Reasoning about only knowing with many agents*

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120–6099
halpern@almaden.ibm.com
408-927-1787

Abstract

We extend two notions of "only knowing", that of Halpern and Moses [1984], and that of Levesque [1990], to many agents. The main lesson of this paper is that these approaches do have reasonable extensions to the multi-agent case. Our results also shed light on the single-agent case. For example, it was always viewed as significant that the HM notion of only knowing was based on S5, while Levesque's was based on K45. In fact, our results show that the HM notion is better understood in the context of K45. Indeed, in the single-agent case, the HM notion remains unchanged if we use K45 (or KD45) instead of S5. However, in the multi-agent case, there are significant differences between K45 and S5. Moreover, all the results proved by Halpern and Moses for the single-agent case extend naturally to the multi-agent case for K45, but not for S5.

1 Introduction

There has been over twelve years of intensive work on various types of nonmonotonic reasoning. Just as with the work on knowledge in philosophy in the 1950's and 1960's, the focus has been on the case of a single agent reasoning about his/her environment. However, in most applications, this environment includes other agents. Surprisingly little of this work has focused on the multi-agent case. To the extent that we can simply represent the other agents' beliefs as propositions (so that "Alice believes that Tweety flies" is a proposition just like "Tweety flies"), then there is no need to treat the other agents in a special way. However, this is no longer the case if we want to reason about the other agents' reasoning. In fact, we need to reason about other agents' reasoning when doing multi-agent planning; moreover, much of this reasoning will be nonmonotonic (see [Morgenstern 1990] for examples).

*The work of the author is sponsored in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.

In this paper, we show how to extend to the multi-agent case two related approaches to nonmonotonic reasoning, both based on the notion of "only knowing": that of Halpern and Moses [1984] (hereafter called the HM notion) and that of Levesque [1990]. The main lesson of the paper is that, despite some subtleties, both approaches do have reasonable extensions to the multi-agent case. Our results also shed light on the single-agent case. For example, it was always viewed as significant that the HM notion of only knowing was based on S5, while Levesque's was based on K45.[1] In fact, our results show that the HM notion is better understood in the context of K45. Indeed, in the single-agent case, the HM notion remains unchanged if we use K45 (or KD45) instead of S5. However, in the multi-agent case, there are significant differences between K45 and S5. Moreover, as we show here, all the results proved by Halpern and Moses for the single-agent case extend naturally to the multi-agent case for K45, but not for S5.

2 The HM notion of "all I know"

The intuition behind the HM notion is straightforward: In each world of a (Kripke) structure, an agent considers a number of other worlds possible. In the case of a single agent whose knowledge satisfies S5 (or K45 or KD45), we can identify a world with a truth assignment, and a structure with a set of truth assignments. Truth in these logics is with respect to *situations* (W, w), consisting of a structure W, representing the set of truth assignments (worlds) that the agent considers possible, and a truth assignment w, intuitively representing the "real world".[2] The more worlds an agent considers possible, the less he knows. Thus, (W, w) is the situation where α is all that is known if (1) $(W, w) \models L\alpha$ (so that the agent knows α) and (2) if $(W', w') \models L\alpha$, then

[1] Due to lack of space, we are forced to assume that the reader is familiar with standard notions of modal logic. Details can be found in [Hughes and Cresswell 1968; Halpern and Moses 1992].

[2] For KD45, we require that W be nonempty; for S5, we require in addition that $w \in W$.

$W' \subseteq W$. If there is no situation (W, w) satisfying (1) and (2), then α is said to be *dishonest*; intuitively, it cannot then be the case that "all the agent knows" is α. A typical dishonest formula is $Lp \vee Lq$. To see that this formula is dishonest, let W_p consist of all truth assignments satisfying p, let W_q consist of all truth assignments satisfying q, and let w satisfy $p \wedge q$. Then $(W_p, w) \models Lp \vee Lq$, and $(W_q, w) \models Lp \vee Lq$. Thus, if $Lp \vee Lq$ were honest, there would have to be a situation (W, w') such that $(W, w') \models Lp \vee Lq$ and $W \supseteq W_p \cup W_q$ It is easy to see that no such situation exists. Notice that in the case of one agent, the notions of honesty and "all I know" coincide for K45, KD45, and S5.

We want to extend this intuition to the multi-agent case and—in order to put these ideas into better perspective—to other modal logics. We consider six logics, three that do not have negative introspection, K_n, T_n, and $S4_n$, and three that do, $K45_n$, $KD45_n$, and $S5_n$.[3] Below, when we speak of a modal logic $\mathcal{S}$, we are referring to one of these six logics; we refer to $K45_n$, $KD45_n$ and $S5_n$ as *introspective logics*, and K_n, T_n, and $S4_n$ as *non-introspective logics* (despite the fact that positive introspection holds in $S4_n$). As we shall see, "all I know" behaves quite differently in the two cases.

There are philosophical problems involved in dealing with a notion of "all I know" for the non-introspective logics. What does it mean for an agent to say "all I know is α" if he cannot do negative introspection, and so does not know what he doesn't know. Fortunately, there is another interpretation of this approach that makes sense for arbitrary modal logics. Suppose that a says to b, "All i knows is α" (where i is different from a and b). If b knows in addition that i's reasoning satisfies the axioms of modal logic $\mathcal{S}$, then it seems reasonable for b to say that i's knowledge is described by the "minimal" model satisfying the axioms of $\mathcal{S}$ consistent with $L_i\alpha$, and for b to view a as dishonest if there is no such minimal model.

Of course, the problem lies in defining what it means for a model to be "minimal". Once we consider multi-agent logics, or even nonintrospective single-agent logics, we can no longer identify a possible world with a truth assignment. It is not just the truth assignment at a world that matters; we also need to consider what other worlds are accessible from that world. This makes it more difficult to define a reasonable notion of minimality. To deal with this problem, we define a canonical collection of objects that an agent can consider possible. These will act like the possible worlds in the single-agent case. The kind of objects we consider depends on whether we consider the introspective or the non-introspective logics, for reasons that will become clearer below. We start with the non-introspective case.

Fix a set Φ of primitive propositions, and agents $1, \ldots, n$. We define a *(rooted) k-tree (over Φ)* by induction on k: A 0-tree consists of a single node, labeled by a truth assignment to the primitive propositions in Φ. A $(k+1)$-tree consists of a root node labeled by a truth assignment, and for each agent i, a (possibly empty) set of directed edges labeled by i leading to roots of distinct k-trees.[4] We say a node w' is the *i-successor* of a node w in a tree if there is an edge labeled i leading from w to w'. The *depth* of a node in a tree is the distance of the node from the root. We say that the $k+1$-tree T_{k+1} is an *extension* of the k-tree T_k if T_{k+1} is the result of adding some successors to the depth-k leaves in T_k. Finally, an ω-tree T_ω is a sequence $\langle T_0, T_1, \ldots \rangle$, where T_k is a k-tree, and T_{k+1} is an extension of T_k, for $k = 0, 1, 2, \ldots$. (We remark that ω-trees are closely related to the knowledge structures of [Fagin, Halpern, and Vardi 1991; Fagin and Vardi 1986], although we do not pursue this connection here.)

We now show that with each situation we can associate a unique ω-tree. We start by going in the other direction. We can associate with each k-tree T $(k \neq \omega)$ a Kripke structure $M(T)$ defined as follows: the nodes of T are the possible worlds in $M(T)$, the $\mathcal{K}_i$ accessibility relation of $M(T)$ consists of all pairs (w, w') such that w' is an i-successor of w in T, and the truth of a primitive proposition at a world w in $M(T)$ is determined by the truth assignment labeling w.

We define the *depth* of a formula by induction on structure. Intuitively, the depth measures the depth of nesting of the K_i operators. Thus, we have $depth(p) = 0$ for a primitive proposition p; $depth(\neg\varphi) = depth(\varphi)$; $depth(\varphi \wedge \psi) = \max(depth(\varphi), depth(\psi))$; $depth(K_i\varphi) = 1 + depth(\varphi)$. If M and M' are (arbitrary) structures, w is a world in M, and w' a world in M', we say that (M, w) and (M', w') are *equivalent up to depth k*, and write $(M, w) \equiv_k (M', w')$ if, for all formulas φ with $depth(\varphi) \leq k$, we have $(M, w) \models \varphi$ iff $(M', w') \models \varphi$. For convenience, if w_0 is the root of T, we take $M(T) \models \varphi$ to be an abbreviation for $(M(T), w_0) \models \varphi$, and write $(M, w) \equiv_k M(T)$ rather than $(M, w) \equiv_k (M(T), w_0)$.

Proposition 2.1: *Fix a situation (M, w). For all k, there is a unique k-tree $T_{M,w,k}$ such that such that $(M, w) \equiv_k M(T_{M,w,k})$. Moreover, $T_{M,w,k+1}$ is an extension of $T_{M,w,k}$.*

Let $T_{M,w}$ be the ω-tree $\langle T_{M,w,0}, T_{M,w,1}, T_{M,w,2}, \ldots \rangle$. By Proposition 2.1, $T_{M,w}$ can be viewed as providing a canonical way of representing the situation (M, w) in terms of trees. We use (ω-)trees as a tool for defining what agent i considers possible in (M, w). Thus, we define i's possibilities at (M, w), denoted $Poss_i(M, w)$, to be $\{T_{M,w'} : (w, w') \in \mathcal{K}_i\}$.

[3] The subscript n in all these logics is meant to emphasize the fact that we are considering the n-agent version of the logic. We omit it when considering the single-agent case. Details and axiomatizations can be found in [Halpern and Moses 1992].

[4] Since we are allowing a node to have no successors, any k-tree is also a $(k+1)$-tree.

Intuitively, for α to be i-honest, there should be a situation (M, w) for which i has the maximum number of possibilities. Formally, if $\mathcal{S}$ is a non-introspective logic, we say that α is *$\mathcal{S}$-i-honest* if there is an $\mathcal{S}$-situation (M, w), called an *$\mathcal{S}$-i-maximum situation for α*, such that $(M, w) \models L_i\alpha$, and for all $\mathcal{S}$-situations (M', w'), if $(M', w') \models L_i\varphi$, then $Poss_i(M', w') \subseteq Poss_i(M, w)$. If α is $\mathcal{S}$-i-honest, we say that agent i knows β if all he knows is α, and write $\alpha \mid\!\sim^i_{\mathcal{S}} \beta$, if $(M, w) \models L_i\beta$ for some $\mathcal{S}$-i-maximum situation for α.[5]

How reasonable are our notions of honesty and $\mid\!\sim^i_{\mathcal{S}}$? The following results give us some justification for these definitions. The first gives us a natural characterization of honesty.

Theorem 2.2: *If $\mathcal{S}$ is a non-introspective logic, then the formula α is $\mathcal{S}$-i-honest iff $L_i\alpha$ is $\mathcal{S}$-consistent, and for all formulas $\varphi_1, \ldots, \varphi_k$, if $\models_{\mathcal{S}} L_i\alpha \Rightarrow (L_i\varphi_1 \vee \ldots \vee L_i\varphi_k)$, then $\models_{\mathcal{S}} L_i\alpha \Rightarrow L_i\varphi_j$, for some $j \in \{1, \ldots, k\}$.*

Thus, a typical dishonest formula in the case of T_n or $\mathrm{S4}_n$ is $L_ip \vee L_iq$, where p and q are primitive propositions. If α is $L_ip \vee L_iq$, then $L_i\alpha \Rightarrow (L_ip \vee L_iq)$ is valid in T_n and $\mathrm{S4}_n$, although neither $L_i\alpha \Rightarrow L_ip$ nor $L_i\alpha \Rightarrow L_iq$ is valid. However, the validity of $L_i\alpha \Rightarrow (L_ip \vee L_iq)$ depends on the fact that $L_i\alpha \Rightarrow \alpha$. This is not an axiom of K_n. In fact, it can be shown that $L_ip \vee L_iq$ is K_n-i-honest. Thus, what is almost the archetypical "dishonest" formula is honest in the context of K_n. As the following result shows, this is not an accident.

Theorem 2.3: *All formulas are K_n-i-honest.*

A set S of formulas is an *$\mathcal{S}$-i-stable set* if there is some $\mathcal{S}$-situation (M, w) such that $S = \{\varphi : (M, w) \models K_i\varphi\}$. We say the situation (M, w) *corresponds* to the stable set S. This definition is a generalization of the one given by Moore [1985] (which in turn is based on Stalnaker's definition [1980]); Moore's notion of stable set corresponds to a K45-stable set in the single-agent case. (See [Halpern 1993] for some discussion as to why this notion of stable set is appropriate.) Since a stable set describes what can be known in a given situation, we would expect a formula to be honest if it is in a minimum stable set. This is indeed true.

Theorem 2.4: *If $\mathcal{S}$ is a non-introspective logic, then α is $\mathcal{S}$-i-honest iff there is an $\mathcal{S}$-i-stable set S^α containing α which is a subset of every $\mathcal{S}$-i-stable set containing α. Moreover, if α is stable, then $\alpha \mid\!\sim^i_{\mathcal{S}} \beta$ iff $\beta \in S^\alpha$.*

This characterization of honesty is closely related to one given in [Halpern and Moses 1984]; we discuss this in more detail below.

[5] There may be more than one $\mathcal{S}$-i-maximum situation for α; two $\mathcal{S}$-i-maximum situations for α may differ in what $j \neq i$ considers possible. However, if (M, w) and (M', w') are two $\mathcal{S}$-i-maximum situations for α, then $(M, w) \models L_i\beta$ iff $(M', w') \models L_i\beta$.

Our next result gives another characterizion of what agent i knows if "all agent i knows is α", for an honest formula α. Basically, it shows that all agent i knows are the logical consequences of his knowledge of α. Thus, "all agent i knows" is a monotonic notion for the non-introspective logics.

Theorem 2.5: *If $\mathcal{S}$ is a non-introspective logic and α is $\mathcal{S}$-i-honest, then $\alpha \mid\!\sim^i_{\mathcal{S}} \beta$ iff $\models_{\mathcal{S}} L_i\alpha \Rightarrow L_i\beta$.*

This completes our discussion of the non-introspective logics. We must take a slightly different approach in dealing with the introspective logics. To see the difficulties if we attempt to apply our earlier approach without change to the introspective case, consider the single-agent case. Suppose Φ consists of two primitive propositions, say p and q, and suppose that all the agent knows is p. Surely p should be honest. Indeed, according to the framework of Halpern and Moses [1984], there is a maximum situation where p is true where the structure consists of two truth assignments: one where both p and q are true, and the other where p is true and q is false. Call this structure M. There is, of course, another structure where the agent knows p. This is the structure where the only truth assignment makes both p and q true. Call this structure M'. Let w be the world where both p and q are true. We can easily construct $T_{M,w}$ and $T_{M',w}$; the trouble is that $Poss_1(M, w)$ and $Poss_1(M', w)$ are incomparable. What makes them incomparable is introspective knowledge: In (M, w), the agent does not know q; so, because of introspection, he knows that he does not know q. On the other hand, in (M', w), the agent does not know this. These facts are reflected in the trees. We need to factor out the introspection somehow. In the single-agent case considered, this was done by considering only truth assignments, not trees. We need an analogue for the multi-agent case.

We define an *i-objective k-tree* to be a k-tree whose root has no i-successors. We define a *i-objective ω-tree* to be an ω-tree all of whose components are i-objective. Given a k-tree T, let T^i be the result of removing all the i-successors of the root of T (and all the nodes below it). Given an ω-tree $T = \langle T_0, T_1, \ldots \rangle$, let $T^i = \langle T^i_0, T^i_1, \ldots \rangle$. The way we factor out introspection is by considering i-objective trees. Intuitively, this is because the i-objective tree corresonding to a situation (M, w) eliminates all the worlds that i considers possible in that situation. Notice that in the case of one agent, the i-objective trees are precisely the possible worlds.

We define $IntPoss_i(M, w) = \{T^i : T \in Poss_i(M, w)\}$. (*IntPoss* stands for *int*rospective *poss*ibilities.) The following result assures us that we have not lost anything in the introspective logics by considering $IntPoss_i$ instead of $Poss_i$.

Lemma 2.6: *If M is an $\mathcal{S}$-structure, and $\mathcal{S}$ is an introspective logic, then $Poss_i(M, w)$ is uniquely determined*

by $\text{IntPoss}_i(M, w)$.

In the case of the introspective logics, we now repeat all our earlier definitions using *IntPoss* instead of *Poss*. Thus, for example, we say that that α is $\mathcal{S}$-*i*-honest if there is an $\mathcal{S}$-situation (M, w) such that $(M, w) \models L_i\alpha$, and for all $\mathcal{S}$-situations (M', w'), if $(M', w') \models L_i\varphi$, then $IntPoss_i(M', w') \subseteq IntPoss_i(M, w)$. We make the analogous change in the definition of $\mid\!\sim^i_{\mathcal{S}}$. Since *i*-objective trees are truth assignments in the single-agent case, it is easy to see that these definitions generalize those for the single-agent case given in [Halpern and Moses 1984].

We now want to characterize honesty and "all agent i knows" for the introspective logics. There are some significant differences from the non-introspective case. For example, as expected, the primitive proposition p is $\mathcal{S}$-1-honest even if $\mathcal{S}$ is introspective. However, due to negative introspection, $\neg L_1 q \Rightarrow L_1 \neg L_1 q$ is $\mathcal{S}$-valid, so we have $\models_{\mathcal{S}} L_1 p \Rightarrow (L_1 q \vee L_1 \neg L_1 q)$. Moreover, we have neither $\models_{\mathcal{S}} L_1 p \Rightarrow L_1 q$ nor $\models_{\mathcal{S}} L_1 p \Rightarrow L_1 \neg L_1 q$. Thus, the analogue to Theorem 2.2 does not hold.

We say a formula is *i-objective* if it is a Boolean combination of primitive propositions and formulas of the form $L_j\varphi$, $j \neq i$, where φ is arbitrary. Thus, $q \wedge L_2 L_1 p$ is 1-objective, but $L_1 p$ and $q \wedge L_1 p$ are not. Notice that if there is only one agent, say agent 1, then the 1-objective formulas are just the propositional formulas. As the following result shows, the analogue of Theorem 2.2 holds for KD45_n and K45_n, provided we stick to *i*-objective formulas.

Theorem 2.7: *For $\mathcal{S} \in \{\text{KD45}_n, \text{K45}_n\}$, the formula α is $\mathcal{S}$-i-honest iff for all i-objective formulas $\varphi_1, \ldots, \varphi_k$, if $\models_{\mathcal{S}} L_i\alpha \Rightarrow (L_i\varphi_1 \vee \ldots \vee L_i\varphi_k)$ then $\models_{\mathcal{S}} L_i\alpha \Rightarrow L_i\varphi_j$, for some $j \in \{1, \ldots, k\}$.*

This result does not hold for S5_n; for example, $\models_{\text{S5}_n} L_1 p \Rightarrow (L_1 q \vee L_1 L_2 \neg L_2 L_1 q)$ (this follows from the fact that $\models_{\text{S5}_n} \neg L_1 q \Rightarrow L_1 L_2 \neg L_2 L_1 q$). However, it is easy to see that $\not\models_{\text{S5}_n} L_1 p \Rightarrow L_1 q$ and $\not\models_{\text{S5}_n} L_1 p \Rightarrow L_1 L_2 \neg L_2 L_1 q$. Since p is S5_n-1-honest, Theorem 2.7 fails for S5_n.

Theorem 2.7 is a direct extension of a result in [Halpern and Moses 1984] for the single-agent case. Two other characterizations of honesty and "all I know" are given by Halpern and Moses, that can be viewed as analogues to Theorems 2.4 and 2.5. As we now show, they also extend to K45_n and KD45_n, but not S5_n.

One of these characterizations is in terms of stable sets. The direct analogue of Theorem 2.4 does not hold for the introspective logics. In fact, as was already shown in [Halpern and Moses 1984] for the single-agent case, any two consistent stable sets are incomparable with respect to set inclusion. Again, the problem is due to introspection. For suppose we have two consistent $\mathcal{S}$-*i*-stable sets S and S' such that $S \subset S'$, and $\varphi \in S' - S$. By definition, there must be situations (M, w) and (M', w'), corresponding to S and S' respectively, for which we have $(M, w) \models L_i\varphi$ and $(M', w') \not\models L_i\varphi$. By introspection, we have $(M, w) \models L_i L_i \varphi$ and $(M', w') \models L_i \neg L_i \varphi$. This means that $L_i\varphi \in S$ and $\neg L_i\varphi \in S'$. Since $S \subset S'$, we must also have $L_i\varphi \in S$, which contradicts the assumption that S' is consistent.

We can get an analogue of Theorem 2.4 if we consider *i*-objective formulas. Define the *i-kernel* of an $\mathcal{S}$-*i*-stable set S, denoted $\ker_i(S)$, to consist of all the *i*-objective formulas in S.

Theorem 2.8: *For $\mathcal{S} \in \{\text{KD45}_n, \text{K45}_n\}$, a formula α is $\mathcal{S}$-i-honest iff there is an $\mathcal{S}$-i-stable set S^i_α containing α such that for all i-stable sets S containing α, we have $\ker_i(S^i_\alpha) \subseteq \ker_i(S)$. Moreover, α is $\mathcal{S}$-i-honest, then $\alpha \mid\!\sim^i_{\mathcal{S}} \beta$ iff $\beta \in S^i_\alpha$.*

As we show in the full paper, Theorem 2.8 does not hold for S5_n. This is not an artifact of our definition of honesty for S5_n, since in fact we can show that for *no* formula α is there an S5_n-*i*-stable set containing α whose *i*-kernel is a minimum.

Finally, let us consider the analogue to Theorem 2.5. In contrast to the non-introspective case, inference from "all agent i knows" is nonmonotonic for the introspective logics. For example, we have $p \mid\!\sim_{\mathcal{S}} \neg L_1 q$, even though $\not\models_{\mathcal{S}} L_1 p \Rightarrow L_1 \neg L_1 q$. This seems reasonable: if all agent 1 knows is p, then agent 1 does not know q and (by introspection) knows that he does not know this. As shown in [Halpern and Moses 1984], there is an elegant algorithmic characterization of "all agent i knows" in the single-agent case. We extend it to the multi-agent case here. We recursively define a set $D^i_{\mathcal{S}}(\alpha)$ that intuitively consists of all the formulas agent i knows, given that agent i knows only α (and reasons using modal logic $\mathcal{S}$):

$$\varphi \in \mathcal{D}^i_{\mathcal{S}}(\alpha) \text{ iff } \models_{\mathcal{S}} (L_i\alpha \wedge \varphi^{\alpha,i}) \Rightarrow L_i\varphi,$$

where $\varphi^{\alpha,i}$ is the conjunction of $L_i\psi$ for all subformulas $L_i\psi$ of φ for which $\psi \in D^i_{\mathcal{S}}(\alpha)$, and $\neg L_i\psi$ for all subformulas $L_i\psi$ for which $\psi \notin D^i_{\mathcal{S}}(\alpha)$ (where φ is considered a subformula of itself). Thus, the algorithm says that the agent knows φ if it follows from knowing α, together with the formulas that were decided by recursive applications of the algorithm. Then we have:

Theorem 2.9: *For $\mathcal{S} \in \{\text{KD45}_n, \text{K45}_n\}$, the formula α is i-honest iff $D^i_{\mathcal{S}}(\alpha)$ is (propositionally) consistent. If α is $\mathcal{S}$-i-honest, then $\alpha \mid\!\sim^i_{\mathcal{S}} \beta$ iff $\beta \in D^i_{\mathcal{S}}(\alpha)$.*

While the analogue to Theorem 2.9 does not hold for S5_n, the algorithm is correct for honest formulas.

Theorem 2.10: *If α is S5_n-i-honest, then $\alpha \mid\!\sim^i_{\text{S5}_n} \beta$ iff $\beta \in D^i_{\text{S5}_n}(\alpha)$.*

We now characterize the complexity of computing honesty and "all i knows".

Theorem 2.11: *For $\mathcal{S} \in \{\text{T}_n, \text{S4}_n : n \geq 1\} \cup \{\text{KD45}_n, \text{K45}_n, \text{S5}_n : n \geq 2\}$, the problem of computing whether α is $\mathcal{S}$-i-honest is PSPACE-complete.*

Of course, the problem of computing whether α is K_n-i-honest is trivial: the answer is always "Yes".

Theorem 2.12: *For $\mathcal{S} \in \{K_n, T_n, S4_n : n \geq 1\} \cup \{KD45_n, K45_n, S5_n : n \geq 2\}$, if α is $\mathcal{S}$-i-honest, then the problem of deciding if $\alpha \mid\!\sim^i_{\mathcal{S}} \beta$ is PSPACE-complete.*

We close this section by briefly comparing our approach to others in the literature. Fagin, Halpern, and Vardi [1991] define a notion of *i-no-information extension* that can also be viewed as characterizing a notion of "all agent i knows" in the context of $S5_n$. However, it is defined only for a limited set of formulas. It can be shown that these formulas are always $S5_n$-i-honest in our sense, and, if α is one of these formulas, we have $\alpha \mid\!\sim^i_{S5_n} \beta$ iff β is true in the i-no-information extension of α. The fact that these two independently motivated definitions coincide (at least, in the cases where the i-no-information extension is defined) provides further evidence for the reasonableness of our definitions.

Vardi [1985] defines a notion of "all agent i knows" for $S4_n$, using the knowledge-structures approach of [Fagin, Halpern, and Vardi 1991], and proves Theorem 2.5 for $S4_n$ in the context of his definition. It is not hard to show that our definition of honesty coincides with his for $S4_n$. However, the knowledge structures approach does not seem to extend easily to the introspective logics. Moreover, using our approach leads to much better complexity results. For example, all that Vardi was able to show was that honesty was (nonelementary-time) decidable.

Parikh [1991] defines a notion of "all that is known" for $S5_n$ much in the spirit of the definitions given here. Among other things, he also starts with k-trees (he calls them *normal models*), although he does not use i-objective trees. However, rather than focusing on all that some fixed agent i knows as we have done, Parikh treats all agents on an equal footing. This leads to some technical differences between the approaches. He was also able to obtain only nonelementary-time algorithms for deciding whether a formula was honest in his sense.

3 Levesque's notion of "only knowing"

Despite the similarity in philosophy and terminology, Levesque's notion of "only knowing" differs in some significant ways from the HM notion (see [Halpern 1993] for a discussion of this issue). Nevertheless, some of the ideas of the previous section can be applied to extending it to many agents.

Levesque considers a K45 notion of belief, and introduces a modal operator O, where $O\alpha$ is read "only believes α". The O operator is best understood in terms of another operator introduced by Levesque denoted N. While $L\alpha$ says "α is true at all the worlds that the agent considers possible", $N\alpha$ is viewed as saying "α is true at all the worlds that the agent does *not* consider possible". Then $O\alpha$ is defined as an abbreviation for $L\alpha \wedge N\neg\alpha$. Thus, $O\alpha$ holds if α is true at all the worlds that the agent considers possible, and only these worlds. We can read $L\alpha$ as saying "the agent knows at least α", while $N\neg\alpha$ says "the agent knows at most α" (for if he knew more, than he would not consider possible all the worlds where α is true).

In the case of a single agent, since worlds are associated with truth assignments, it is easy to make precise what it means that the agent does not consider a world possible: it is impossible if it is not one of the truth assignments the agents considers possible. Thus, Levesque defines:

$$(W, w) \models N\alpha \text{ if } (W, w') \models \alpha \text{ for all } w' \notin W.$$

Two important features of this definition are worth mentioning here. First, the set of all worlds is absolute, and does not depend on the situation: it is the set of all truth assignments. Thus, the set of impossible worlds given that W is the set of worlds that the agent considers possible is just the complement of W (relative to the set of all truth assignments). Second, when evaluating the truth of α at an "impossible world" w', we do not change W, the set of worlds that the agent considers possible. (We remark that it is this second point that results in the main differences between this notion of "all I know" and the HM notion; see [Halpern 1993].)

Of course, the problem in extending Levesque's notion to many agent lies in coming up with an analogue to "the worlds that the agent does not consider possible". This is where our earlier ideas come into play.

Before we go into details on the multi-agent case, we mention one important property of this notion of "only knowing". Moore [1985] defines a *stable expansion of* α to be a (K45-)stable set S such that S is the closure under propositional reasoning of $\{\alpha\} \cup \{L\alpha : L\alpha \in S\} \cup \{\neg L\alpha : \neg L\alpha \in S\}$. Notice that for any stable set S, there is a unique set W_S of truth assignments such that $\varphi \in S$ iff $(W_S, w) \models L\varphi$ for all $w \in W_S$. Levesque shows that S is a stable expansion of α iff $(W_S, w) \models O\alpha$ for all $w \in W_S$.

We now turn to extending Levesque's definitions to the multi-agent case. We first extend the language of knowledge by adding modal operators N_i and O_i for each agent $i = 1, \ldots, n$. Following Lakemeyer, we call the full language $\mathcal{ONL}_n$. We say that a formula in $\mathcal{ONL}_n$ is *basic* if it does not involve the modal operators O_i or N_i. Finally, we take the language $\mathcal{ONL}_n^-$ to be the sublanguage of $\mathcal{ONL}_n$ where no O_j or N_j occurs in the scope of an O_i, N_i, or L_i, for $i \neq j$. In analogy to Levesque, we define $O_i\alpha$ as the conjunction $L_i\alpha \wedge N_i\neg\alpha$. The problem is to define $N_i\alpha$. As in the single-agent case, we want $N_i\alpha$ to mean that α is true at all the worlds that i does not consider possible. So what are the worlds that i does not consider possible?

Perhaps the most straightforward way of making sense of this, used by Lakemeyer [1993], is to define N_i in terms of the complement of the $\mathcal{K}_i$ relation. We

briefly outline this approach here. Given a structure $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \pi)$, let $\mathcal{K}_i(w) = \{w' : (w, w') \in \mathcal{K}_i\}$. $\mathcal{K}_i(w)$ is the set of worlds that agent i considers possible at w. We write $w \approx_i w'$ if $\mathcal{K}_i(w) = \mathcal{K}_i(w')$. Thus, if $w \approx_i w'$, then agent i's possibilities are the same at w and w'. Finally, Lakemeyer defines:

$(M, w) \models_{Lak} N_i\alpha$ if $(M, w') \models \alpha$ for all w' such that $(w, w') \notin \mathcal{K}_i$ and $w \approx_i w'$.

By restricting attention to worlds w' such that $w \approx_i w'$, Lakemeyer is preserving the second property of Levesque's definition, namely, that when evaluating the truth of a formula at an impossible world, we keep the set of agent i's possibilities unchanged. However, this definition does not capture the first property of Levesque's definition, that the set of impossible worlds is absolute. Here it is relative to the structure. To get around this problem, Lakemeyer focuses on a certain *canonical model*, which intuitively has "all" the possibilities.[6] It is only in this model that the N_i (and thus the O_i) operators seem to have the desired behavior. (We discuss to what extent they really do have the desired behavior in this canonical model below.)

We want to define N_i and O_i in a reasonable way in all models. We proceed as follows:

$(M, w) \models N_i\alpha$ if $(M', w') \models \alpha$ for all (M', w') such that $T^i_{M',w'} \notin IntPoss_i(M, w)$ and $IntPoss_i(M, w) = IntPoss_i(M', w')$.

The analogues to Lakemeyer's definitions should be obvious: we replace $(w, w') \notin \mathcal{K}_i$ by $T^i_{M',w'} \notin IntPoss_i(M, w)$ and $w \approx_i w'$ by $IntPoss_i(M, w) = IntPoss_i(M', w')$.

What evidence do we have that this definition is reasonable? One piece of evidence is that we can extend to the multi-agent case Levesque's result regarding the relationship between only knowing and stable expansions. To do this, we first need to define the notion of stable expansion in the context of many agents. We say that S is a K45$_n$-*i-stable expansion of* α if S is a K45$_n$-i-stable set and S is the closure under K45$_n$ of $\{\alpha\} \cup \{L_i\alpha : L_i\alpha \in S\} \cup \{\neg L_i\alpha : \neg L_i\alpha \in T\}$.[7]

Next, we need to associate a situation with each K45$_n$-i-stable set, as we were able to do in the single-agent case. Given a set S of basic formulas, we say that the K45$_n$-situation (M, w) *i-models* S if, for all basic formulas φ, we have $(M, w) \models L_i\varphi$ iff $\varphi \in S$. In analogy to the single-agent case, the situation that we would like to associate with a stable set S is one that i-models S. There is, however, a complication. In the single-agent case, a stable set determines the set of possible truth assignments. That is, given a stable set S, there is a unique set W_S such that (for any w) we have $(W_S, w) \models L\varphi$ iff $\varphi \in S$. The analogue does not hold in the multi-agent case. That is, given a stable set S, there is not a unique set $\mathcal{W}$ of i-objective ω-trees such that if (M, w) i-models S, then $IntPoss_i(M, w) = \mathcal{W}$. As we show in the full paper, two structures can agree on all basic formulas, and still differ with regard to formulas of the form $N_i\alpha$ or $O_i\alpha$ under $\models$.[8] A similar phenomenon was encountered by Levesque [1990] when considering only knowing in the first-order case. We solve our problem essentially the same way he solved his. We say that (M, w) is a *maximum i-model of the stable set* S if (M, w) is an i-model of S and for every i-model (M', w') of S, we have $IntPoss_i(M', w') \subseteq IntPoss_i(M, w)$.

Lemma 3.1: *Every* K45$_n$-*i-stable set has a maximum i-model.*

Theorem 3.2: *Suppose S is a* K45$_n$-*i-stable set and (M, w) is a maximum i-model of S. Then S is an i-stable expansion of α iff $(M, w) \models O_i\alpha$.*

We remark that an analogous result is proved by Lakemeyer [1993], except that he restricts attention to situations in the canonical model.

More evidence as to the reasonableness of our definitions is given by considering the properties of the operators N_i and O_i. As usual, we say that φ is valid, and write $\models \varphi$, if $(M, w) \models \varphi$ for all situations (M, w). We write $\models_{Lak} \varphi$ if φ is valid under Lakemeyer's semantics *in the canonical model*; we remark that $\models_{Lak}$ is the notion of validity considered by Lakemeyer, since he is only interested in the canonical model.

Theorem 3.3: *For all formulas φ, if $\models \varphi$ then $\models_{Lak} \varphi$. If $\varphi \in \mathcal{ONL}_n^-$, we have $\models \varphi$ iff $\models_{Lak} \varphi$.*

This theorem says that Lakemeyer's notion of validity is stronger than ours, although the two notions agree with respect to formulas in the sublanguage $\mathcal{ONL}_n^-$. In fact, Lakemeyer's notion of validity is strictly stronger than ours. Lakemeyer shows that $\models_{Lak} \neg O_i \neg O_j p$; under his semantics, it is impossible for i to only know that it is not the case that j only knows p. This seems counterintuitive. Why should this be an unattainable state of knowledge? Why can't j just tell i that it is not the case that he (j) only knows p?

We would argue that the validity of this formula is an artifact of Lakemeyer's focus on the canonical model. Roughly speaking, we would argue that the canonical model is not "canonical" enough. Although it includes all the possibilities in terms of basic formulas, it does

[6] This canonical model is built using standard modal logic techniques (cf. [Halpern and Moses 1992; Hughes and Cresswell 1968]); the worlds in this canonical model consist of all maximally K45$_n$-consistent subsets of formulas.

[7] In Moore's definition of stable expansion, we could have used closure under K45 instead of closure under deductive reasoning. The two definitions are equivalent in the single-agent case, but modal reasoning is necessary in the multi-agent case so that agent i can capture j's introspective reasoning.

[8] This can be viewed as indicating that basic formulas are not expressive enough to describe ω-trees. If we had had allowed infinite disjunctions and conjunctions into the language, then a stable set would determine the set of trees.

not include all the possibilities in terms of the extended language. The formula $O_i \neg O_j p$ is easily seen to be satisfiable under our semantics.

Lakemeyer provides a collection of axioms that he proves are sound with respect to $\models_{Lak}$, and complete for formulas in $\mathcal{ONL}_n^-$. He conjectures that they are not complete with respect to the full language. It is not hard to show that all of Lakemeyer's axioms are sound with respect to our semantics as well. It follows from Theorem 3.3 and Lakemeyer's completeness result that these axioms are complete with respect to $\mathcal{ONL}_n^-$ for our semantics too. It also follows from these observations that, as Lakemeyer conjectured, his proof system is not complete. This follows since everything provable in his system must be valid under our semantics, and $\neg O_i \neg O_j p$ is not valid under our semantics (although it is valid under his).

4 Discussion

We have shown how to extend two notions of only knowing to many agents. The key tool in both of these extensions was an appropriate canonical representation of the possibilities of the agents. Although we gave arguments showing that the way we chose to represent an agent's possibilities was reasonable, it would be nice to have a more compelling theory of "appropriateness". For example, why is it appropriate to use arbitrary trees for the non-introspective logics, and i-objective trees for the introspective logics? Would a different representation be appropriate if we had changed the underlying language? Perhaps a deeper study of the connections between ω-trees and the knowledge structures of [Fagin and Vardi 1986; Fagin, Halpern, and Vardi 1991] may help clarify some of these issues.

Another open problem is that of finding a complete axiomatization for $\mathcal{ONL}_n$. We observed that Lakemeyer's axioms were not complete with respect to his semantics. In fact, it seems that these axioms are essentially complete for $\mathcal{ONL}_n$ under our semantics.[9] We hope to report on these results in the future.

Acknowledgements: I would like to thank Ron Fagin, Gerhard Lakemeyer, Grisha Schwarz, and Moshe Vardi for their helpful comments on an earlier draft of this paper.

[9] The reason we say "essentially complete" here is that one of the axioms has the form

$N_i\alpha \Rightarrow \neg L_i\alpha$ for all basic i-objective α falsifiable in $\mathrm{K45}_n$.

We need to extend this axiom to formulas that are not basic. But the axiom system $\mathrm{K45}_n$ does not apply to non-basic formulas. We deal with this problem by extending the language so that we can talk about satisfiability within the language. The axiom then becomes

$$\neg Con(\alpha) \Rightarrow (N_i\alpha \Rightarrow \neg L_i\alpha),$$

where $Con(\alpha)$ holds if α is satisfiable.

References

Fagin, R., J. Y. Halpern, and M. Y. Vardi (1991). A model-theoretic analysis of knowledge. *Journal of the ACM 91*(2), 382–428. A preliminary version appeared in *Proc. 25th IEEE Symposium on Foundations of Computer Science*, 1984.

Fagin, R. and M. Y. Vardi (1986). Knowledge and implicit knowledge in a distributed environment: preliminary report. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, San Mateo, CA, pp. 187–206. Morgan Kaufmann.

Halpern, J. Y. (1993). A critical reexamination of default logic, autoepistemic logic, and only knowing. In *Proceedings, 3rd Kurt Gödel Colloquium*. Springer-Verlag.

Halpern, J. Y. and Y. Moses (1984). Towards a theory of knowledge and ignorance. In *Proc. AAAI Workshop on Non-monotonic Logic*, pp. 125–143. Reprinted in *Logics and Models of Concurrent Systems*, (ed., K. Apt), Springer-Verlag, Berlin/New York, pp. 459–476, 1985.

Halpern, J. Y. and Y. Moses (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence 54*, 319–379.

Hughes, G. E. and M. J. Cresswell (1968). *An Introduction to Modal Logic*. London: Methuen.

Lakemeyer, G. (1993). All they know: a study in multi-agent autoepestemic reasoning. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*. Unpublished manuscript.

Levesque, H. J. (1990). All I know: A study in autoepistemic logic. *Artificial Intelligence 42*(3), 263–309.

Moore, R. C. (1985). Semantical considerations on nonmonotonic logic. *Artificial Intelligence 25*, 75–94.

Morgenstern, L. (1990). A theory of multiple agent nonmonotonic reasoning. In *Proc. National Conference on Artificial Intelligence (AAAI '90)*, pp. 538–544.

Parikh, R. (1991). Monotonic and nonmonotonic logics of knowledge. *Fundamenta Informaticae 15*(3,4), 255–274.

Stalnaker, R. (1980). A note on nonmonotonic modal logic. Technical report, Dept. of Philosophy, Cornell University. A slightly revised version will appear in *Artificial Intelligence*.

Vardi, M. Y. (1985). A model-theoretic analysis of monotonic knowledge. In *Proc. Ninth International Joint Conference on Artificial Intelligence (IJCAI '85)*, pp. 509–512.