A Logic of Knowledge and Belief for Recursive Modeling: Preliminary Report

Piotr J. Gmytrasiewicz and Edmund H. Durfee
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, Michigan 48109

Abstract

To make informed decisions in a multiagent environment, an agent needs to model itself, the world, and the other agents, including the models that those other agents might be employing. We present a framework for recursive modeling that uses possible worlds semantics, and is based on extending the Kripke structure so that an agent can model the information it thinks that another agent has in each of the possible worlds, which in turn can be modeled with Kripke structures. Using recursive nesting, we can define the propositional attitudes of agents to distinguish between the concepts of knowledge and belief. Through the Three Wise Men example, we show how our framework is useful for deductive reasoning, and we suggest that it might provide a meeting ground between decision theoretic and deductive methods for multiagent reasoning.

Introduction

In this paper, we develop a preliminary framework for recursive modeling in multiagent situations based on logics of knowledge and belief. If an intelligent agent is engaged in an interaction with another agent, it will have to reason about the other's knowledge, beliefs, and view of the world in order to interact with the other agent effectively. Reasoning about knowledge and belief is thus important not only for philosophy, but also for distributed and multiagent systems.

Presently, there seems to be no consensus among philosophers and AI researchers as to what particular properties concepts like knowledge and belief should have. As a result, a whole family of logics have appeared, with basically the same formalism but with differing sets of axioms. We summarize this formalism in the first section. After this, we go on to extend this formalism to the multiple agent case in a way that can be used for recursive modeling.

We describe how our framework can define propositional attitudes of agents in a way that provides for a natural distinction between the concepts of knowledge and belief. The intuition that we are able to formalize, suggested by Hintikka [Hintikka, 1962], is that statements about knowledge, unlike statements about belief, contain an element of commitment to this knowledge on the side of the agent making the statement.

We then compare our framework to other approaches and discuss the practical issues of creating the recursive hierarchy of models. Finally, we outline our framework's application to nested deductive reasoning using as an example the Three Wise Men puzzle, and we suggest how is might also be applied to coordination and communication using decision theory.

Classical Model

This section largely follows the presentation in [Halpern and Moses, 1990; Halpern and Moses, 1991]. The classical model for reasoning about knowledge and belief is the *possible worlds* model. The basic intuition here is that an agent has a limited view of the world and cannot be sure in what state the world really is. Hence, there are several states of the world that an agent considers possible, called possible worlds. A formal tool for reasoning about possible worlds is a *Kripke structure*. A Kripke structure, M, for an agent is (S, π, R), where S is a set of possible worlds, π is a truth assignment to the primitive propositions for each possible world (i.e. $\pi(s)(p) \in \{true, false\}$ for each state $s \in S$ and each primitive proposition p), and R is a binary relation on the possible worlds, called the possibility relation. The truth assignment π can used to define a relation, $\models$,between a proposition, p, and a possible world, s, of a structure M, as follows:

$(M, s) \models$ iff $\pi(s)(p) = true$.

Let us examine an example (based on [Halpern and Moses, 1991]). Let p denote a primitive proposition, then $\pi(s)(p) = true$ describes the situation in which p holds in state s of structure M. Let us take an example set of possible worlds consisting of s, t and

[0]This research was supported, in part, by the Department of Energy under contract DG-FG-86NE37969, and by the National Science Foundation under grant IRI-9015423 and PYI award IRI-9158473.

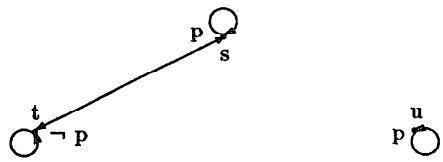

Figure 1: **Diagram of a Kripke Structure**

u: $S = \{s, t, u\}$. Assume that proposition p is true in states s and u but false in t (so that $\pi(s)(p) = \pi(u)(p) = true$ and $\pi(t)(p) = false$) and that a particular agent cannot tell the states s and t apart, so that $R = \{(s,s),(s,t),(t,s),(t,t),(u,u)\}$. This situation can be diagrammed, as in Figure 1, where the possibility relation between worlds is depicted as a vector, as between s and t, denoting that in the state s the agent considers state t possible. Now, in state s, the agent is uncertain whether it is in s or in t, and since p holds in s and does not in t, we can conclude that the agent does not know p. In state u, the agent can be said to know that p is true, since the only state accessible from u is u itself and p is true in u. Considerations of this sort lead us to the modal operator K, denoting knowledge. According to the classical definition, an agent in some state is said to know a fact if this fact is true in all of the worlds that the agent considers possible.

In multiagent situations different agents might have different possibility relations. The model proposed in [Halpern and Moses, 1990; Halpern and Moses, 1991] for the case of multiple agents, named 1 through n, is a Kripke structure $M = (S, \pi, R_1, R_2, ..., R_n)$. Thus, the possibility relation of each of the agents is included directly in M. While a straightforward extension of a single agent case, we have found this representation problematic when one wants to consider agents reasoning about other agents. Specifically, we would like the possibility relation that agent 1 ascribes to agent 2 to potentially differ from agent 2's true possibility relation. Thus, each agent might have many possibility relations associated with it, depending on who's perspective is being considered. As we detail next, our own approach for treating with multiple agents involves a nesting, rather than an indexing, of possible worlds that permits different viewpoints to coexist and that allows a distinction between the concepts of knowledge and belief. After describing our approach in the next section, we compare it to related work in more detail.

Personal Recursive Kripke Structures

Our formalism views an agent's knowledge from its own, personal perspective so that the formalism can be used by an agent when interacting with others. Let us consider a set of n interacting agents, named 1 through n. Without loss of generality, we will consider the situation from the perspective of agent 1, which is in a world about which it has limited information. We will call the representation of this information that 1 has its *view*. Based on its view of the world, 1 can form a set of possible worlds that are consistent with its limited view, and represent them in its Kripke structure. Each of these worlds can be described by a set Φ, of primitive propositions p.

Since there are other agents around, the agent should wonder about their views of the world. In the formalism we are proposing, agent 1 forms its model of the other agents' views in each of the worlds it considers possible. Thus, each of the possible worlds s_k is *augmented* with structures representing the knowledge the agent attributes to each of the other agents in *this* world. It is natural to postulate that these structures themselves be Kripke structures. We are getting a recursively nested Kripke structure of agent 1: $RM^1 = (S, \pi, R)$, where the elements of S are augmented possible worlds: $s'_k = (s_k, RM^2_k, ..., RM^i_k, ..., RM^n_k)$. The first element, s_k, is a classical possible world described by a set of primitive propositions; the other elements are recursively defined Kripke structures of the other agents, corresponding to their limited views of the possible world s_k. Thus, $RM^i_k = (S^i_k, \pi^i_k, R^i_k)$ in which S^i_k is the set of augmented possible worlds of agent i in the world s_k. In the above formulation, the π relation is, as before, the truth assignment to the primitive propositions for each possible world, s_k. The binary relation R in RM^1 is a possibility relation defined over the set of augmented possible worlds S. The truth assignment π can be used to define a binary relation, $\models$, between a proposition, p, and a possible world, s_k, of a structure RM^1, as follows:

$(RM^1, s_k) \models p$ iff $\pi(s_k)(p) = true$.

The personal recursive Kripke structure RM defined above can serve to define a number of concepts useful in multiagent reasoning, in a manner analogous to one used in the case of classical Kripke structures (we follow the spirit of [Hintikka, 1962; Hughs and Cresswell, 1972]). These concepts are referred to as *propositional attitudes*.

Propositional Attitudes of a Single Agent

Based on its recursive Kripke structure, $RM^1 = (S, \pi, R)$, agent 1 can say that it *knows* that p holds, written as $K_1 p$, if

- $(RM^1, s_k) \models p$ for s_k in all $s'_k \in S$, i.e., if p is true in all of the possible worlds consistent with agent 1's view of the world.

In these circumstances, agent 1 can also say that it believes p, and thus, there is no distinction between the concepts of knowledge and belief when agent 1 reasons or communicates facts about its own view of the world. It is, then, the same for agent 1 to assert "I know p", as to assert "I believe p". Our convention of equating the concepts of knowledge and belief in this case differs from some of the established conventions that differentiate between these two concepts based on the

properties of the possibility relation R. In particular, "knowledge" is sometimes reserved only for assertions that an agent makes that are true in the actual world (as assessed by some correct and omniscient agent). The uniqueness of our approach stems from the fact that we consider the agent's knowledge from its own, personal perspective. Because the real world, and its complete description, cannot be known with certainty by the agent, it cannot be sure that the real world is among the worlds that it considers possible. Consequently, there is no way that the agent can tell its knowledge and belief apart.

In the remainder of this paper, therefore, we will use K_1p to denote agent 1's making a statement, p, based on its Kripke structure, RM^1, with the understanding that K_1p is always equivalent to B_1p. Later, however, we will show how the difference between knowledge and belief arises intuitively when an agent makes assertions about other agents. Now we continue with the propositional attitudes of a single agent:

Agent 1 can say that it *knows whether* p holds, written as W_1p, if

- K_1p or $K_1\neg p$.

Further, agent 1 can say that the proposition p is *possible*, written as P_1p, if

- $\exists s'_k \in S$ such that $(RM^1, s_k) \models p$, i.e., p is true in at least one of the worlds consistent with agent 1's view of the world.

And, agent 1 can say that the proposition p is *contingent*, written as C_1p, if

- $\neg W_1p$, i.e., agent 1 does not know whether p.

Propositional Attitudes of Other Agents

To reason about the knowledge and beliefs of others, agent 1, with its structure $RM^1 = (S, \pi, R)$, can inspect the structures of the other agents, $RM^i_k = (S^i_k, \pi^i_k, R^i_k)$, in its augmented possible worlds. Thus, this kind of reasoning always pertains to what agent 1 *thinks* other agents are thinking. A number of propositional attitudes describing other agents can be defined as follows.

Belief

Agent 1 can say that agent i believes p in a possible world s_k, written $K_1B_i^{s_k}p$, if

- $(RM^1, s^i_{k,l}) \models p$ for $s^i_{k,l}$ in all $s'^i_{k,l} \in S^i_k$, i.e., p holds in all of the worlds that agent 1 thinks that agent i considers possible in s_k.

Agent 1 can say that agent i believes a fact p, denoted as K_1B_ip, if

- $K_1B_i^{s_k}p$ for s_k in all $s'_k \in S$, i.e., if agent i believes p in all of agent 1's possible states of the world.

The definitions of possibility and contingency for other agents can be constructed analogously to belief.

Note that the definitions of the propositional attitude of belief of agent i above did not contain any reference to what agent 1 knows (believes) of the world. Therefore, we can say that if agent 1 makes statements about agent i's beliefs, the propositional attitude of agent 1 would not be revealed. This can be contrasted with agent 1 speaking about agent i in terms of knowledge, as we now see.

Knowledge

Agent 1 can say that agent i knows that p holds in possible world s_k, written as $K_1K_i^{s_k}p$, if

- $(RM^1, s_k) \models p$, i.e., p holds in s_k, and if
- $(RM^1, s^i_{k,l}) \models p$ for $s^i_{k,l}$ in all $s'^i_{k,l} \in S^i_k$, i.e., p holds in all of the worlds that agent 1 thinks that agent i considers consistent with s_k.

Agent 1 can say that agent i knows a fact p, written as K_1K_ip, if

- $K_1K_i^{s_k}p$ for s_k in all $s'_k \in S$, i.e., agent i knows p in all of agent 1's possible states of the world. Let us note that the above also implies that agent 1 knows p.

Analogously, agent 1 can say that agent i knows whether a fact p holds in possible world s_k, written as $K_1W_i^{s_k}p$, if

- $K_1K_i^{s_k}p$ or $K_1K_i^{s_k}\neg p$.

And agent 1 can say that agent i knows whether a fact p holds, written K_1W_ip, if

- K_1K_ip or $K_1K_i\neg p$.

Relations Between Knowledge and Belief

It is important to note that the definitions agent 1 uses to characterize agent i in terms of knowledge involve a comparison between i's view of the world and agent 1's view. Thus, an agent that makes statements about the knowledge of other agents expresses its own commitment to this knowledge. Statements about others' beliefs, on the other hand, do not involve this commitment, and the notions of knowledge and belief *differ*. Our definitions, therefore, capture the "knowledge as a justified, true belief" paradigm of modal logic.

To investigate the relation between the concepts of knowledge and belief a little further, let us introduce some helpful notation. We will call the relation between the possible worlds s_k in the augmented worlds, $s'_k = (s_k, ..., RM^i_k, ...)$ belonging to the set S of structure $RM^1 = (S, \pi, R)$, and the possible worlds $s^i_{k,l}$ in the augmented worlds $s'^i_{k,l}$ belonging to the set S^i_k of the structure $RM^i_k = (S^i_k, \pi^i_k, R^i_k)$, a *subordination*[1] relation for agent i in the world s_k: $Sub_i^{s_k} = \{(s_k, s^i_{k,l})\}$. The worlds, s_k and $s^i_{k,l}$, that are connected via a subordination relation will be called a parent world, and a child world, respectively. Thus, the subordination relation connects parent worlds to children, that themselves can be parents of other worlds, and so on.

Theorem 1. If the subordination relation in a personal recursive Kripke structure, RM^1, is reflexive for all agents, then the concepts of knowledge and belief, that agent 1 uses, are equivalent.

Proof: The definitions of $K_1K_i^{s_k}p$ and $K_1B_i^{s_k}p$ in the previous section ensure that $K_1K_i^{s_k}p$ implies $K_1B_i^{s_k}p$.

[1] Our choice of this term is motivated by such relations investigated in [Hughs and Cresswell, 1972; Hughs and Cresswell, 1984].

To establish the implication in the other direction note that, if the subordination relation is reflexive then the world s_k is also one of the $s^i_{k,l}$ worlds. Since $K_1 B_i^{s_k} p$ demands that $(RM, s^i_{k,l}) \models p$ for all $s^i_{k,l}$ worlds, it follows that $(RM, s_k) \models p$. Thus, $K_1 B_i^{s_k} p$ implies $K_1 K_i^{s_k} p$. The equivalence of $K_1 K_i^{s_k} p$ and $K_1 B_i^{s_k} p$ for all of the worlds s_k ensures the equivalence of $K_1 K_i p$ and $K_1 B_i p$.

Restated in terms of *views*, Theorem 1 says that, if an agent can be sure that another agent's view of any possible world is guaranteed to be *correct*, then the distinction between knowledge and belief ceases to exist. It is, therefore, the possibility of the other agent's view to be an *incorrect* description of a world, as opposed to being only a partial description of it, that allows for the intuitively appealing distinction between knowledge and belief.

Given that our formulations have shown that distinctions between knowledge and belief arise when one agent reasons about another, we might ask what happens when an agent treats itself in this way. An agent doing so amounts to *introspection*. We call the corresponding concepts introspective knowledge, $K_1 K_1 p$, and introspective belief, $K_1 B_1 p$.

To enable introspection, we can formally modify the personal recursive Kripke structure, $RM^1 = (S, \pi, R)$, of agent 1, and include in an augmented world, s'_k, a structure, RM^1_k, representing the information contained in the view agent 1 would have in each of its possible worlds, s_k. The augmented worlds contained in S are now: $s'_k = (s_k, RM^1_k, RM^2_k, ..., RM^i_k, ..., RM^n_k)$. The fact that RM^1_k is to represent the view the agent would have in s_k suggests that RM^1_k describe a portion of RM^1 visible[2] from s_k. If this is taken to be the case, we obtain the following theorems:

Theorem 2. If the accessibility relation, R, in the personal recursive Kripke structure, $RM^1 = (S, \pi, R)$, of agent 1 is reflexive, then the introspective knowledge of a proposition, $K_1 K_1 p$, is equivalent to introspective belief, $K_1 B_1 p$, for this agent.

Proof: Introspective knowledge, $K_1 K_1 p$, clearly implies introspective belief, $K_1 B_1 p$. If R is reflexive, then, using the notation above, every world s_k belongs to the set of $s^1_{k,l}$ worlds in RM^1_k. Introspective belief, $K_1 B_1 p$, demands that p be true in all worlds $s^1_{k,l}$, and thus also in s_k, for every such s_k. For reflexive R, therefore, $K_1 B_1 p$ implies $K_1 K_1 p$.

Theorem 3. If the accessibility relation, R, in the personal recursive Kripke structure, $RM^1 = (S, \pi, R)$, of agent 1 is universal, then the introspective knowledge and introspective belief of an agent are equivalent to the agent's knowledge (and, of course, belief.).

Proof: Introspective knowledge, $K_1 K_1 p$, clearly implies knowledge, $K_1 p$. Note that, for R universal, the set of possible worlds, s_k, in the set S is the same as the set of the worlds, $s^1_{k,l}$, accessible from s_k. Therefore, knowledge of a proposition, $K_1 p$, demanding that p hold in all of the worlds s_k, implies that p holds in all of the worlds $s^1_{k,l}$. For a universal R, therefore, knowledge, $K_1 p$, implies introspective knowledge, $K_1 K_1 p$. In this case, according to Theorem 2, introspective knowledge is also equivalent to introspective belief.

The theorems above provide a certain amount of guidance as to the properties of relations holding among the possible worlds that one can reasonably postulate in practical situations. It seems that it may be desirable to be able to make a distinction between the concepts of knowledge and belief used by agents to describe other agents. Thus, we should not demand that the subordination relation, holding between the parent and the children possible worlds, be reflexive. On the other hand, it seems desirable to demand that the accessibility relations, R, holding among the sibling worlds themselves, be not only reflexive, but also universal. This property ensures that the introspective knowledge of an agent will be no different than its knowledge.

The nonreflexive subordination relation, together with a universal relation among the sibling worlds in a personal recursive Kripke structure, provides it with a unique composition. It consists of clusters of the sibling worlds, interconnected via a universal accessibility relation, overlayed over a tree whose branches consist of the monodirectional subordination relation[3]. Of course, while the above composition does provide for a reasonable set of properties, other models may be equally interesting. Some of them might not provide for equivalence among introspective knowledge, introspective belief and knowledge, and it remains to be investigated whether they correspond to any realistic situations.

Comparison to Related Work

As we mentioned before, the Kripke structure suggested for n agents in [Halpern and Moses, 1991] is a tuple $M = (S, \pi, R_1, R_2, ..., R_n)$. Unlike our definition, the possibility relation of each of the agents is included directly in M. Important consequences of this are revealed when the the agents' knowledge about each other's knowledge is considered. It is suggested that, in order for the agents to be able to consider somebody else's knowledge, they have to have access to their possibility relation. So, in $M = (S, \pi, R_1, R_2, ..., R_n)$, agent 1 can peek into R_2 and claim what agent 2 knows or not. Also, in order for agent 1 to find out what agent 2 knows about agent 1, the possibility relation R_1 has to be consulted, which is the one summarizing agent 1's knowledge itself.

In general, one can say that viewing the knowledge of

[2] We say that the world s_1 sees the world s_2 if $(s_1, s_2) \in R$.

[3] Our nomenclature is again motivated by some of the models analyzed in [Hughs and Cresswell, 1984].

the agents via the structure $M = (S, \pi, R_1, R_2, ..., R_n)$ amounts to taking an *external* view of their knowledge, in that it is an external observer that lists the agents' possible worlds in S and summarizes their knowledge about the real world in relations R_i. It is then counter-intuitive to postulate that the agents themselves can inspect the possibility relation of the other agents. It is also surprising that the agents, wondering how others view them, look into their own possibility relations.

Our approach avoids the above drawbacks; the personal recursive Kripke structure represents the information an agent has about the world from its own perspective, and the information it has about the other agents' knowledge is represented as a model the agent has of the others. The model of the other agents may contain information the original agent has about how it is itself modeled by the other agents, but this may be quite different from the information the initial agent actually has.

The idea that the recursive nesting of knowledge levels is necessary for analyzing the interactions in multiagent systems has been present for quite a while in the area of game theory, and recently received attention in the AI literature, for instance from Fagin and others in [Fagin *et al.*, 1991]. The most obvious difference between their approach and ours is that we use a suitably modified Kripke structures, while the authors of [Fagin *et al.*, 1991], after noting that the classical extension of the Kripke structures to the multiagent case (mentioned above) is inadequate, develop a complementary concept of *knowledge structures*. The motivation and basic intuitions behind knowledge structures is very similar to ours. Thus, knowledge structures represent recursive, potentially infinite, nesting of information that agents have about other agents, just as our recursive Kripke structures do. An important distinction is that knowledge structures, as defined in [Fagin *et al.*, 1991], do not assume a personal perspective from an agent's point of view; they instead contain information of all of the agents in the environment, in addition to the description of the environment itself, and thus amount to an *external* view of the multiagent situation. While the authors provide for the definition of an individual agent's view of the knowledge structure, which should correspond to our personal Kripke structure, its function is unspecified. Another difference is that we are able to provide a clear and intuitive distinction between the concepts of knowledge and belief within our single recursive framework. Our motivation here is very similar to one presented in [Shoham and Moses, 1989]. This work, although using quite a different approach, also attempts to derive the connections between knowledge and belief within a single framework.

The relation between the introspective knowledge and knowledge of agents has received attention in the AI literature, for example from Konolige in [Konolige, 1986], who provides a discussion of some of the properties of introspection: fulfillment and faithfulness. Although Konolige does not employ possible worlds semantics in his considerations, it seems that these properties can be arrived at using our formalism. Establishing further relations between these approaches is a goal of our future research.

The issues of recursive nesting of beliefs are also of interest in [Wilks *et al.*, 1991; Wilks and Bien, 1983], but we find this other work most relevant to the heuristic construction of the recursive models, described in the next section.

Construction of Personal Recursive Kripke Structures

The personal recursive Kripke structure provides a formal model that the agents engaged in a multiagent interaction can use to reason about the other agents' knowledge and belief. Within this framework they can construct models of the other agents' knowledge. As we mentioned before, the concept that might be useful for constructing the models is an agent's view of the world. A view is essentially a partial description of the world that an agent has, given its knowledge base, its location in the environment, sensors it has, etc.

To illustrate what we mean by a view, let us consider an example of two agents, 1 and 2, facing each other. Imagine that each of the agents is wearing a hat and can see the hat of the other agent, while being unable to see its own hat. The problem agents are facing is to determine whether their own hat is black or white. Assume that both hats are black, and that it is known that hats can be only black or white. Agent 1's view of this situation may be a partial description of the two hats; agent 1 knows that agent 2 is wearing a black hat, but the color of its own hat is unknown to agent 1 and might be represented by a "?", for instance. In a frame-like language this information may be represented as slots with values assigned to them:

Agent 1' view:
Agent-1-hat – ?
Agent-2-hat – black

Out of its incomplete description of the world, agent 1 can construct two possible worlds that are consistent with its view:

Possible World 1:	Possible World 2:
Agent-1-hat – black	Agent-1-hat – white
Agent-2-hat – black	Agent-1-hat – black

Agent 1 can also construct agent 2's views of these worlds:

Agent 2's view of PW1:	Agent 2's view of PW2:
Agent-1-hat – black	Agent-1-hat – white
Agent-2-hat – ?	Agent-2-hat – ?

These views lead, in turn, to agent 2's possible worlds in each case, as depicted in Figure 2, where the agents' views were also included.

Let us note a few things about constructing views and possible worlds. First, agent 1 chose to describe the world in terms of primitive propositions denoting

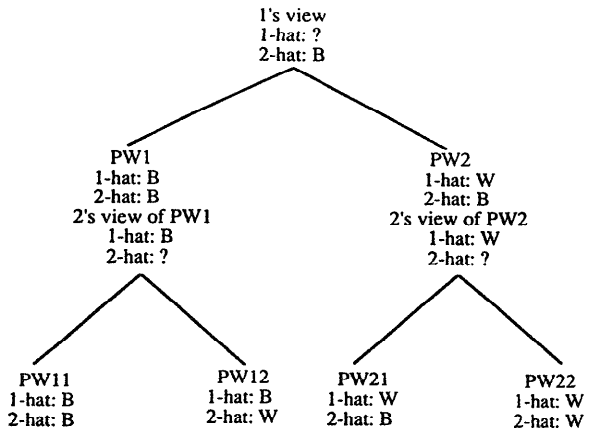

Figure 2: **Recursive Kripke Structure of Agent 1**

agents' hats being black or white. The reason it chose these particular propositions is that these are the *relevant* ones for this situation. Thus, agent 1 did not include propositions describing the number of hairs on their heads, because this information is clearly irrelevant. What made the colors of the hats relevant and the number of hairs irrelevant is, in the case of this example, clear in the statement of the problem they face, along with the fact that there is no apparent connection between the number of hairs and hats being black or white.

Now, let us assume for a minute that agent 1 received confidential information stating that the hat of the agent with more hairs is black. In this case, it would be advisable for agent 1 to include the information about the number of hairs in its view. This information, then, would find its way into agent 1's possible worlds, but clearly agent 1 would not be justified in including this information in agent 2's views, since the information agent 1 received was confidential and agent 2 is not *aware* of it.

The problems of relevance and awareness are difficult issues that have to be dealt with when one engages in recursive modeling. In the simple example above, they were easily and intuitively resolved, but in real life situations things may be more difficult. In these cases, strong heuristics for properly determining relevance and awareness are needed. It seems that the work by Yoric Wilks and his colleagues [Wilks *et al.*, 1991; Wilks and Bien, 1983] on belief ascription addresses these issues. They propose a number of heuristics, including a relevance heuristic, percolation heuristic, and pushing down environments. They capture the intuitive assumption that the other agents are aware of everything that I am aware of, unless my beliefs are atypical or confidential (as was the confidential information above).

Let us also note that the construction of the personal recursive Kripke structure described above does not bottom out, and the recursion describing the nesting of the knowledge and belief seems to go on forever. While this is an uncomfortable prospect, in the next section we will show that in many practical applications the agents can reach useful conclusions with the recursive Kripke structures cached down to a finite level.

Applications of Personal Recursive Kripke Structures

There are possibly a number of ways the information contained in a personal recursive Kripke structure can be used in multiagent reasoning. A class of problems that can be tackled deductively using this information includes the Three Wise Men problem, together with similar ones: Muddy Children, Cheating Husbands, etc., described in [Moses *et al.*, 1983].

For brevity, we will sketch the solution of the scaled-down version of the Three Wise Man puzzle, easily generalizable to the rest of the problems. The Two Wise Men puzzle describes two "wise" agents that, as described before, wear hats so that they can see the other's hat but not their own. The ruler of the kingdom the agents live in, intent on testing their wisdom, announces: "At least one of you is wearing a black hat". Then, he asks agent 2: "Do you know whether your hat is black or white?". Agent 2's answer is "No". The King then asks agent 1 the same question. And 1's answer is "My hat is black".

To trace the reasoning of agent 1, its recursive Kripke structure, developed down to the second level of modeling will be needed. We depict it in Figure 2, showing the state of knowledge of agent 1 before the King's announcement. PW1 and PW2 stand for the two relevant worlds agent 1 considers possible. They are described by propositions stating that the hat of agent 2 is black, p (or white, $\neg p$), and that the hat of agent 1 is black, q (or white, $\neg q$). In each of these worlds, the views of agent 2 are created by agent 1, and these lead to two worlds agent 1 thinks agent 2 considers possible, described by the same set of propositions.

After the King announces: "At least one of you is wearing a black hat" the state of knowledge of agent 1 changes, since agent 1 knows that agent 2 considers impossible all of the worlds in which both hats are white. By deduction, PW22 is impossible. Thus, in the possible world in which the hat of agent 1 is white, PW2, agent 2 knows that its own hat is black: $K_1 K_2^{PW2} p$, which also implies that it knows whether p: $K_1 W_2^{PW2} p$. In the possible world in which the hat of agent 1 is black, PW1, agent 2 does not know whether its hat is black or white: $K_1 \neg W_2^{PW1} p$. In this situation, the answer of agent 2 that it does not know whether its hat is black or white solves the puzzle for agent 1, since it deductively identifies PW2 as impossible and PW1 as the only possible world.

The solution of the Three Wise Man puzzle involves

the use of the recursive structure developed down to the third level, while n muddy children require n levels and the deduction is analogous. In the example problems discussed above, the crucial part of their solution is the definitions of propositional attitudes of other agents in various possible worlds. These concepts enable the reasoner to move upward in the tree of recursive models and deductively eliminate some of the possible worlds as new information warrants.

We have previously studied similar propagation of information upward in a recursive tree of payoff matrices in [Gmytrasiewicz *et al.*, 1991a; Gmytrasiewicz *et al.*, 1991b]. In fact, the recursive hierarchy of payoff matrices is a personal recursive Kripke structure, with the information describing the possible worlds cast in the form of payoff matrices. In this work, we applied decision and game theory to facilitate coordination, cooperation, and communication among autonomous agents. Unlike the deductive reasoning used in the Three Wise Men puzzle, our previous work employed the intentionality principle and expected utility calculations. We have noticed that the two approaches actually complement each other. The decision-theoretic modeling is built within a formal framework of reasoning about knowledge and belief of other agents. Since the deductive powers of this formalism are capable of dealing only with a limited spectrum of problems (in the Three Wise Men puzzle family), they are complemented with the capabilities of decision-theoretic reasoning when it comes to predicting other agents' actions and to effective communication. Moreover, the decision-theoretic calculations require that probabilities be assigned to possible worlds [Halpern, 1989]. Our ongoing work includes tying these two approaches together more formally.

Conclusion

We have developed a preliminary framework based on a possible worlds semantics, modeled by the personal recursive Kripke structure, that autonomous agents can use to organize their knowledge. Our model can serve as a semantic model for a logic of knowledge and belief, creating a natural and intuitive distinction between these concepts. This logic can be used to deductively reason about the knowledge and beliefs of the other agents, as in the Three Wise Men puzzle. We suggest that our model can also be used as a basis for the type of decision-theoretic reasoning in multiagent environments that we have found useful for studying coordination, cooperation, and communication. Our future work will address extending the logical framework (axiomatization, completeness, consistency, and complexity of decision procedures), and will explore the relationships between our deductive and decision-theoretic recursive models.

Acknowledgments

The authors would like to thank Yoav Shoham, Joseph Halpern, and the anonymous reviewers for their helpful comments on many aspects of this work.

References

[Fagin *et al.*, 1991] Ronald Fagin, Joseph Y. Halpern, and Moshe Y. Vardi. A model-theoretic analysis of knowledge. *Journal of the ACM*, (2):382–428, April 1991.

[Gmytrasiewicz *et al.*, 1991a] Piotr J. Gmytrasiewicz, Edmund H. Durfee, and David K. Wehe. A decision-theoretic approach to coordinating multiagent interactions. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 62–68, August 1991.

[Gmytrasiewicz *et al.*, 1991b] Piotr J. Gmytrasiewicz, Edmund H. Durfee, and David K. Wehe. The utility of communication in coordinating intelligent agents. In *Proceedings of the National Conference on Artificial Intelligence*, pages 166–172, July 1991.

[Halpern and Moses, 1990] Joseph Y. Halpern and Yoram Moses. A guide to the modal logics of knowledge and belief. Technical Report 74007, IBM Corporation, Almaden Research Center, 1990.

[Halpern and Moses, 1991] Joseph Y. Halpern and Yoram Moses. Reasoning about knowledge: a survey circa 1991. Technical Report 50521, IBM Corporation, Almaden Research Center, 1991.

[Halpern, 1989] Joseph Y. Halpern. An analysis of first-order logics of probability. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1375–1382, August 1989.

[Hintikka, 1962] Jaakko Hintikka. *Knowledge and Belief.* Cornell University Press, 1962.

[Hughs and Cresswell, 1972] G. E. Hughs and M. J. Cresswell. *An Introduction to Modal Logic.* Methuen and Co., Ltd., London, 1972.

[Hughs and Cresswell, 1984] G. E. Hughs and M. J. Cresswell. *An Introduction to Modal Logic.* Methuen and Co., Ltd., London, 1984.

[Konolige, 1986] Kurt Konolige. *A Deduction Model of Belief.* Morgan Kaufmann, 1986.

[Moses *et al.*, 1983] Y. Moses, D. Dolev, and J. Y. Halpern. Cheating husbands and other stories: a case study in common knowledge. Technical report, IBM, Almaden Research Center, 1983.

[Shoham and Moses, 1989] Yoav Shoham and Yoram Moses. Belief as defeasible knowledge. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1168–1172, Detroit, Michigan, August 1989.

[Wilks and Bien, 1983] Y. Wilks and J. Bien. Beliefs, points of view, and multiple environments. *Cognitive Science*, 7:95–119, April 1983.

[Wilks *et al.*, 1991] Y. Wilks, J. Barden, and J. Wang. Your metaphor or mine: Belief ascription and metaphor interpretation. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 945–950, August 1991.