Toward formalizing common-sense psychology: an analysis of the false-belief task

Konstantine Arkoudas and Selmer Bringsjord

Cognitive Science and Computer Science Departments, RPI
arkouk@rpi.edu, brings@rpi.edu

Abstract. Predicting and explaining the behavior of others in terms of mental states is indispensable for everyday life. It will be equally important for artificial agents. We present an inference system for representing and reasoning about mental states, and use it to provide a formal analysis of the false-belief task. The system allows for the representation of information about events, causation, and perceptual, doxastic, and epistemic states (vision, belief, and knowledge), incorporating ideas from the event calculus and multi-agent epistemic logic. Unlike previous AI formalisms, our focus here is on mechanized proofs and proof programmability, not on metamathematical results. Reasoning is performed via cognitively plausible inference rules, and automation is achieved by general-purpose inference *methods*. The system has been implemented as an interactive theorem prover and is available for experimentation.[1]

1 Introduction

Predicting and explaining the behavior of other people is indispensable for everyday life. The ability to ascribe mental states to others and to reason about such mental states is pervasive and invaluable. All social transactions—from engaging in commerce and negotiating to making jokes and empathizing with other people's pain or joy—require at least a rudimentary grasp of common-sense psychology (CSP). Artificial agents without an ability of this sort would be severely handicapped in their interactions with humans. This could present problems not only for artificial agents trying to interpret human behavior, but also for artificial agents trying to interpret the behavior of one another. When a system exhibits a complex but rational behavior, and detailed knowledge of its internal structure is not available, the best strategy for predicting and explaining its actions might be to analyze its behavior in intentional terms, i.e., in terms of mental states such as beliefs and desires (regardless of whether the system *actually* has genuine mental states). Mentalistic models are likely to be particularly apt for agents trying to manipulate the behavior of other agents.

[1] That prover, along with code that makes it possible to engineer autonomous synthetic characters (residing on servers in our lab) that have avatars in *Second Life*, has also been used to allow such characters to pass the false-belief task. For demonstrations, visit www.cogsci.rpi.edu/research/rair/asc_rca/SLDemos.

Any computational treatment of CSP will have to integrate action and cognition. Agents must be able to reason about the causes and effects of various events, whether they are non-intentional physical events or intentional events brought about by their own agency. More importantly, they must be able to reason about what others believe or know about such events. To that end, our system combines and adapts ideas drawn from the event calculus and from multi-agent epistemic logics. It is based on multi-sorted first-order logic extended with subsorting, epistemic operators for perception, belief, and knowledge, and mechanisms for reasoning about causation and action. Using subsorting, we formally model agent actions as types of events, which enables us to use the resources of the event calculus to represent and reason about agent actions. The usual axioms of the event calculus are encoded as common knowledge, suggesting that people have an understanding of the basic folk laws of causality (innate or acquired), and are indeed aware that others have such an understanding.

It is important to be clear about what we hope to accomplish through the present work. In general, any logical system or methodology capable of representing and reasoning about intentional notions such as knowledge can have at least three different uses. First, it can serve as a tool for the specification, analysis, and verification of rational agents. Second, in tandem with some appropriate reasoning mechanism, it can serve as a knowledge representation framework, i.e., it can be used *by* artificial agents to represent their own "mental states"—and those of other agents—and to deliberate and act in accordance with those states and their environment. Finally, it can be used to provide formal models of certain interesting cognitive phenomena. One intended contribution of our present work is of the third sort, namely, to provide a formal model of false-belief attributions, and, in particular, a description of the logical competence of an agent capable of passing a false-belief task. It addresses questions such as the following: What sort of principles is it plausible to assume that an agent has to deploy in order to be able to succeed on a false-belief task? What is the depth and complexity of the required reasoning? Can such reasoning be automated, and if so, how? These questions have not been taken up in detail in the relevant discussions in cognitive science and the philosophy of mind, which have been couched in overly abstract and rather vague terms. Formal computational models such as the one we present here can help to ground such discussions, to clarify conceptual issues, and to begin to answer important questions in a concrete setting.

Although the import of such a model is primarily scientific, there can be interesting engineering implications. For instance, if the formalism is sufficiently expressive and versatile, and the posited computational mechanisms can be automated with reasonable efficiency, then the system can make contributions to the first two areas mentioned above. We believe that our system has such potential for two reasons. First, the combination of epistemic constructs such as common knowledge with the conceptual resources of the event calculus for dealing with causation appears to afford great expressive power, as demonstrated by our formalization. A key technical insight behind this combination is the modelling of agent actions as events via subsorting. Second, procedural abstraction

mechanisms appear to hold significant promise for automation; we discuss this issue later in more detail.

The remainder of this paper is structured as follows. The next section gives the formal definition of our system. Section 3 represents the false-belief task in our system, and section 4 presents a model of the reasoning that is required to succeed in such a task, carried out in a modular fashion by collaborating methods. Section 5 discusses some related work and concludes.

2 A calculus for representing and reasoning about mental states

The syntactic and semantic problems that arise when one tries to use classical logic to represent and reason about intentional notions are well-known. Syntactically, modelling belief or knowledge relationally is problematic because one believes or knows arbitrarily complex propositions, whereas the arguments of relation symbols are terms built from constants, variables, and function symbols. (The objects of belief could be encoded by strings, but such representations are too low-level for most purposes.) Semantically, the main issue is the referential opacity (or intensionality) exhibited by propositional-attitude operators. In intensional contexts one cannot freely substitute one coreferential term for another. Broadly speaking, there are two ways of addressing these issues. One is to use a modal logic, with built-in syntactic operators for intentional notions. The other is to retain classical logic but distinguish between an object-language and a meta-language, representing intentional discourse at the object level. Each approach has its advantages and drawbacks. Retaining classical logic has the important advantage of efficiency, in that (semi-)automated deduction systems for classical logic, such as resolution provers—which have made impressive strides over the last decade—can be used for reasoning. This is the option we have chosen in some previous work [3]. One disadvantage of this approach is that when the object language is first-order (includes quantification), then notions such as substitutions and alphabetic equivalence must be explicitly encoded. Depending on the facilities provided by the meta-language, this does not need to be overly onerous, but it does require extra effort. The modal-logic approach has the advantage of solving the syntactic and referential-opacity problems directly, without the need to distinguish an object-language and a meta-language. That is the approach we have taken in this work.

The specification of the syntax of our system appears in figure 1, which describes the various sorts of our universe (S), the signatures of certain built-in function symbols (f), and the abstract syntax of terms (t) and propositions (P). The symbol $\sqsubseteq$ denotes subsorting. Propositions of the form $\mathbf{S}(a, P)$, $\mathbf{B}(a, P)$, and $\mathbf{K}(a, P)$ should be understood as saying that agent a sees that P is the case, believes that P, and knows that P, respectively. Propositions of the form $\mathbf{C}(P)$ assert that P is commonly known. Sort annotations will generally be omitted, as they are easily deducible from the context. We write $P[x \mapsto t]$ for the proposition obtained from P by replacing every free occurrence of x by t, assuming that t

$$
\begin{array}{rl}
S ::= & \texttt{Object} \mid \texttt{Agent} \mid \texttt{ActionType} \mid \texttt{Action} \sqsubseteq \texttt{Event} \mid \texttt{Moment} \mid \texttt{Boolean} \mid \texttt{Fluent} \\
f ::= & \begin{array}{l} \mathit{action} : \texttt{Agent} \times \texttt{ActionType} \rightarrow \texttt{Action} \\ \mathit{initially} : \texttt{Fluent} \rightarrow \texttt{Boolean} \\ \mathit{holds} : \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\ \mathit{happens} : \texttt{Event} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\ \mathit{clipped} : \texttt{Moment} \times \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\ \mathit{initiates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\ \mathit{terminates} : \texttt{Event} \times \texttt{Fluent} \times \texttt{Moment} \rightarrow \texttt{Boolean} \\ \mathit{prior} : \texttt{Moment} \times \texttt{Moment} \rightarrow \texttt{Boolean} \end{array} \\
t ::= & x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
P ::= & t : \texttt{Boolean} \mid \neg P \mid P \wedge Q \mid P \vee Q \mid P \Rightarrow Q \mid P \Leftrightarrow Q \mid \\
& \forall\, x : S \;.\; P \mid \exists\, x : S \;.\; P \mid \mathbf{S}(a, P) \mid \mathbf{K}(a, P) \mid \mathbf{B}(a, P) \mid \mathbf{C}(P)
\end{array}
$$

Fig. 1. The specification of sorts, function symbols, terms, and propositions.

is of a sort compatible with the sort of the free occurrences in question, and taking care to rename P as necessary to avoid variable capture. We use the infix notation $t_1 < t_2$ instead of $\mathit{prior}(t_1, t_2)$.

We express the following standard axioms of the event calculus as common knowledge:

$$[A_1]\ \mathbf{C}(\forall\, f, t \;.\; \mathit{initially}(f) \wedge \neg \mathit{clipped}(0, f, t) \Rightarrow \mathit{holds}(f, t))$$
$$[A_2]\ \mathbf{C}(\forall\, e, f, t_1, t_2 \;.\; \mathit{happens}(e, t_1) \wedge \mathit{initiates}(e, f, t_1) \wedge t_1 < t_2 \wedge \neg \mathit{clipped}(t_1, f, t_2) \Rightarrow \mathit{holds}(f, t_2))$$
$$[A_3]\ \mathbf{C}(\forall\, t_1, f, t_2 \;.\; \mathit{clipped}(t_1, f, t_2) \Leftrightarrow [\exists\, e, t \;.\; \mathit{happens}(e, t) \wedge t_1 < t < t_2 \wedge \mathit{terminates}(e, f, t)])$$

suggesting that people have a (possibly innate) understanding of basic causality principles, and are indeed aware that everybody has such an understanding. In addition to $[A_1]$—$[A_3]$, we postulate a few more axioms pertaining to what people know or believe about causality. First, agents know the events that they intentionally bring about themselves—that is part of what "action" means. In fact, this is common knowledge. The following axiom expresses this:

$$[A_4]\ \mathbf{C}(\forall\, a, d, t \;.\; \mathit{happens}(\mathit{action}(a, d), t) \Rightarrow \mathbf{K}(a, \mathit{happens}(\mathit{action}(a, d), t)))$$

The next axiom states that it is common knowledge that if an agent a believes that a certain fluent f holds at t and he does not believe that f has been clipped between t and t', then he will also believe that f holds at t':

$$[A_5]\ \mathbf{C}(\forall\, a, f, t, t' \;.\; \mathbf{B}(a, \mathit{holds}(f, t)) \wedge \mathbf{B}(a, t < t') \wedge \neg \mathbf{B}(a, \mathit{clipped}(t, f, t')) \Rightarrow \mathbf{B}(a, \mathit{holds}(f, t')))$$

The final axiom states that if a believes that b believes that f holds at t_1 and a believes that nothing has happened between t_1 and t_2 to change b's mind, then a will believe that b will not think that f has been clipped between t_1 and t_2:

$$\begin{array}{l}[A_6]\ \forall\, a, b, t_1, t_2, f \;.\; [\mathbf{B}(a, \mathbf{B}(b, \mathit{holds}(f, t_1))) \wedge \mathbf{B}(a, \neg\exists\, e, t \;.\; \mathbf{B}(b, \mathit{happens}(e, t)) \wedge \\ \qquad \mathbf{B}(b, t_1 < t < t_2) \wedge \mathbf{B}(b, \mathit{terminates}(e, f, t)))] \Rightarrow \mathbf{B}(a, \neg\mathbf{B}(b, \mathit{clipped}(t_1, f, t_2)))\end{array}$$

This captures a form of closed-world reasoning, for it could well be the case that, in fact, b has come to believe that something has happened between t and t' that

terminated f, and therefore no longer believes that f holds. But if a believes that there have been no such events, then it is reasonable for a to assume that b will not believe that f has been clipped.

In addition to the usual introduction and elimination rules for first-order predicate logic with equality, we will make use of the following inference rules:

$$\frac{}{\mathbf{C}(\mathbf{S}(a,P) \Rightarrow \mathbf{K}(a,P))}\ [R_1] \qquad \frac{}{\mathbf{C}(\mathbf{K}(a,P) \Rightarrow \mathbf{B}(a,P))}\ [R_2]$$

$$\frac{\mathbf{C}(P)}{\mathbf{K}(a_1,\mathbf{K}(a_2,\mathbf{K}(a_3,P)))}\ [R_3] \qquad \frac{\mathbf{K}(a,P)}{P}\ [R_4]$$

$[R_1]$ says that it is common knowledge that visual perception is a justified source of knowledge. In other words, it is commonly known that if I *see* that P, I know P.[2] $[R_2]$ says that it is commonly known that knowledge requires belief, while $[R_3]$ captures an essential property of common knowledge. Usually common knowledge of a proposition P is taken to mean that everybody knows that P, everybody knows that everybody knows that P, and so on ad infinitum. This is captured by recursive rules that allow us to "unfold" the common-knowledge operator arbitrarily many times. However, this viewpoint is quite problematic for finite knowers of limited cognitive capacity. After three or four levels of nesting, iterated knowledge claims become unintelligible. Because in the present setting we are concerned with cognitive plausibility, we refrain from characterizing common knowledge in the customary strong form, imposing instead limit of three levels of iteration, as indicated in $[R_3]$.[3] $[R_4]$ is a veracity rule for knowledge.

The following rules can now be readily derived:

$$\frac{\mathbf{C}(P)}{\mathbf{K}(a_1,\mathbf{K}(a_2,P))}\ [DR_1] \qquad \frac{\mathbf{C}(P)}{\mathbf{K}(a,P)}\ [DR_2]$$

$$\frac{\mathbf{C}(P)}{P}\ [DR_3] \qquad \frac{\mathbf{S}(a,P)}{\mathbf{K}(a,P)}\ [DR_4] \qquad \frac{\mathbf{K}(a,P)}{\mathbf{B}(a,P)}\ [DR_5]$$

We next have the following three rules:

$$\frac{}{\mathbf{C}(\mathbf{K}(a,P_1 \Rightarrow P_2) \Rightarrow \mathbf{K}(a,P_1) \Rightarrow \mathbf{K}(a,P_2))}\ [R_5]$$

$$\frac{}{\mathbf{C}(\mathbf{B}(a,P_1 \Rightarrow P_2) \Rightarrow \mathbf{B}(a,P_1) \Rightarrow \mathbf{B}(a,P_2))}\ [R_6] \qquad \frac{}{\mathbf{C}(\mathbf{C}(P_1 \Rightarrow P_2) \Rightarrow \mathbf{C}(P_1) \Rightarrow \mathbf{C}(P_2))}\ [R_7]$$

From these we can easily derive the so-called Kripke ("K") rules for knowledge, belief, and common knowledge:

$$\frac{\mathbf{K}(a,P_1 \Rightarrow P_2) \quad \mathbf{K}(a,P_1)}{\mathbf{K}(a,P_2)}\ [DR_6]$$

We likewise have derived rules $[DR_7]$ and $[DR_8]$ for belief and common knowledge, respectively (omitted here). We also assume that a few straightforward tautologies are common knowledge, and the self-explanatory $[R_{11}]$:

[2] We currently ignore the issue of perceptual illusions.

[3] Although there is not enough space here for a full discussion, we point out that third-order epistemic and doxastic states (as opposed to n-order for $n > 3$) are often held to be at a level of iteration sufficient for general accounts of human thinking, e.g., see Dennett (1978). This is not to say that fairly realistic scenarios involving iteration of 4 or even 5 levels cannot be devised, but in the present paper we have used 3 for the purpose of modeling the false-belief task.

$$\frac{}{\mathbf{C}((\forall\ x\ .\ P) \Rightarrow P[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}([P_1 \Leftrightarrow P_2] \Rightarrow \neg P_2 \Rightarrow \neg P_1)}\ [R_9]$$

$$\frac{}{\mathbf{C}([P_1 \wedge \cdots \wedge P_n \Rightarrow P] \Rightarrow [P_1 \Rightarrow \cdots \Rightarrow P_n \Rightarrow P])}\ [R_{10}] \qquad \frac{\mathbf{B}(a, P_1) \quad \mathbf{B}(a, P_2)}{\mathbf{B}(a, P_1 \wedge P_2)}\ [R_{11}]$$

Note that usually it is postulated that *every* tautology is common knowledge. If we took that as a principle, the presentation of the system could be somewhat simplified. However, such a principle (and other "logical omniscience" principles like it) is wildly implausible, as has often been pointed out. Since we do not accept such unrestricted principles, we only posit certain specific tautologies that are intuitively deemed as obvious. While this is not a general solution, it nevertheless averts the cognitive implausibility of the unrestricted rules, and also serves to isolate the logical knowledge that we need to attribute to agents for a specific reasoning problem.

The following rules are now readily derived:[4]

$$\frac{\mathbf{K}(a, \forall\ x\ .\ P)}{\mathbf{K}(a, P[x \mapsto t])}\ [DR_9] \qquad \frac{\mathbf{B}(a, \forall\ x\ .\ P)}{\mathbf{B}(a, P[x \mapsto t])}\ [DR_{10}] \qquad \frac{\mathbf{C}(\forall\ x\ .\ P)}{\mathbf{C}(P[x \mapsto t])}\ [DR_{11}] \qquad \frac{\mathbf{B}(a_1, \mathbf{K}(a_2, P))}{\mathbf{B}(a_1, \mathbf{B}(a_2, P))}\ [DR_{12}]$$

$$\frac{\mathbf{K}(a_1, \mathbf{K}(a_2, P_1 \Rightarrow P_2)) \quad \mathbf{K}(a_1, \mathbf{K}(a_2, P_1))}{\mathbf{K}(a_1, \mathbf{K}(a_2, P_2))}\ [DR_{13}] \qquad \frac{\mathbf{B}(a_1, \mathbf{B}(a_2, P_1 \Rightarrow P_2)) \quad \mathbf{B}(a_1, \mathbf{B}(a_2, P_1))}{\mathbf{B}(a_1, \mathbf{B}(a_2, P_2))}\ [DR_{14}]$$

$$\frac{\mathbf{K}(a_1, \mathbf{K}(a_2, P_1 \Leftrightarrow P_2)) \quad \mathbf{K}(a_1, \mathbf{K}(a_2, \neg P_2))}{\mathbf{K}(a_1, \mathbf{K}(a_2, \neg P_1))}\ [DR_{15}] \qquad \frac{\mathbf{B}(a_1, \mathbf{B}(a_2, P_1 \Leftrightarrow P_2)) \quad \mathbf{B}(a_1, \mathbf{B}(a_2, \neg P_2))}{\mathbf{B}(a_1, \mathbf{B}(a_2, \neg P_1))}\ [DR_{16}]$$

$$\frac{\mathbf{K}(a_1, \mathbf{K}(a_2, [P_1 \wedge \cdots \wedge P_n] \Rightarrow P)) \quad \mathbf{K}(a_1, \mathbf{K}(a_2, P_1)) \quad \cdots \quad \mathbf{K}(a_1, \mathbf{K}(a_2, P_n))}{\mathbf{K}(a_1, \mathbf{K}(a_2, P))}\ [DR_{17}]$$

$$\frac{\mathbf{B}(a_1, \mathbf{B}(a_2, [P_1 \wedge \cdots \wedge P_n] \Rightarrow P)) \quad \mathbf{B}(a_1, \mathbf{B}(a_2, P_1)) \quad \cdots \quad \mathbf{B}(a_1, \mathbf{B}(a_2, P_n))}{\mathbf{B}(a_1, \mathbf{B}(a_2, P))}\ [DR_{18}]$$

$$\frac{\mathbf{B}(a, P_1 \wedge P_2 \wedge P_3 \Rightarrow P_4) \quad \mathbf{B}(a, P_1) \quad \mathbf{B}(a, P_2) \quad \mathbf{B}(a, P_3)}{\mathbf{B}(a, P_4)}\ [DR_{19}]$$

The system presented in this section has been implemented in the form of a denotational proof language, as a language similar to the Athena system [2].

3 Encoding the false-belief task

False-belief scenarios can be regarded as the drosophila of computational theories of mind. Experiments with false beliefs were first carried out by Wimmer and Perner [12]. In a typical scenario, a child (we will call her Alice) is presented with a story in which a character (we will call him Bob) places an object (say, a cookie) in a certain location l_1, say in a particular kitchen cabinet. Then Bob leaves, and during his absence someone else (say, Charlie) removes the object from its original location l_1 and puts it in a different location l_2 (say, a kitchen drawer). Alice is then asked to predict where Bob will look for the object when he gets back, the right answer, of course, being the original location—the cabinet. In this section we show how to formalize this scenario in our calculus. In the next section we will present a formal explanation as to how Alice can come to acquire the correct belief about Bob's false belief.

We introduce the sort `Location` and the following function symbols specifically for reasoning about the false-belief task:

[4] Derivation proofs are omitted, but can be obtained (along with the computer implementation of the system) by contacting the authors.

$$places : \texttt{Object} \times \texttt{Location} \rightarrow \texttt{ActionType}$$
$$moves : \texttt{Object} \times \texttt{Location} \times \texttt{Location} \rightarrow \texttt{ActionType}$$
$$located : \texttt{Object} \times \texttt{Location} \rightarrow \texttt{Fluent}$$

Intuitively, $action(a, places(o, l))$ signifies a's action of placing object o in location l, while $action(a, moves(o, l_1, l_2))$ is a's action of moving object o from location l_1 to location l_2. It is common knowledge that placing o in l initiates the fluent $located(o, l)$:

$$[D_1] \quad \mathbf{C}(\forall\, a, t, o, l \;.\; initiates(action(a, places(o, l)), located(o, l), t))$$

It is likewise known that if an object o is located at l_1 at a time t, then the act of moving o from l_1 to l_2 results in o being located at l_2:

$$[D_2] \quad \mathbf{C}(\forall\, a, t, o, l_1, l_2 \;.\; holds(located(o, l_1), t) \Rightarrow initiates(action(a, moves(o, l_1, l_2)), located(o, l_2), t))$$

If, in addition, the new location is different from the old one, the move terminates the fluent $located(o, l_1)$:

$$[D_3] \quad \mathbf{C}(\forall\, a, t, o, l_1, l_2 \;.\; holds(located(o, l_1), t) \wedge l_1 \neq l_2 \Rightarrow terminates(action(a, moves(o, l_1, l_2)), located(o, l_1), t))$$

The following axiom captures the constraint that an object cannot be in more than one place at one time; this is also common knowledge:

$$[D_4] \quad \mathbf{C}(\forall\, o, t, l_1, l_2 \;.\; holds(located(o, l_1), t) \wedge holds(located(o, l_2), t) \Rightarrow l_1 = l_2)$$

We introduce three time moments that are central to the narrative of the false-belief task: *beginning*, *departure*, and *return*. The first signifies the time point when Bob places the cookie in the cabinet, while *departure* and *return* mark the points when he leaves and comes back, respectively. We assume that it's common knowledge that these three time points are linearly ordered in the obvious manner:

$$[D_5] \quad \mathbf{C}(beginning < departure < return).$$

We also introduce two distinct locations, *cabinet* and *drawer*:

$$[D_6] \quad \mathbf{C}(cabinet \neq drawer).$$

Finally, we introduce a domain `Cookie` as a subsort of `Object`, and declare a single element of it, *cookie*. It is a given premise that, in the beginning, Alice sees Bob place the cookie in the cabinet:

$$[D_7] \quad \mathbf{S}(Alice, happens(action(Bob, places(cookie, cabinet)), beginning)).$$

4 Modeling the reasoning underlying false-belief tasks, and automating it via abstraction

At this point we have enough representational and reasoning machinery in place to infer the correct conclusion from a couple of obvious premises. However, a monolithic derivation of the conclusion from the premises would be unsatisfactory, as it would not give us a story about how such reasoning can be dynamically put together. Agents must be able to reason about the behavior of other agents

efficiently. It is not at all obvious how efficiency can be achieved in the absence of mechanisms for abstraction, modularity, and reusability.

We can begin to address both issues by pursuing further the idea of derived inference rules, and by borrowing a page from classic work in cognitive science and production systems. Suppose that we had a mechanism which enabled the derivation of not only *schematic* inference rules, such as the ones that we presented in section 2, but derived inference rules allowing for arbitrary computation and search. We could then formulate *generic* inference rules, capable of being applied to an unbounded (potentially infinite) number of arbitrarily complex concrete situations.

Our system has a notion of *method* that allows for that type of abstraction and encapsulation. Methods are derived inference rules, not just of the schematic kind, but incorporating arbitrary computation and search. They are thus more general than the simple if-then rules of production systems, and more akin to the knowledge sources (or "demons") of blackboard systems [8]. They can be viewed as encapsulating specialized expertise in deriving certain types of conclusions from certain given information. They can be parameterized over any variables, e.g., arbitrary agents or time points.

A key role in our system is played by an associative data structure (shared by all methods) known as the *assumption base*, which is an efficiently indexed collection of propositions that represent the collective knowledge state at any given moment, including perceptual knowledge. The assumption base is capable of serving as a communication buffer for the various methods. Finally, the control executive is itself a method, which directs the reasoning process incrementally by invoking various methods triggered by the contents of the assumption base.

We describe below three general-purpose methods for reasoning in the calculus we have presented. With these methods, the reasoning for the false-belief task can be performed in a handful of lines—essentially with one invocation of each of these methods. We stress that these methods are not ad hoc or hardwired to false-belief tasks. They are generic, and can be reused in any context that requires reasoning about other minds and satisfies the relevant preconditions. In particular, the methods do not contain or require any information specific to false-belief tasks.

- *Method 1*: This method, which we call M_1, shows that when an agent a_1 sees an agent a_2 perform some action-type α at some time point t, a_1 knows that a_2 knows that a_2 has carried out α at t. M_1 is parameterized over a_1, a_2, α, and t:
  1. The starting premise is that a_1 sees a_2 perform α at t:

$$\mathbf{S}(a_1, happens(action(a_2, \alpha), t)) \tag{1}$$

  2. Therefore, a_1 knows that the corresponding event has occurred at t:

$$\mathbf{K}(a_1, happens(action(a_2, \alpha), t)) \tag{2}$$

     This follows from the preceding premise and $[DR_4]$.

3. From $[A_4]$ and $[DR_2]$ we obtain:

$$\mathbf{K}(a_1, \forall\, a, \alpha, t\ .\ happens(action(a, \alpha), t) \Rightarrow \mathbf{K}(a, happens(action(a, \alpha), t))) \tag{3}$$

4. From (3) and $[DR_9]$ we get:

$$\mathbf{K}(a_1, happens(action(a_2, \alpha), t) \Rightarrow \mathbf{K}(a_2, happens(action(a_2, \alpha), t))) \tag{4}$$

5. From (4), (2), and $[DR_6]$ we get:

$$\mathbf{K}(a_1, \mathbf{K}(a_2, happens(action(a_2, \alpha), t))) \tag{5}$$

– *Method 2*: The second method, M_2, shows that when (1) it is common knowledge that a certain event e initiates a fluent f; (2) an agent a_1 knows that an agent a_2 knows that e has happened at a time t_1; (3) it is commonly known that $t_1 < t_2$; and (4) a_1 knows that a_2 knows that nothing happens between t_1 and t_2 to terminate the fluent f; then a_1 knows that a_2 knows that f holds at t_2. M_2 is parameterized over a_1, a_2, e, f, t_1, and t_2. We omit the definition due to space issues.

– *Method 3*: The last method, M_3, shows that when (1) it is common knowledge that t_1 is prior to t_2; (2) an agent a_1 knows that an agent a_2 knows that a fluent f holds at t_1; and (3) a_1 believes that nothing happened between t_1 and t_2 that would cause a_2 to believe that f no longer holds; then a_1 believes that a_2 believes that f holds at t_2:

1. The starting premises are:
   - $P_1 : \mathbf{C}(t_1 < t_2); P_2 : \mathbf{K}(a_1, \mathbf{K}(a_2, holds(f, t_1)))$;
     $P_3 : \mathbf{B}(a_1, \neg\exists e, t\ .\ \mathbf{B}(a_2, happens(e, t)) \wedge \mathbf{B}(a_2, t_1 < t < t_2) \wedge \mathbf{B}(a_2, terminates(e, f, t)))$.
2. From premise P_2, $[DR_5]$, and $[DR_{12}]$, we get:

$$\mathbf{B}(a_1, \mathbf{B}(a_2, holds(f, t_1))) \tag{6}$$

3. From $[A_6]$, $[DR_3]$, and universal specialization we get:

$$[\mathbf{B}(a_1, \mathbf{B}(a_2, holds(f, t_1))) \wedge \mathbf{B}(a_1, \neg\exists\, e, t\ .\ \mathbf{B}(a_2, happens(e, t)) \wedge \mathbf{B}(a_2, t_1 < t < t_2) \wedge \mathbf{B}(a_2, terminates(e, f, t)))] \Rightarrow \mathbf{B}(a_1, \neg\mathbf{B}(a_2, clipped(t_1, f, t_2))) \tag{7}$$

4. By P_3, (7), (6), conjunction introduction, and modus ponens, we get:

$$\mathbf{B}(a_1, \neg\mathbf{B}(a_2, clipped(t_1, f, t_2))) \tag{8}$$

5. From $[A_5]$, $[DR_{11}]$, and $[DR_2]$ we get:

$$\mathbf{K}(a_1, [\mathbf{B}(a_2, holds(f, t_1)) \wedge \mathbf{B}(a_2, t_1 < t_2) \wedge \neg\mathbf{B}(a_2, clipped(t_1, f, t_2))] \Rightarrow \mathbf{B}(a_2, f)) \tag{9}$$

6. From (9) and $[DR_5]$ we get:

$$\mathbf{B}(a_1, [\mathbf{B}(a_2, holds(f, t_1)) \wedge \mathbf{B}(a_2, t_1 < t_2) \wedge \neg\mathbf{B}(a_2, clipped(t_1, f, t_2))] \Rightarrow \mathbf{B}(a_2, holds(f, t_2)) \tag{10}$$

7. From P_1, $[DR_1]$, $[DR_5]$, and $[DR_{12}]$ we get:

$$\mathbf{B}(a_1, \mathbf{B}(a_2, t_1 < t_2)) \tag{11}$$

8. From (10), (6), (11), (8), and $[DR_{19}]$ we get:

$$\mathbf{B}(a_1, \mathbf{B}(a_2, holds(f, t_2))) \tag{12}$$

The correct conclusion for the false-belief task, produced by our implementation in a fraction of a second, is now obtained in the following manner:

1. Method M_1 fires, invoked with Alice, Bob, the action type *places*(*cookie*, *cabinet*), and time point *beginning*.
2. Axiom $[D_1]$ is repeatedly instantiated (via $[DR_{11}]$) with *Bob*, *cookie*, and *cabinet*.
3. Method M_2 fires, invoked with *Alice*, *Bob*, the action that *Bob* has placed the cookie in the cabinet, the fluent that the cookie is located in the cabinet, and the two time points *beginning* and *departure*.
4. Method M_3 fires, invoked with Alice, Bob, the fluent that the cookie is located in the cabinet, and the two time points *departure* and *return*.

5 Related work and conclusions

We have presented a formal system for representing and reasoning about certain important mental states, and used it to provide a formal analysis of false-belief tasks. Such tasks have been extensively discussed, particularly in the debate between theory-theory and simulation [6], but there are few rigorous models to be found. The only computational treatments of which we are aware are by Bello and Cassimatis [4] and by Watt [15]. Neither is based on a formal inference system. Goodman et al. [10] present a rational analysis of false belief reasoning based on causal Bayesian models.

Technically, our system is a multi-sorted multi-modal first-order logic. There is a growing recognition of the importance of quantification in epistemic contexts. Propositional multi-modal logics are just not sufficiently expressive. For instance, they cannot capture the difference between de dicto and de re knowledge. The versatility of first-order logic is necessary, alongside constructs such as common knowledge.

Our approach has been thoroughly proof-theoretic; we have not given a model-theoretic semantics for our logic. Coming up with an appropriate formal semantics for propositional attitudes is exceedingly difficult, and should not hold back experimentation with and implementation of various proof systems. The usual possible-world semantics [9] are mathematically elegant and well-understood, and they can be a useful tool in certain situations (e.g., in security protocol analysis). But they are notoriously implausible from a cognitive viewpoint.[5] The element of justification, for instance, which is central in our intuitive conception of knowledge, is entirely lacking from the formal semantics of epistemic logic. Indeed, knowledge, belief, desire, intention, provability, etc.,

[5] In an apt assessment of the situation, Anderson [1] wrote that epistemic logic "has been a pretty bleak affair." Fagin et al. [9] describe various attempts to deal with some of the problems arising in a possible-worlds setting, none of which has been widely accepted as satisfactory.

all receive the exact same formal analysis in possible-world semantics. That is simply not tenable.

At any rate, even in the standard Kripke framework, the question of how to combine quantification with epistemic constructs (particularly with common knowledge) is a difficult open problem: there have been no complete recursive axiomatizations, and indeed such logics are not even recursively enumerable [17]. Some decidable fragments have been investigated, such as the space of monodic formulas [14], but such restrictions limit expressivity, which in our view is a more important consideration. Indeed, we see no reason to insist on a computationally tractable—or even decidable—formalism, or on a complete logic, at the expense of expressivity. First-order logic is undecidable, but it is routinely used for the analysis and verification of a wide variety of extensional systems, by deploying interactive theorem-proving systems. Higher-order logic is both undecidable and incomplete, but it too is used widely for similar purposes. Things need not be different when it comes to the representation, analysis, and verification of rational agents. Our concern here has been to design and implement an expressive logic that can be readily used for such purposes; and to gain experience with constructing machine-checkable proofs in that logic, and particularly with writing powerful proof tactics in it.

LORA [18] is a multi-sorted language that extends first-order branching-time temporal logic with modal constructs for beliefs, desires, and intentions (drawing on the seminal work of Cohen and Levesque [5], and particularly on the BDI paradigm that followed it [12]), as well as a dynamic logic for representing and reasoning about actions. It does not have any constructs for perception or for common knowledge, and does not allow for the representation of events that are not actions. Its semantics for the propositional attitudes are standard Kripke semantics, with the possible worlds being themselves branching time structures. We are not aware of any implementations of LORA.

CASL (Cognitive Agents Specification Language) [13] is another system which combines an action theory, defined in terms of the situation calculus, with modal operators for belief, desire, and intention. Like LORA, CASL does not have any constructs for perception or for group knowledge (shared, distributed, or common). Also like LORA, the semantics of all intensional operators in CASL are given in terms of standard possible worlds. They are, in fact, explicitly defined in the higher-order logic PVS [11] by quantifying over states. Insofar as both LORA and CASL base their treatment of intensional operators on Kripke structures, they inherit all the conceptual difficulties associated with them. An advantage of CASL from our viewpoint is that it is implemented and allows for mechanized proofs, given in PVS. However, PVS is not readily programmable, and the use of sequents complicates the formulation of tactics. The natural deduction style of our framework is much more conducive to that task.

While preliminary experience with our implementation is encouraging, more work is needed to determine to what extent methods can facilitate reasoning in such a multi-modal first-order system. One issue related to efficiency is parallelism. The current implementation of our system is single-threaded, and is thus unable to realize one of the chief advantages of blackboard systems, namely, the independence of the specialists from one another and the fact that they can run

concurrently. (Of course, even with concurrency, the top control mechanism still needs to coordinate the writing into the blackboard, and that can be expected to become a bottleneck in some cases.) We plan to implement a multi-threaded version of the system in the future.

References

1. C. A. Anderson. The Paradox of the Knower. *Journal of Philosophy*, 80(6):338–355, 1983.
2. K. Arkoudas. Athena. `http://www.pac.csail.mit.edu/athena`.
3. K. Arkoudas and S. Bringsjord. Metareasoning for multi-agent epistemic logics. In *Fifth International Conference on Computational Logic In Multi-Agent Systems (CLIMA 2004)*, volume 3487 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 111–125. Springer, New York, 2005.
4. P. Bello and N. Cassimatis. Some Unifying Principles for Computational Models of Theory-of-Mind. 2006.
5. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
6. M. Davies and T. Stone, editors. *Folk Psychology: The Theory of Mind Debate*. Blackwell Publishers, 1995.
7. Daniel Dennett. Conditions of personhood. In *Brainstorms: Philosophical Essays on Mind and Psychology*, pages 267–285. Bradford Books, Montgomery, VT, 1978.
8. R. Engelmore and T. Morgan, editors. *Blackboard Systems*. Addison-Wesley, 1988.
9. R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about knowledge*. MIT Press, Cambridge, Massachusetts, 1995.
10. N. D. Goodman, E. B. Bonawitz, C. L. Baker, V. K. Mansinghka, A. Gopnik, H. Wellman, L. Schulz, and J. B. Tenenbaum. Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the Twenty-Eight Annual Conference of the Cognitive Science Society*, 2006.
11. S. Owre, N. Shankar, and J. M. Rushby. The PVS specification language (draft). Research report, Computer Science Laboratory, SRI International, Menlo Park, California, February 1993.
12. A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI-architecture. In *Proceedings of Knowledge Representation and Reasoning (KR&R-91)*, pages 473–484, 1999.
13. S. Shapiro, Y. Lespérance, and H. J. Levesque. The cognitive agents specification language and verification environment for multiagent systems. In *The First International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002*, pages 19–26, 2002.
14. H. Sturm, F. Wolter, and M. Zakharyaschev. Common Knowledge and Quantification. *Economic Theory*, 19:157–186, 2002.
15. S. N. K. Watt. *Seeing things as people*. PhD thesis, Knowledge Media Institute, Open University, UK, 1997.
16. H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128, 1983.
17. F. Wolter. First Order Common Knowledge Logics. *Studia Logica*, 65:249–271, 2000.
18. M. Wooldridge. *Rational Agents*. MIT Press, 2000.