A Synthesis of Bayesian and Logical Approaches to the False Belief Task

Mathias Winther Madsen

March 6, 2014

1 Introduction

The false belief task is a classic experiment in developmental psychology which shows that children learn to attribute false beliefs to others between roughly the age of 3 or 4 [Wimmer and Perner, 1983]. Perhaps because of its close relationship to rationality and truth, this experiment has also received a lot of attention from researchers in logic, probability theory, and artificial intelligence, and it has been analyzed be means of a number of computational formalisms.

On a very coarse-grained level, these formalisms can be split in two categories: A probabilistic family, which uses Bayes Nets, neural networks, and similar formalisms [Berthiaume et al., 2008, Goodman et al., 2006]; and logical family, which uses epistemic logic, event calculi, hybrid logic, and other discrete-mathematics formalisms [van Ditmarsch and Labuschagne, 2007, Stenning and van Lambalgen, 2007, Arkoudas and Bringsjord, 2009, Braüner, 2013].

Some advantages of the probabilistic approach is that

- it can deal more easily with quantitative evidence because it builds probabilities directly into in the model;
- it comes equipped with a theory of learning and thus suggests a natural explanation for why children change behavior as they age;
- with a proper choice of prior distributions, it can also explain why a child starts its life using a simple and mostly correct heuristic but later in life switches to a more complicated but also more adequate model.

Some advantages of the logical approach is that

- it can represent much more complicated patters of belief, including beliefs about beliefs and non-rationalizable beliefs;
- it can give systematic accounts of how such beliefs models are constructed in terms of various event models.

These desirable properties unfortunately rarely appear together. The event calculus model designed by Arkoudas and Bringsjord [2009], for instance, uses a very sophisticated model of the processes by which the child comes to form a cognitive representation of the world, but their analysis does not contain any graded quantitative differences. This makes it hard to apply to noisy or ambiguous statistical data without substantial additional assumptions.

In the strictly Bayesian model of Goodman et al. [2006], on the other hand, "belief" is modeled quite simplistically as a single binary variable, and it is not clear how this generalizes to more complicated settings with multiple agents or multiple issues. Also, the Bayesian networks used in that paper are simply stipulated by the researchers, with no principled account of how the child would construct such models. They consequently write in their conclusion:

> The present account of the false belief transition is incomplete in important ways. After all, our agent had only to choose the best of two known models. This begs an understanding of the dynamics of rational revision near threshold and when the space of possible models is far larger. (p. 1387)

There are thus many excellent ideas in both the logical and the probabilistic literature on false beliefs, but less cross-pollination than one might hope (perhaps partly because of the mistaken belief that probability theory and logic are exclusive alternatives). The purpose of this paper is attempt at an integration of some of these ideas so as to reap the benefits of both formalisms.

My strategy will be, following van Benthem et al. [2009], to extract the core ideas from dynamic logic in the cleanest possible form and then enrich it with probabilities. On the formal level, I will generally follow their example, but my discussion will hopefully clarify some of the many connections and similarities between their logical apparatus and the corresponding probabilistic concepts, and my intention is that this paper should be readable without any prior knowledge of dynamic logic. After completing this discussion of the general framework, I will return to the false belief task and see what we can say about it from the perspective of "probabilized" dynamic logic.

The contribution of this paper is thus not a radically new idea or specialized inference system. I put together a number of modules that are already available, but too rarely combined. My aim is to reduce the number of formalisms in the world, not to increase it. Because of this synthetic approach, I suspect that about half of this paper will seem like old news to everybody in the intended audience, but I also suspect that it will not be the same half for everybody.

2 Generating Probabilistic Models

One of the important ideas that have come out of the literature on logic and artificial intelligence in the last half a century is that we don't have to be content writing models that describe specific situations — we can write meta-models

that automatically adapt to the world as circumstances change. Such meta-models prescribe how an agent's representation of the world should be updated in the face of new events [McCarthy and Hayes, 1969, Kowalski and Sergot, 1986, Baltag et al., 1998].

For situations involving multiple agents, such a system should ideally be able to handle two different kinds of events [Herzig and De Lima, 2006]:

- Ontic events, like flicking a switch or moving an object.
- Epistemic events, like measuring a temperature or telling somebody what time it is.

The design principle that underlies many recent "dynamic logic" solutions to these problems is that both of these kinds of events should be encoded in a single general-purpose format which can also be used to represent the initial situation [Baltag et al., 1998]. When this principle is respected, the operation of updating a model can often be described very compactly: You update a prior uncertainty model M_i with an event E by taking a Cartesian product

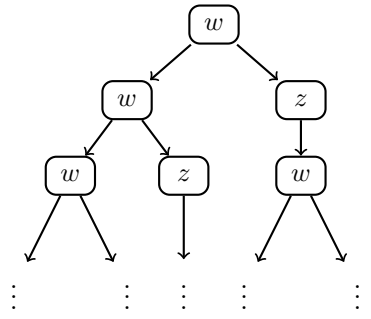
$$M_{i+1} = M_i \times E,$$

and then doing a little bit of bookkeeping to handle interaction effects.

More recently, these systems have been adjusted to accommodate probabilistic uncertainty too [van Benthem et al., 2009]. In the remainder of this section, I will introduce the tricks which form the foundation of dynamic logic and try to make their relation to probability theory as clear as possible. In section 3, I apply these ideas to the false belief task.

2.1 Probabilistic Kripke Frames

In order to represent an agent's knowledge state in an uncertain world, we need a representational scheme with a clear relationship to probability theory. A naive encoding of beliefs about beliefs causes a problem, however:



We are often not not only interested in what agents believe about the world and about other agents, but also in what those agents believe other agents believe about themselves and about yet other agents. A direct encoding of all of these layers would require an infinite model, so we need a compression strategy.

The classical solution to this problem is to use a cyclic graph instead of trees [Kripke, 1963]. The nodes in these graphs represent "possible worlds," and each possible world itself includes, in the probabilistic context that I am interested in here, a probability distribution over the full set of possible worlds. By allowing for this mild kind of circularity in the belief model, we are able to fold up the infinite tree of distributions into a smaller but loopy network.

In any particular "possible world" in this graph, the agents are able to reason probabilistically about the beliefs of the other agents by inferring which world they might be in, and which world the other agents might think they are in.

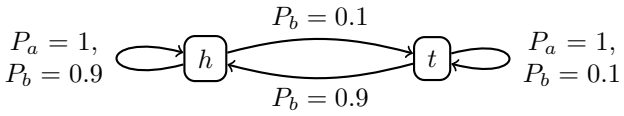
In the remainder of this paper, I will assume that a logical language L and and a set of agents A is fixed. We then have:

Definition 1. A (probabilistic) **Kripke frame** over a set Ω is a specification, for each $w \in \Omega$, of

1. a valuation function $V_w : L \to \{0, 1\}$ determining the truth value $V_w(\varphi)$ of any non-modal (ontic) formula φ at w;
2. for each agent $a \in A$, a probability distribution $P_a(v \,|\, w)$ over $v \in \Omega$.

Probabilistic Kripke frames are consistent with the axioms of probability theory if $P_a(x \,|\, y) = P_a(x \,|\, z)$ whenever $P_a(y \,|\, z) > 0$. All frames in this paper will satisfy this rationality constraint.

Kripke frames can be visualized as directed graphs, as in



Alternatively, they can also be tabulated, as in

w	$P_a(h \,\|\, w)$	$P_a(t \,\|\, w)$	$P_b(h \,\|\, w)$	$P_b(t \,\|\, w)$
h	1	0	0.9	0.1
t	0	1	0.9	0.1

Both of these frames represent the same coin flipping situation: One in which agent a knows whether the coin came up h or t, while b does not (but b strongly suspects that we are in case h).

In any particular possible world w, we can use such frames to compute the probabilities of various higher-order belief statements like "a knows that h is the case," or "$\Pr(\Box_a h \,|\, w)$." These probabilities are defined recursively:

Definition 2. Let a Kripke frame over a set Ω be given. Then

$$\Pr(\Box_a \varphi \,|\, w) := \sum_{v \in \Omega} \Pr(\varphi \,|\, v)\, P_a(v \,|\, w)$$

These concepts define a very general uncertainty calculus, and in the following subsection, I will discuss how it applies to the representation of events.

2.2 Event Frames and Product Updates

We need to prescribe a method for handling two types of events, ontic and epistemic. We start with the ontic, since they are simpler. These events are constructed out of the following uncertainty-free building block:

Definition 3. An **ontic action** e is mapping from the set of valuation functions $\{0,1\}^L$ to itself.

The idea is here that the new valuation $V'_w = e(V_w)$ describes what is and is not the case in world w "after e happens." It is useful to think about e as specifying some snippet of computer code to be executed, and V_w as a table that keeps track of the current value of all the variables in the program.

Events can have uncertain consequences. We capture this uncertainty too in the form of Kripke frames:

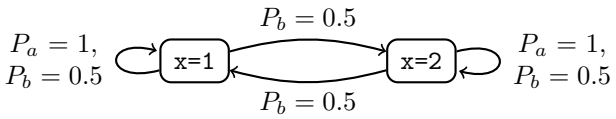
Definition 4. An **event frame** is a Kripke frame over a set of ontic actions.

Because event frames are specified in the same format as the prior situation, they combine with situations according to a relatively simple update rule:

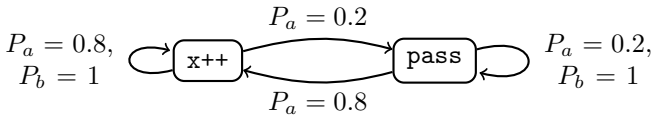
Definition 5. The **product update** of a Kripke frame over a set Ω with an event frame over a set E is the Kripke frame over $\Omega \times E$ for which

1. the valuation at world (w, e) is $e(V_w)$;
2. for all agents a, $P_a((v,f) \,|\, (w,e)) := P_a(v \,|\, w)\, P_a(f \,|\, e)$.

As an example, suppose that the prior situation is that a variable either has the value `x=1` or `x=2`, and suppose that a knows this, but b doesn't:

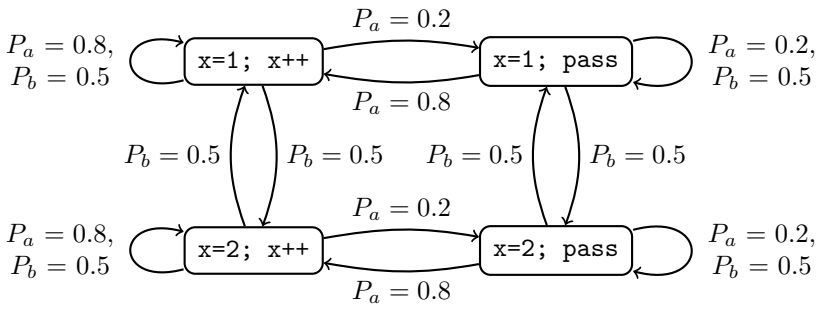


Beginning from this prior situation, we now execute a piece of code which probably increments `x` by one (`x++`), but might also do nothing (`pass`). We'll assume that b knows what the code does, but that a doesn't:



The posterior situation that corresponds to this prior situation and event

frame is then the "product" of those two Kripke frames:



If we are confident that the model will never be queried for information about the events that occurred, but only about the value of `x`, we can also simplify it a bit by executing each of the little scripts in the event frame. A state/action pair like "`x=2; x++`" would then reduce to "`x=3`," and so forth.

As this example shows, the product update is in many ways the most general and least committed way of revising a model: It consists in combining all contingencies that might obtain in all ways they might be combined. it also tends, as the name suggests, to increase the model size exponentially as new events occur.

Notice that the probability of each state/action combination is computed as a product of the state probability and the action probability. This means in effect that the state and the action are treated as stochastically independent by this rule. If we want any interdependence between the two [e.g., preconditions; Reiter, 1991], we have to encode these in the form of conditional statements (`if ... then ...`) in the atomic events.

2.3 Announcements and Epistemic Events

The update rule defined above describes how a model should be updated in the light of ontic actions: They change the states of affairs but leave the probability distributions as unchanged as possible. Epistemic actions do the opposite: They change the distributions, but leave the truth values.

The atomic (uncertainty-free) version of an epistemic action is a partition:

Definition 6. An **epistemic action** is a partition of a set of possible worlds.

An epistemic actions can be thought of as issues or questions, e.g., "What is the value of the variable X?" Each class in the partition is then a resolution of that issue. The reason that we use questions instead of answers as a building block is that we want to enable agents to learn that somebody has received an answer to a question without actually learning that answer.

Definition 7. Suppose a Kripke frame over a set Ω is given. The Kripke frame produced by **private announcement** of the epistemic action π to an agent a

is then a frame identical to the old one except that a's probability distributions $P_a(v \mid w)$ are updated to

$$P_a(v \mid w, \pi) := \frac{P_a(v \mid w)\, \pi(v \mid w)}{\sum_{u \in \Omega} P_a(u \mid w)\, \pi(u \mid w)},$$

where $\pi(v \mid w) = 1$ if w and v are in the same class of the partition π, and $\pi(v \mid w) = 0$ otherwise.

A private announcement is thus a conditioning operation in which each partition class thus becomes its own renormalized sample space. In the coin flipping example on page 4, for instance, a's knowledge can be interpreted as an updated version of b's knowledge after the private announcement of the partition $\{\{h\}, \{t\}\}$. This corresponds to an event of the type "the value of X is revealed to a."

With this concept in place, we can now define the update rule for uncertain epistemic events in the same manner as for uncertain ontic events:

Definition 8. An **announcement frame** is a Kripke frame over a set of epistemic actions.

Definition 9. Suppose a Kripke frame over a set Ω is given along with an announcement frame over a set Π. The **announcement frame update** is then the Kripke frame over the product set $\Omega \times \Pi$ for which

1. the valuation at world (w, π) is V_w;
2. for all agents a,

$$P_a((v, \psi) \mid (w, \pi)) := P_a(v \mid w, \psi)\, P_a(\psi \mid \pi),$$

where $P_a(v \mid w, \psi)$ is the probability distribution at w produced by private announcement of ψ to a (as in Def. 7).

The update rule thus consists in a definition of the joint probabilities for the state-action combinations; only now, the actions are epistemic instead of ontic. By this rule, we can deal with statements like "a is not sure whether the value of X has been revealed to b."

3 An Application to the False Belief Task

I will now construct the model that is associated with one possible version of story of the false belief task. Using the prior probabilities encoded in the model of Goodman et al. [2006], this story can be translated as follows:

1. There is a slight probability that Sally does not want a cookie ($\varepsilon = 1/11$).
   (a) Whether she does is announced to her, but not to the child.

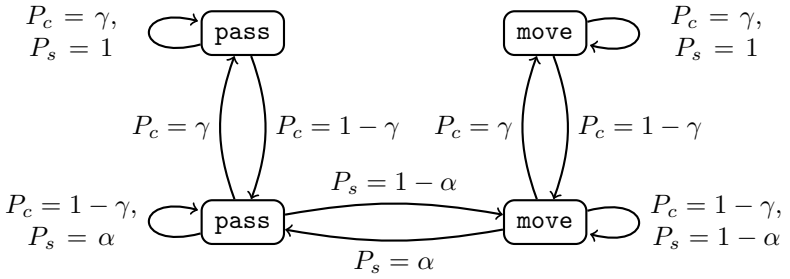
2. The cookie is put in a basket (and this is announced to everybody).

3. The cookie is possibly moved to a box. (Goodman et al. [2006] provide no probabilities for this, but we might arbitrarily set it to $\alpha = 1/10$.)

   (a) This displacement is announced to the child, and it might also be announced to Sally too (with probability $\gamma = 1/6$).

      i. Whether it was is not announced to the child.

The most complicated part of this story is the last event, since it involves uncertainty about uncertainty. It is modeled by the following frame:



The posterior model produced by this event frame essentially consists of two copies of this picture, one where Sally wants a cookie, and one where she doesn't.

Notice that the parameter γ here measures how "flat" the child's model of the world is. If $\gamma = 1$, the child will assume that Sally is a "mindreader" having access to the exact same information as him- or herself.

Let us define the sentence "Sally looks in the box" (*look*) to be the logical conjunction of "Sally wants the cookie" (*want*) and "Sally believes that the cookie was moved" ($\Box_s move$). The proposition $\Box_s move$ is related to, but not identical to the fact that the displacement was announced to Sally (*tell*). Across the four possible worlds where the cookie was moved ($[\![move]\!]$), the child now has the probability distribution shown in Table 3.

In these possible words, the child has the following conditional probabilities:

1. The probability that Sally will look in the box is $(1-\varepsilon)\gamma + (1-\varepsilon)(1-\gamma)\alpha$. With the parameter settings above, this number evaluates to $5/22 \approx 22.7\%$. Holding α and ε fixed, it is larger than $1/2$ when $\gamma > 1/2$.

2. Given that Sally looked in the box, the probability that she somehow learned where the cookie was (perhaps by mind-reading) is

$$\frac{(1-\varepsilon)\gamma}{(1-\varepsilon)\gamma + (1-\varepsilon)(1-\gamma)\alpha},$$

   which evaluates to $2/3$. The remaining probability is accounted for by the possibility that she guessed that the cookie moved because things just move sometimes (with probability α).

w	*want*	*tell*	$P_c(w \mid [\![move]\!])$	$P_c(\Box_s move)$	$P_c(look)$
w_1	1	1	$(1-\varepsilon)\gamma$	1	1
w_2	1	0	$(1-\varepsilon)(1-\gamma)$	α	α
w_3	0	1	$\varepsilon\gamma$	1	0
w_4	0	0	$\varepsilon(1-\gamma)$	α	0

Table 3: The child's probability distribution in the partition class $[\![move]\!]$.

3. Given that Sally does not look in the box, the probability that she did not have the same information as the child is

$$\frac{(1-\varepsilon)(1-\gamma)(1-\alpha)+\varepsilon(1-\gamma)}{(1-\varepsilon)(1-\gamma)(1-\alpha)+\varepsilon\gamma+\varepsilon(1-\gamma)},$$

which evaluates to $^{50}/_{51} \approx 98.0\%$. This event thus strongly favors the more complicated model and can lead the child to decrease γ.

4. Given that Sally did not look in the box, the probability that she did not want the cookie is

$$\frac{(1-\varepsilon)(1-\gamma)(1-\alpha)}{(1-\varepsilon)(1-\gamma)(1-\alpha)+\varepsilon\gamma+\varepsilon(1-\gamma)},$$

which evaluates to $^{2}/_{17} \approx 11.8\%$. When γ increases, this lack-of-desire explanation increases in probability, as the child will assign less and less probability to a lack-of-knowledge explanation.

We thus find many of the observations by Goodman et al. [2006] replicated here: There is a single parameter which controls to which extent the child is able to take the perspective of another person, and this parameter should change value as the child ages. There is, in addition, a correlation between the value of this parameter and the type of explanation that the child will give for certain events.

The natural extension of this line of thought would be to apply the model to more complicated situations with multiple agents and more uncertainty. This is beyond the scope of this paper, but because the behavior of this model is defined in such generality, such an extension would be almost a purely mechanical.

References

Konstantine Arkoudas and Selmer Bringsjord. Propositional attitudes and causation. *International Journal of Software and Informatics*, 3(1):47–65, 2009.

Alexandru Baltag, Lawrence S. Moss, and Sławomir Solecki. The Logic of Public Announcements, Common Knowledge, and Private Suspicions. In *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, pages 43–56, 1998.

Vincent G. Berthiaume, Kristine H. Onishi, and Thomas R. Shultz. A Computational Developmental Model of the Implicit False Belief Task. In B. C. Love, K. McRae, and V. M. Sloutsky, editors, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 825–30, 2008.

Torben Braüner. Hybrid-Logical Reasoning in False-Belief Tasks. In Burkhard C. Schipper, editor, *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 186–195. 2013.

Noah D. Goodman, Chris L. Baker, Elizabeth Baraff Bonawitz, Vikash K. Mansinghka, Alison Gopnik, Henry Wellman, Laura Schulz, and Joshua B. Tenenbaum. Intuitive Theories of Mind: A Rational Approach to False Belief. In *Proceedings of the twenty-eighth annual conference of the cognitive science society*, pages 1382–1387, 2006.

Andreas Herzig and Tiago De Lima. Epistemic Actions and Ontic Actions: A Unified Logical Framework. In Solange Oliveira Rezende Jaime Simão Sichman, Helder Coelho, editor, *Advances in Artificial Intelligence – IBERAMIA-SBIA 2006*, pages 409–418. 2006.

Robert Kowalski and Marek Sergot. A Logic-based Calculus of Events. *New Generation Computing*, 4:67–94, 1986.

Saul A. Kripke. Semantical Considerations on Modal Logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

John McCarthy and Patrick Hayes. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In D. Michie and B. Meltzer, editors, *Machine Intelligence*, volume 4, pages 463–502. 1969.

Raymond Reiter. The Frame Problem in the Situation Calculus: A Simple Solution (Sometimes) and a Completeness Result for Goal Regression. In Vladimir Lifschitz, editor, *Artificial and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, pages 359–380. Academic Press, 1991.

Keith Stenning and Michiel van Lambalgen. Logic in the study of psychiatric disorders: executive function and rule-following. *Topoi*, 26(1):97–114, 2007.

Johan van Benthem, Jelle Gerbrandy, and Barteld Kooi. Dynamic Update with Probabilities. *Studia Logica*, 93(1):67–96, 2009.

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.

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.