in the beginning of Section 5. We will not show the detailed calculations, as we will save that for the next example, which formalises a more complex false-belief task.

Example 6 (Formalising the second-order chocolate task). We now consider a compact version of the *second-order chocolate task* presented in Flobbe et al. (2008); Arslan et al. (2013). It has the following steps:

0. John and Mary are in a room. There is a chocolate bar in the room.
1. John puts the chocolate into a drawer.
2. John leaves the room.
3. John starts peeking into the room through the window, without Mary seeing.
4. Mary transfers the chocolate from the drawer to a box.

The child taking the test is now asked "where does Mary believe that John believes the chocolate to be?" It is a second-order task since this question concerns second-order belief attribution (Mary's beliefs about John's beliefs). The correct answer is "in the drawer", since Mary is not aware that John was peeking while she moved the chocolate. It is immediate that step 1 and 4 above are ontic actions, and steps 2 and 3 are observability changing actions. Let us use atomic propositions d for the "the chocolate is in the *d*rawer" and x for "the chocolate is in the bo*x*." We use agent symbols J for John and M for Mary. Step 1, "John puts the chocolate into the drawer", must then be the ontic action $do(J,d)$. Step 2, "John leaves the room", must be the observability change $oc(\neg J{\lhd}M \wedge \neg M{\lhd}J)$ (John stops observing Mary and Mary stops observing John). Step 3 is again an observability change, but this time it is simply $oc(J{\lhd}M)$: John starts observing Mary. Finally, step 4 is the ontic action $do(M, \neg d \wedge x)$.

Figure 9 calculates the result of executing the action sequence of steps 1–4 in the initial state described by step 0. The actions in the figure show the applied instances of $do(i,\phi)$ and $oc(\phi)$ calculated from Figure 8. To simplify, we have replaced labels of the form $j : j{\lhd}j$ by j, and omitted labels of the from $j : \neg j{\lhd}j$. This can be done as we are only working with normal models ($i{\lhd}i$ is universally true for all i). To simplify further, in states (actions) we have omitted worlds (events) that are not accessible from the actual world (event) by any sequence of agents, that is, we have deleted worlds (events) that are not in the same connected component as the actual world (event). This clearly does not change what is true in the actual world (event) of that state (action).

Before going into the detailed calculations of Figure 9, let us have a look at the resulting model s_5. This is the model in which it should be checked where Mary believes John believes the chocolate to be. Clearly we have

$$s_5 \models B_M B_J d,$$

so the agent's answer will be "in the drawer", hence passing the false-belief test. But s_5 can do more than just answer this question, in fact it is a full description of

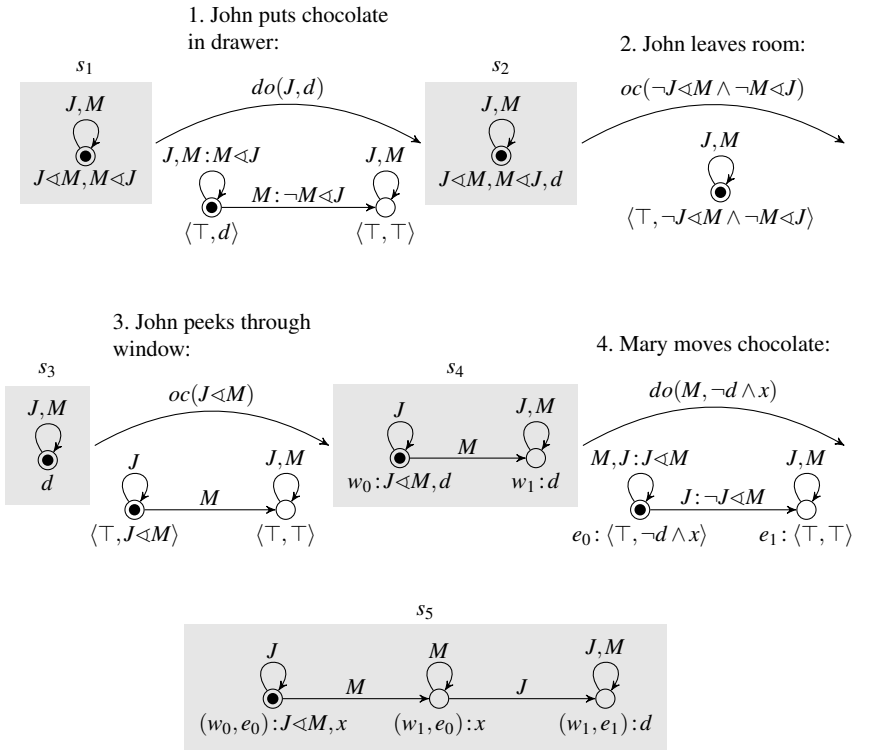


Fig. 9 The DEL-formalisation of the second-order chocolate task

the final situation, including all beliefs to arbitrary order. Concerning observability, we can for instance see that

$$s_5 \models J \lhd M \wedge B_M \neg J \lhd M \wedge B_J B_M \neg J \lhd M :$$

John sees Mary, Mary believes he does not, and John knows this.[3] We can also imagine a third-order version of the task, where the question is "Where does John believe that Mary believes that John believes the chocolate to be", and by consulting s_5 we immediately get the answer "in the drawer":

$$s_5 \models B_J B_M B_J d.$$

The most interesting part of the calculation in Figure 9 is the last step, $s_5 = s_4 \otimes do(M, \neg d \wedge x)$, so we will explain this in more detail. Calculating the product $s_4 \otimes$

[3] Strictly speaking, we should say "John believes this" instead of "John knows this", since our modality is a belief modality. To improve readability, we however allow ourselves to slightly abuse the term and use "knows" instead of "believes" when the formula believed is also true.

$do(M, \neg d \wedge x)$ follows the same strategy as in Example 4. First we find the matching world-event pairs which, in this case, are all four world-event combinations (w_0, e_0), (w_0, e_1), (w_1, e_0) and (w_1, e_1), since both e_0 and e_1 have trivial preconditions. See Figure 9 where $do(M, \neg d \wedge x)$ is the event model of step 4. The world-event pair (w_0, e_1) is not shown in s_5 in Figure 9, as it turns out not to be accessible from the actual world (w_0, e_0). In the world-event pairs containing e_0, the postcondition of e_0 is enforced, that is, d is made false and x true. The other world-event pairs simply inherit their label from the first element of the pair. Hence the four worlds of the resulting model s_5 are $(w_0, e_0) : J \lhd M, x$; $(w_0, e_1) : J \lhd M; d$, $(w_1, e_0) : x$; $(w_1, e_1) : d$. Now for the interesting part, the edges. At (w_0, e_0) we get a J-loop, since there is J-loop at w_0 and *the condition for having a J-loop at e_0 is $J \lhd M$, which is satisfied in w_0*. This should be contrasted with the situation at (w_1, e_0): Here we also have a J-loop at the world of the pair, w_1, but now *the condition $J \lhd M$ for having a J-loop at the event of the pair is* not *satisfied in the world of the pair*. At (w_1, e_0) we hence only get an M-loop (since both w_1 and e_0 unconditionally have such a loop). We leave the calculation of the rest of the edges to the (enthusiastic) reader.

Let us try to analyse the formalisation of the second-order chocolate task a bit deeper. For $n \geq 1$, we say that $\phi \in \mathcal{L}(P, \mathcal{A})$ is an *nth-order false belief* in the state s if for some sequence $i_1, \ldots, i_n \in \mathcal{A}$ the following holds:

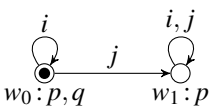
$$s \models \neg\phi \wedge E\neg\phi \wedge E^2\neg\phi \wedge \cdots \wedge E^{n-1}\neg\phi \wedge B_{i_1} B_{i_2} \cdots B_{i_n} \phi$$

That is, ϕ is false, everybody beliefs this to depth $n-1$, but agent i_1 falsely believes that agent i_2 believes that... agent i_n believes that ϕ is true. Note that there is a second-order false belief concerning d (and x) in s_5 of Figure 9, since

$$s_5 \models \neg d \wedge E\neg d \wedge B_M B_J d.$$

In s_4, there are no false beliefs about d. In fact, d is even common belief in s_4: $s_4 \models C_{\{M,J\}} d$. To get from common belief of d in s_4 to a second-order false belief concerning d in s_5, we only had to apply an instance of a generic edge-conditioned action with 2 events ($do(M, \neg d \wedge x)$). This situation is much better than what can be achieved with standard actions (standard event models). The following propositions show that there is no standard action with 2 events that can create a second-order false belief concerning a proposition p from a state in which p is common belief. The first proposition considers only product updates of the state s_4 of Figure 9. The second proposition generalises the result.

Proposition 1. *Let s be a state isomorphic to s_4 of Figure 9, that is a state of the following form, with $\mathcal{A} = \{i, j\}$:*

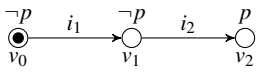


Let a be a standard action (standard event model, as defined in Definition 2) with only two events, and assume $s \otimes a$ is serial. Then p is not a second-order false belief in $s \otimes a$.

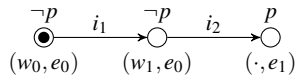
Proof. Let v_0 denote the actual world of $s \otimes a$. Let e_0 and e_1 denote the two events of a, with e_0 being the actual event. Then $v_0 = (w_0, e_0)$. We will make a proof by contradiction, that is, we first assume p is a second-order false belief. This means that for some choice of $(i_1, i_2) \in \{(i, j), (j, i)\}$ we have:

$$s \otimes a \models \neg p \wedge E\neg p \wedge B_{i_1} B_{i_2} p.$$

We can conclude that $s \otimes a$ must contain a path of the following form:

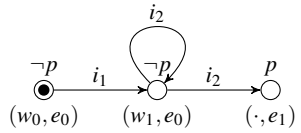


where p is false in v_0 and v_1, and true in v_2. Since $v_0 = (w_0, e_0)$ and p is true in w_0, e_0 must be an event that makes p false, that is, $\neg p$ is a conjunct of $post(e_0)$. Since p is true in v_2, v_2 can then not be a world-event pair of the form $(\cdot, e_0)$. It must hence be of the form $(\cdot, e_1)$, and e_1 must therefore be an event that *does not* make p false, that is $\neg p$ is *not* a conjunct of $post(e_1)$. Since p is false in v_1, it follows that v_1 must be a world-event pair of the form $(\cdot, e_0)$. Hence $v_1 = (w_0, e_0)$ or $v_1 = (w_1, e_0)$. We can immediately eliminate the possibility $v_1 = (w_0, e_0)$, since in that case we would have $v_1 = v_0$, and thus (v_1, v_2) would be an i_2-edge from the actual world of $s \otimes a$ to a world where p is true, contradicting that $E\neg p$ holds in $s \otimes a$. Hence $v_1 = (w_1, e_0)$. Thus the path above has the following form:



From the i_1-edge from (w_0, e_0) to (w_1, e_0), we can conclude that in a there is an i_1-loop at e_0 (cf. the definition of product update). Similarly, from the i_2-edge from (w_1, e_0) to $(\cdot, e_1)$, we can conclude that there is an i_2-edge from e_0 to e_1 in a.

Due to the seriality of $s \otimes a$, there also has to be an outgoing i_2-edge from (w_0, e_0). This edge must end in a world of the form $(\cdot, e_0)$, since $E\neg p$ holds in $s \otimes a$. Hence we can conclude that there must be an i_2-loop at e_0 in a. Since there is also an i_2-loop at w_1 in s, we can conclude that there must be an i_2-loop at (w_1, e_0) in $s \otimes a$. That is, $s \otimes a$ must contain a submodel of the following form:



This immediately contradicts the original assumption that $s \otimes a \models B_{i_1} B_{i_2} p$, and the proof is complete.

Proposition 2. *Let s be a state and a a standard action (standard event model, as defined in Definition 2) such that, for some $p \in P$,*

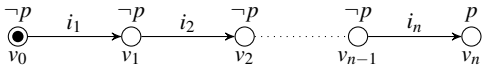
1. *p is common belief in s, that is, $s \models C_{\mathcal{A}} p$.*
2. *s is functional, that is, for each world w of s and each agent $i \in \mathcal{A}$, there is at most one world w' with wR_iw'.*[4]
3. *a contains only two events.*
4. *$s \otimes a$ is serial.*

For all $n > 1$, if p is an nth-order false belief in $s \otimes a$, then some formula ϕ is an nth-order false belief in s.

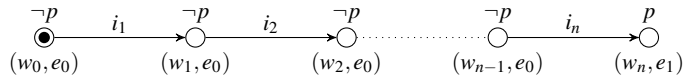
Proof. Let v_0 denote the actual world of $s \otimes a$. Let e_0 and e_1 denote the two events of a, with e_0 being the actual event. Let w_0 denote the actual world of s. Then $v_0 = (w_0, e_0)$. Assume p is an nth-order false belief for some $n > 1$. Then for some sequence $i_1, i_2, \ldots, i_n \in \mathcal{A}$ we have

$$s \otimes a \models \neg p \wedge E\neg p \wedge E^2\neg p \wedge \cdots \wedge E^{n-1}\neg p \wedge B_{i_1}B_{i_2}\cdots B_{i_n}p.$$

Hence there must exist a path in $s \otimes a$ of the following form



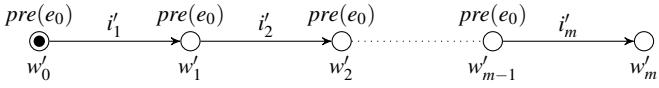
where p is false in v_i, $i < n$, and true in v_n. Since $s \models C_{\mathcal{A}} p$, p is true in w_0 of s. Since p is false in $v_0 = (w_0, e_0)$, e_0 must be an event that makes p false, that is, $\neg p$ is a conjunct of $post(e_0)$. Since p is true in v_n, v_n can not be a world-event pair of the form $(\cdot, e_0)$. It must hence be of the form $(\cdot, e_1)$, and e_1 must therefore be an event that *does not* make p false, that is, $\neg p$ is *not* a conjunct of $post(e_1)$. Since p is false in all of v_i with $1 < i < n$, all of these must be world-event pairs of the form $(\cdot, e_0)$. Hence the path has the following form:



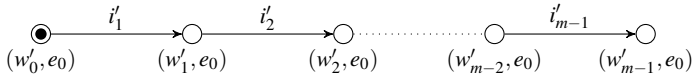
We now show that a contains an i-loop at e_0 for each $i \in \mathcal{A}$. Since $s \otimes a$ is serial, for all $i \in \mathcal{A}$, there must be an outgoing i-edge from (w_0, e_0) to some world v' in $s \otimes a$. The world v' can not have the form $(\cdot, e_1)$, as p would then be true in v', which contradicts that p is an nth-order false belief with $n > 1$ (recall that p is common belief in s, and e_1 is an event that does not make p false). Hence v' must have the form $(\cdot, e_0)$. We now have that $s \otimes a$ contains an i-edge from (w_0, e_0) to a world-event pair of the form $(\cdot, e_0)$, from which we can conclude that a contains an i-loop at e_0.

[4] Note that all states considered so far in this paper have been functional, and that the property of being functional is preserved under any sequence of updates with $do(i, \phi)$ and $oc(\phi)$ actions.

We wish to show that $\neg pre(e_0)$ is an nth-order false belief in s, which will complete the proof. First we prove $s \models E^m pre(e_0)$ for all $m < n$. To this end, let there be given a path

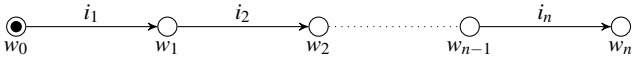


in s with $w'_0 = w_0$ and such that w'_k satisfies $pre(e_0)$ for all $k < m$. We need to show that w'_m also satisfies $pre(e_0)$. Since each w'_k with $k < m$ satisfies $pre(e_0)$, and since we have already shown that a contains an i-loop at e_0 for every $i \in \mathcal{A}$, we can conclude that $s \otimes a$ must contain a path of the following form:

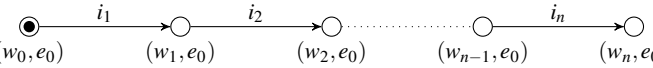


Since $s \otimes a$ is serial, (w'_{m-1}, e_0) must have an i'_m-successor. This successor must have the form (w''_m, e_0), since $s \otimes a \models E^m \neg p$. Then w''_m must also be an i'_m-successor of w'_{m-1} in s, and since s is functional, we get $w''_m = w'_m$. Since $(w''_m, e_0) = (w'_m, e_0)$ is a world of $s \otimes a$, we must have that w'_m satisfies $pre(e_0)$, as required.

We now have proven that $s \models E^m pre(e_0)$ for all $m < n$. The only thing left is to prove $s \models B_{i_1} B_{i_2} \cdots B_{i_n} \neg pre(e_0)$. Consider any path in s of the following form:



We need to prove that w_n satisfies $\neg pre(e_0)$. To obtain a contradiction, assume the opposite. Combined with what we have already shown, we must now have that w_m satisfies $pre(e_0)$ for all $m \leq n$, and since e_0 contains an i-loop for every $i \in \mathcal{A}$, $s \otimes a$ must contain the following path

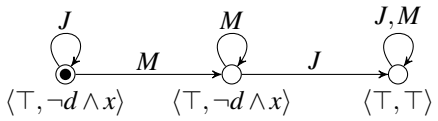


Since e_0 makes p false, all these worlds satisfy $\neg p$, and we hence have a contradiction with our original assumption that $s \otimes a \models B_{i_1} B_{i_2} \cdots B_{i_n} p$.

Both of the propositions above immediately implies that there is no standard event model a with two events such that

$$s_5 = s_4 \otimes a.$$

Since $s_5 = s_4 \otimes do(M, \neg d \wedge x)$, this is a clear difference to the situation with edge-conditioned event models. The result might seem minor, but it is significant for our *faithfulness* aim for the following reason. The propositions imply that the smallest standard event model that can produce s_5 from s_4 is this:



The problem with this event model is that it is already a 'second-order model' that fully encodes the structure of the model s_5 we wish to obtain. Hence if we had to formalise the second-order chocolate task using standard event models, we would have to formalise the step "Mary moves the chocolate" as this event model that already fully encodes the final structure achieved at the end of the story! This would certainly be very far from achieving the *faithfulness* criterion introduced in Section 2. So indeed the edge-conditioned event models make a real difference to the formalisation of false-belief tasks. The fact that we can build a single generic edge-conditioned event model, $do(i,\phi)$, with only two events, that is both appropriate to create a first-order false belief about t from common belief of t in the Sally-Anne task and create a second-order false belief about d from common belief of d in the second-order chocolate task, we find to be a significant step in the direction of achieving *faithful* and *generic* formalisations of false-belief tasks.

7 Conclusion, related work and future work

In this paper we have shown how to formalise two false-belief tasks—a first- and a second-order one—in an extension of dynamic epistemic logic. In the end, we were able to express the formalisations rather compactly as a simple initial state followed by a sequence of generic actions:

- **Sally-Anne task**:

$$\underset{S\lhd A, A\lhd S}{\overset{S,A}{\circlearrowright}}, do(S,t), oc(\neg S\lhd A\wedge\neg A\lhd S), do(A,\neg t\wedge x), oc(S\lhd A\wedge A\lhd S).$$

- **Chocolate task**:

$$\underset{J\lhd M, M\lhd J}{\overset{J,M}{\circlearrowright}}, do(J,d), oc(\neg J\lhd M\wedge\neg M\lhd J), oc(J\lhd M), do(M,\neg d\wedge x).$$

We started out expressing two overall criteria for our formalisations of false-belief tasks: robustness and faithfulness. To be robust, the formalism should be able to formalise false-belief tasks of arbitrary order. We claim to have such robustness in our current formalism, but proving it formally is future work. Nevertheless, we *have* been able to show that we could go from a formalisation of a first-order false-belief task to a second-order one at no extra cost, which as discussed above is *not* the case

in standard DEL (and not in most other frameworks either). To have faithfulness, we required that it should be relatively straightforward to get from the informal action descriptions of the false-belief task to the corresponding formalised actions. We believe we have taken a big step closer towards achieving this. If the (semi-)informal description says "agent i makes ϕ true" it is our action $do(i,\phi)$. If the informal description says, e.g., "now agent i starts observing j without agent j noticing" it is $oc(i \lhd j)$. The formalisation step can of course still not be fully automated, but we are much closer than if we just had to build all the relevant event models from scratch, which was where this paper started.

There is of course also a limit to the types of false-belief tasks that can be dealt with using only *do* and *oc*. In particular, a lot of the existing false-belief tasks involve untruthful announcements such as the 'ice-cream task' in Perner and Wimmer (1985), the 'birthday puppy task' in Sullivan et al. (1994) and the 'clown-in-the-park task' in Wahl and Spada (2000). These can not be dealt with in the current framework. To be able to deal with untruthful announcements and the revision of false beliefs, we need another type of model called plausibility models (Baltag and Smets 2008). We plan to show how these models can be used to formalise the aforementioned false-belief tasks in a future paper.

In our approach, observability amounts to 'who sees who', that is, it is a relation between agents. Other approaches to modelling observability can be found in e.g. Brenner and Nebel (2009); Hoek et al. (2011); Baral et al. (2012); van Ditmarsch et al. (2013); Herzig et al. (2015); Bolander et al. (2015). In these approaches, observability is instead connected either to propositions (Brenner and Nebel 2009; Hoek et al. 2011; Herzig et al. 2015), particular actions (Baral et al. 2012) or *all* actions (van Ditmarsch et al. 2013; Bolander et al. 2015). The paper Seligman et al. (2013) uses a similar approach to observability as we do, but in a more complex 2-dimensional dynamic epistemic logic. In the papers by Brenner and Nebel (2009) and Baral et al. (2012), observability is encoded using axioms instead of being encoded into the states as we do. For us, it is very important to encode observability directly into the states to be able to deal with higher-order observability ('Mary does not see John seeing her').

Even though edge-conditioned event models is an original idea of this paper, they are close in spirit to the *generalised arrow updates* of Kooi and Renne (2011). However, arrow updates are rather an *alternative* to event models, whereas our edge-conditioned event models is a straightforward *generalisation* of event models. Furthermore, arrow updates are purely epistemic (without postconditions), and would hence not be able to represent the ontic actions of the false-belief tasks. Generalised arrow updates are however more general than edge-conditioned event models along a different dimension. We only employ what Kooi and Renne (2011) call *source conditions*: To check whether an edge (w,v) of a state $(\mathcal{M},w_0)$ becomes an edge $((w,e),(v,f))$ in the product update $(\mathcal{M},w_0)\otimes(\mathcal{E},e_0)$, we check the truth-value of the edge condition $Q_i(e,f)$ in the *source* w of the edge (w,v). In generalised arrow updates, there is both a source condition, on w, and a target condition, on v. It would be a simple matter to extend our edge-conditioned event models to also allow target conditions, and hence bring edge-conditioned event models and generalised arrow

updates closer together. However, as target conditions were not relevant for the type of actions we wanted to formalise in this paper, we chose to keep things simple and leave them out.

Solving false-belief tasks using DEL as we do in this paper is part of a larger research effort in *epistemic planning*: combining automated planning with DEL to integrate higher-order social cognition into intelligent planning agents (Bolander and Andersen 2011; Andersen et al. 2012). Combining the ideas of the aforementioned papers with the ideas of this paper will allow us to devise algorithms not only for *analysing* false beliefs (as is done in the false-belief tasks), but also for *synthesising* them. It could e.g. be that Anne plans to deceive Sally by asking her to go outside and then she moves the marble meanwhile. This is a case of epistemic planning where the goal is to achieve a state where Sally does not know the location of the marble.

Acknowledgments

This paper is an extended and revised version of a paper presented at the first European Conference on Social Intelligence (ECSI) in Barcelona, 2014 (Bolander 2014). The author wishes to thank Patrick Blackburn and the anonymous reviewers of both the original submission and this extended version for their encouraging feedback and their many helpful comments and suggestions. The author is also very grateful to the editors of this volume for the invitation to contribute. The author acknowledges support from the Carlsberg Foundation (Center for Information and Bubble Studies, CIBS).

References

Mikkel Birkegaard Andersen, Thomas Bolander, and Martin Holm Jensen. Conditional epistemic planning. *Lecture Notes in Artificial Intelligence*, 7519:94–106, 2012. Proceedings of JELIA 2012.

Konstantine Arkoudas and Selmer Bringsjord. Toward formalizing common-sense psychology: An analysis of the false-belief task. In Tu Bao Ho and Zhi-Hua Zhou, editors, *PRICAI*, volume 5351 of *Lecture Notes in Computer Science*, pages 17–29. Springer, 2008.

Burcu Arslan, Niels Taatgen, and Rineke Verbrugge. Modeling developmental transitions in reasoning about false beliefs of others. In *Proc. of the 12th International Conference on Cognitive Modelling*, 2013.

Guillaume Aucher. Private announcement and belief expansion: an internal perspective. *Journal of Logic and Computation*, 22(3):451–479, 2012.

Alexandru Baltag and Larry Moss. Logic for epistemic programs. *Synthese*, 139 (2):165–224, 2004.

Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. In Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT7)*, volume 3 of *Texts in Logic and Games*, pages 13–60. Amsterdam University Press, 2008.

Alexandru Baltag, Lawrence S. Moss, and Slawomir Solecki. The logic of public announcements and common knowledge and private suspicions. In Itzhak Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-98)*, pages 43–56. Morgan Kaufmann, 1998.

Chitta Baral, Gregory Gelfond, Enrico Pontelli, and Tran Cao Son. An action language for reasoning about beliefs in multi-agent domains. In *Proceedings of the 14th International Workshop on Non-Monotonic Reasoning*, volume 4, 2012.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. Does the autistic child have a theory of mind? *Cognition*, 21(1):37–46, 1985.

Thomas Bolander. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In Andreas Herzig and Emiliano Lorini, editors, *Proceedings of the European Conference on Social Intelligence (ECSI-2014)*, volume 1283 of *CEUR Workshop Proceedings*, pages 87–107. CEUR-WS.org, 2014.

Thomas Bolander and Mikkel Birkegaard Andersen. Epistemic planning for single- and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21:9–34, 2011.

Thomas Bolander, Hans van Ditmarsch, Andreas Herzig, Emiliano Lorini, Pere Pardo, and François Schwarzentruber. Announcements to attentive agents. *Journal of Logic, Language and Information*, pages 1–35, 2015.

Torben Braüner. Hybrid-logical reasoning in false-belief tasks. In B.C. Schipper, editor, *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 186–195, 2013.

Cynthia Breazeal, Jesse Gray, and Matt Berin. Mindreading as a foundational skill for socially intelligent robots. In *Robotics Research*, pages 383–394. Springer, 2011.

Michael Brenner and Bernhard Nebel. Continual planning and acting in dynamic multiagent environments. *Autonomous Agents and Multi-Agent Systems*, 19(3): 297–331, 2009.

Liesbeth Flobbe, Rineke Verbrugge, Petra Hendriks, and Irene Krämer. Childrens application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17(4):417–442, 2008. Special issue on formal models for real people, edited by M. Counihan.

Uta Frith. *Autism: Explaining the enigma*. Wiley Online Library, 1989.

Sujata Ghosh, Ben Meijering, and Rineke Verbrugge. Strategic reasoning: Building cognitive models from logical formulas. *Journal of Logic, Language and Information*, 23(1):1–29, 2014.

Andreas Herzig, Emiliano Lorini, and Faustine Maffre. A poor mans epistemic logic based on propositional assignment and higher-order observation. In *Logic, Rationality and Interaction*, volume 9394 of *Lecture Notes in Computer Science*. Springer, 2015.

Jaakko Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, 1962.

Wiebe van der Hoek, Nicolas Troquard, and Michael Wooldridge. Knowledge and control. In *The 10th International Conference on Autonomous Agents and Multiagent Systems–Volume 2*, pages 719–726. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

Barteld Kooi and Bryan Renne. Generalized arrow update logic. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 205–211. ACM, 2011.

Emiliano Lorini, Luca Tummolini, and Andreas Herzig. Establishing mutual beliefs by joint attention: towards a formal model of public events. In *Proc. of CogSci*, pages 1325–1330, 2005.

Josef Perner and Heinz Wimmer. john thinks that mary thinks that attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471, 1985.

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

J. Seligman, F. Liu, and P. Girard. Facebook and the epistemic logic of friendship. In B.C. Schipper, editor, *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 229–238, 2013.

Michal Peter Sindlar. *In the Eye of the Beholder: Explaining Behavior through Mental State Attribution*. PhD thesis, Universiteit Utrecht, 2011.

Keith Stenning and Michiel Van Lambalgen. *Human reasoning and cognitive science*. MIT Press, 2008.

Kate Sullivan, Deborah Zaitchik, and Helen Tager-Flusberg. Preschoolers can attribute second-order beliefs. *Developmental Psychology*, 30(3):395, 1994.

Michael Tomasello. Joint attention as social cognition. *Joint attention: Its origins and role in development*, pages 103–130, 1995.

Johan van Benthem, Jan van Eijck, and Barteld Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

Hans van Ditmarsch and Barteld Kooi. Semantic results for ontic and epistemic change. In Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge, editors, *Logic and the Foundation of Game and Decision Theory (LOFT 7)*, Texts in Logic and Games 3, pages 87–117. Amsterdam University Press, 2008.

Hans Van Ditmarsch and Willem Labuschagne. My beliefs about your beliefs: A case study in theory of mind and epistemic logic. *Synthese*, 155(2):191–209, 2007.

Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. Dynamic epistemic logic with assignment. In Frank Dignum, Virginia Dignum, Sven Koenig, Sarit Kraus, Munindar P. Singh, and Michael Wooldridge, editors, *Autonomous Agents and Multi-agent Systems (AAMAS 2005)*, pages 141–148. ACM, 2005.

Hans van Ditmarsch, Andreas Herzig, Emiliano Lorini, and François Schwarzentruber. Listen to me! public announcements to agents that pay attention—or not. In *Logic, Rationality, and Interaction*, pages 96–109. Springer, 2013.

Rineke Verbrugge. Logic and social cognition. *Journal of Philosophical Logic*, 38 (6):649–680, 2009.

Stefan Wahl and Hans Spada. Childrens reasoning about intentions, beliefs and behaviour. *Cognitive Science Quarterly*, 1(1):3–32, 2000.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.