Multi-Agent Only Knowing

Joseph Y. Halpern
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099, USA
halpern@almaden.ibm.com

Gerhard Lakemeyer
Institute of Computer Science
University of Bonn
Römerstr. 164
D-53117 Bonn, Germany
gerhard@cs.uni-bonn.de

Abstract

Levesque introduced a notion of "only knowing", with the goal of capturing certain types of nonmonotonic reasoning. Levesque's logic dealt with only the case of a single agent. Recently, both Halpern and Lakemeyer independently attempted to extend Levesque's logic to the multi-agent case. Although there are a number of similarities in their approaches, there are some significant differences. In this paper, we reexamine the notion of only knowing, going back to first principles. In the process, we point out some problems with the earlier definitions. This leads us to reconsider what the properties of only knowing ought to be. We provide an axiom system that captures our desiderata, and show that it has a semantics that corresponds to it. The axiom system has an added feature of interest: it includes a modal operator for satisfiability, and thus provides a complete axiomatization for satisfiability in the logic K45.

1 Introduction

Levesque (1990) introduced a notion of "only knowing", with the goal of capturing certain types of nonmonotonic reasoning. In particular, the hope was to capture the type of reasoning that says "If all I know is that Tweety is a bird, and that birds typically fly, then I can conclude that Tweety flies".[1] Levesque's logic dealt with only the case of a single agent. It is clear that in many applications of such nonmonotonic reasoning, there are several agents in the picture. For example, it may be the case that all Jack knows about Jill is that Jill knows that Tweety is a bird and that birds typically fly. Jack may then want to conclude that Jill knows that Tweety flies.

[1]The reader should feel free to substitute "believe" anywhere we say "know". Indeed, the formal logic that we use, which is based on the modal logic K45, is more typically viewed as a logic of belief rather than knowledge.

Recently, both Halpern (1993) and Lakemeyer (1993) independently attempted to extend Levesque's logic to the multi-agent case. Although there are a number of similarities in their approaches, there are some significant differences. In this paper, we reexamine the notion of only knowing, going back to first principles. In the process, we point out some problems with both of the earlier definitions. This leads us to consider what the properties of only knowing ought to be. We provide an axiom system that captures all our desiderata, and show that it has a semantics that corresponds to it. The axiom system has an added feature of interest: it involves enriching the language with a modal operator for satisfiability, and thus provides an axiomatization for satisfiability in K45. Unfortunately, the semantics corresponding to this axiomatization is not as natural as we might like. It remains an open question whether there is a natural semantics for only knowing that corresponds to this axiomatization.

The rest of this paper is organized as follows. In the next section, we review the basic ideas of Levesque's logic and point out an alternative semantics followed by some remarks regarding the use of a finite language instead of an infinite one as in Levesque's case. In Section 3 we review Lakemeyer's approach, which we call the *canonical-model* approach, and discuss some of its strengths and weaknesses. In Section 4, we go through the same process for Halpern's approach. Finally, in Section 5, we consider our new approach. We conclude in Section 6 with some discussion of only knowing.

2 Levesque's Logic of Only Knowing

We begin by reconsidering Levesque's definition. Let $\mathcal{ONL}$ be a propositional modal language with a countably infinite set of primitive propositions (as we shall see, the fact that the number of primitive propositions is countably infinite, rather than finite, has a significant impact), the classical operators $\neg$ and $\vee$ and two modalities, L and N. We freely use other connectives like $\wedge$, $\Rightarrow$, and $\Leftrightarrow$ as syntactic abbreviations of the usual kind. In addition, we take $O\alpha$ to be an abbreviation for $L\alpha \wedge N\neg\alpha$. Here $L\alpha$ should be read as "the agent knows or believes (at least) α", $N\alpha$ should be read as "the agent believes at most $\neg\alpha$" (so that $N\neg\alpha$ is "the agent believes at most α") and $O\alpha$ should be read as "the agent knows only α".

Levesque gave semantics to knowing and only knowing using the standard possible-worlds approach. In the single-agent case, we can identify a *situation* with a pair (W, w), where w is a possible world (represented as a truth assignment to the primitive propositions) and W consists of a set of possible worlds. Intuitively, W is the set of worlds which the agent considers (epistemically) possible, and w describes the real world. We do not require that $w \in W$ or that $W \neq \emptyset$.[2] As usual, we say that the agent knows (at least) α if α is true in all the worlds that the agent considers possible. Formally, the semantics of the modality L and the classical connectives is given as follows.

[2]By requiring that W is nonempty, we get the modal logic KD45; by requiring that $w \in W$, we get S5.

$(W, w) \models p$ if $w \models p$ if p is a primitive proposition.
$(W, w) \models \neg\alpha$ if $(W, w) \not\models \alpha$.
$(W, w) \models \alpha \vee \beta$ if $(W, w) \models \alpha$ or $(W, w) \models \beta$
$(W, w) \models L\alpha$ if $(W, w') \models \alpha$ for all $w' \in W$.

Notice that if $L\alpha$ holds, then the agent may know more than α. For example, Lp does not preclude $L(p \wedge q)$ from holding. We can think of $L\alpha$ as saying that the agent knows *at least* α.

How do we give precise semantics to N? That is, when should we say that $(W, w) \models N\beta$? Intuitively, $N\beta$ is true if β is true at all the worlds that the agent does *not* consider possible. It seems fairly clear from the intuition that we need to evaluate the truth of β in worlds $w' \notin W$, since these are the worlds that the agent considers impossible in (W, w). But if β is a complicated formula involving nested L operators, then we cannot simply evaluate the truth of β at a world w'. We need to have a set of worlds too. In fact, the set of possible worlds we use is still W. That is, while evaluating the truth of β in the impossible worlds, the agent keeps the set of worlds he considers possible fixed. Formally, we define

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

Let us stress three important features of this definition. First, as we have already observed, the set of possibilities is kept fixed when we evaluate $N\alpha$. Second, the set of *conceivable worlds*—the union of the set of "possible" worlds considered when evaluating L and the set of "impossible" worlds considered when evaluating N—is fixed, independent of the situation (W, w); it is always the set of all truth assignments. Finally, for every set of conceivable worlds, there is a model where that set is precisely the set of worlds that the agent considers possible. We will return to these properties for guidance when we discuss possible ways of extending Levesque's semantics to the multi-agent case.

Since $O\alpha$ is an abbreviation for $L\alpha \wedge N\neg\alpha$, we have that

$$(W, w) \models O\alpha \text{ if for all worlds } w',\ w' \in W \text{ iff } (W, w') \models \alpha.$$

We end this review of Levesque's logic by presenting (a slight variant of) his proof theory. We define an *objective* formula to be a propositional formula, i.e., a formula with no modal operators, and a *subjective* formula to be a Boolean combination of formulas of the form $L\varphi$ or $N\varphi$.

Axioms:

A1. All instances of axioms of propositional logic
A2. $L(\alpha \Rightarrow \beta) \Rightarrow (L\alpha \Rightarrow L\beta)$
A3. $N(\alpha \Rightarrow \beta) \Rightarrow (N\alpha \Rightarrow N\beta)$
A4. $\sigma \Rightarrow L\sigma \wedge N\sigma$ for every subjective formula σ
A5. $N\alpha \Rightarrow \neg L\alpha$ if $\neg\alpha$ is a propositionally consistent objective formula

Inference Rules:

MP. From α and $\alpha \Rightarrow \beta$ infer β

Nec. From α infer $L\alpha$ and $N\alpha$.

Axioms **A2–A4** tell us that that L and N separately are both characterized by are chacterized by the axioms of the modal logic K45 (see (Chellas 1980) for more discussion). Actually, **A4** tells us more; it says that L and N are mutually introspective, so that, for example, $L\varphi \Rightarrow NL\varphi$ is valid. Perhaps the most interesting axiom is **A5**, which gives only-knowing its desired properties. Its soundness depends on the fact that the union of the set of worlds considered when evaluating L and the set of worlds considered when evaluating N is the set of all conceivable worlds.[3]

Theorem 2.1: (Levesque 1990) *The proof theory is sound and complete with respect to the semantics.*

It is interesting to note that the assumption that L and N are interpreted with respect to complementary sets of worlds is not forced by the axioms. In particular, for its soundness, Axiom **A5** requires only that the sets considered for L and N cover all conceivable worlds, but they may overlap. The following semantics makes this precise.

Define an *extended situation* to be a triple (W_L, W_N, w), where W_L and W_N are sets of worlds (truth assignments) such that $W_L \cup W_N$ consists of all truth assignments. Define a new satisfaction relation $\models^x$ which is exactly like Levesque's except for L- and N-formulas. For them, we have

$(W_L, W_N, w) \models^x L\alpha$ if $(W_L, W_N, w') \models^x \alpha$ for all $w' \in W_L$.

$(W_L, W_N, w) \models^x N\alpha$ if $(W_L, W_N, w') \models^x \alpha$ for all $w' \in W_N$.

Note that L and N are now treated in a completely symmetric way. We can recover Levesque's semantics by restricting to extended situations where W_N is the complement of W_L. Although we allow more structures, all of Levesque's axioms are easily seen to be sound under $\models^x$. Thus, it follows from Theorem 2.1 that

Theorem 2.2: *Levesque's axiomatization is sound and complete with respect to* $\models^x$.

We can actually prove completeness of Levesque's axioms for $\models^x$ directly, using a standard canonical model construction (cf. Section 3). Moreover, as we show in the full paper, we can use our approach to give a simpler completeness proof for Levesque's semantics than the one given by Levesque. The key idea is to find a model satisfying a formula where W_L and W_N may overlap, and then use the truth assignment to a primitive proposition not mentioned in the formula to force W_L and W_N to be disjoint. We know

[3]Note that, while unusual, the axiom schema **A5** is recursive, since consistency of formulas in classical propositional logic is decidable. Hence the axioms themselves are *recursive*. As noted in (Levesque 1990), this is a problem in the first-order case, however. In fact, Levesque's proof theory for the first-order version of his logic was recently shown to be incomplete (Halpern and Lakemeyer 1995).

that there is bound to be a primitive proposition not mentioned in the formula, given the assumption that the set of primitive propositions is infinite.

Interestingly, if we move to a language with only a finite set of primitive propositions Φ, then $\models$ and $\models^x$ are no longer characterized by the same axioms. Levesque's axioms are still sound and complete for $\models^x$, and sound for $\models$, but they are not complete for $\models$. Levesque's completeness result depends crucially on the fact that there are infinitely many primitive propositions in the language. With only finitely many primitive propositions, it would appear that we need extra axioms to characterize $\models$.[4] For example, if the language has only one primitive proposition, say p, then $\neg L \neg p \Rightarrow N \neg p$ would be valid under $\models$. It turns out, however, that instead of having to add extra axioms to capture properties like these, it suffices to replace **A5** by a much stronger version which essentially tells us that formulas involving N are reducible to formulas involving only L in the finite language case.

To make this precise, first note that worlds, which are truth assignments to the primitive propositions in Φ, are themselves finite if Φ is finite. Hence we can identify a world w with the conjunction of all literals over Φ which are true at w. For example, if $\Phi = \{p, q\}$ and w makes p true and q false, then we identify w with $p \wedge \neg q$. For any objective formula α, let W_α be the set of all worlds that satisfy α. The axiom system AX_{fin} is then obtained from Levesque's system by replacing **A5** by the following axiom (where if W_α is empty, the conjunction over the empty set is taken to be vacuously *true*, as usual):

A5$_{fin}$. $N\alpha \equiv \bigwedge_{w \in W_{\neg\alpha}} \neg L \neg w$ if α is an objective formula.

The axiom is easily seen to be sound since it merely expresses that $N\alpha$ holds at W just in case W contains all worlds that satisfy $\neg\alpha$. Note that this property depends only on the fact that L and N are defined with respect to complementary sets of worlds and, hence, also holds in the case of infinite Φ. However, it is only in the finite case that we can express the axiom in our language (using only finitely many conjunctions). Completeness is also easy to establish. Levesque (1990) has already shown that it can be proved using only the axioms **A1**–**A4** that every formula is equivalent to one without nested modalities. With **A5**$_{fin}$ we then obtain an equivalent formula which does not mention N. In other words, given a formula consistent with respect to AX_{fin}, a satisfying model can be constructed with the usual technique for $K45$ alone.

Theorem 2.3: *AX_{fin} is sound and complete for Levesque's semantics if the number of primitive propositions is finite.*

[4] This was the situation in the logic considered in (Fagin, Halpern, and Vardi 1992). In that paper, a simple axiomatization was provided for the case where Φ was infinite; for each finite Φ, an extra axiom was needed (that depended on Φ).

3 The Canonical-Model Approach

How do we extend this intuition to the multi-agent case? First we extend the language $\mathcal{ONL}$ to the case of many agents. That is, we now consider a language $\mathcal{ONL}_n$, which is just like $\mathcal{ONL}$ except that there are modalities L_i and N_i for each agent i, $1 \leq i \leq n$, for some fixed n. By analogy with the single-agent case, we call a formula *i-subjective* if it is a Boolean combination of formulas of the form $L_i\varphi$ and $N_i\varphi$. What should be the analogue of an objective formula? It clearly is more than just a propositional formula. From agent 1's point of view, a formula like L_2p or even L_2L_1p is just as "objective" as a propositional formula. We define a formula to be *i-objective* if it is a Boolean combination of primitive propositions and formulas of the form $L_j\varphi$ and $N_j\varphi$, $j \neq i$, where φ is arbitrary. Thus, $q \wedge N_2L_1p$ is 1-objective, but L_1p and $q \wedge L_1p$ are not. The i-objective formulas true at a world can be thought of as characterizing what is true apart from agent i's subjective knowledge of the world.

The standard model here is to have a *Kripke structure* with worlds and accessibility relations that describe what worlds the agents consider possible in each world. Formally, a (Kripke) structure or model is a tuple $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where π associates with each world a truth assignment to the primitive propositions and $\mathcal{K}_i$ is agent i's accessibility relation. Given such a Kripke structure M, let $\mathcal{K}_i^M(w) = \{w' : (w, w') \in \mathcal{K}_i\}$.[5] $\mathcal{K}_i^M(w)$ is the set of worlds that agent i considers possible at w in structure M. As usual, we define

$$(M, w) \models L_i\alpha \text{ if } (M, w') \models \alpha \text{ for all } w' \in \mathcal{K}_i^M(w).$$

We focus on structures where the accessibility relations are Euclidean and transitive, where a relation R on W is Euclidean if $(u, v) \in R$ and $(u, w) \in R$ implies that $(v, w) \in R$, and R is transitive if $(u, v) \in R$ and $(v, w) \in R$ implies that $(u, w) \in R$. We call such structures K45$_n$-structures. It is well known (Chellas 1980; Halpern and Moses 1992) that these assumptions are precisely what is required to get belief to obey the K45 axioms (generalized to n agents). We say that a formula consistent with these axioms is K45$_n$-consistent. An infinite set of formulas is said to be K45$_n$-consistent if the conjunction of the formulas in every one of its finite subsets is K45$_n$-consistent.

Now the question is how to define the modal operator N_i. The problem in the multi-agent case is that we can no longer identify a possible world with a truth assignment. In the single-agent case, knowing the set of truth assignments that the agent considers possible completely determines his knowledge. This is no longer true in the multi-agent case. Somehow we must take the accessibility relations into account. A general semantics for an N-like operator was first given by Humberstone (1986) and later by Ben-David and Gafni (1989). Following this approach, we provisionally define

$$(M, w) \models N_i\alpha \text{ if } (M, w') \models \alpha \text{ for all } w' \in W - \mathcal{K}_i^M(w).$$

[5] We use the superscript M since we shall later need to talk about the $\mathcal{K}_i$ relations in more than one model at the same time.

The problem with this definition is that it misses out on the intuition that when evaluating $N_i\alpha$, we keep the set of worlds that agent i considers possible fixed. If $w' \in W - \mathcal{K}_i^M(w)$, there is certainly no reason to believe that $\mathcal{K}_i^M(w) = \mathcal{K}_i^M(w')$.

One approach to solving this problem is as follows: If w and w' are two worlds in M, we write $w \approx_i w'$ if $\mathcal{K}_i^M(w) = \mathcal{K}_i^M(w')$, i.e., if w and w' agree on the possible worlds according to agent i. We then define

$(M, w) \models N_i\alpha$ if $(M, w') \models \alpha$ for all w' such that $w' \in W - \mathcal{K}_i^M(w)$ and $w \approx_i w'$.

While this definition does capture the first of Levesque's properties, it does not capture the second. To see the problem, suppose we have only one agent and a structure M with only one possible world w. Suppose that $(w, w) \in \mathcal{K}_1^M$ and p is true at w. Then it is easy to see that $(M, w) \models L_1 p \wedge N_1 p$, contradicting axiom **A5**. The problem is that since the structure has only one world and it is in $\mathcal{K}_1^M(w)$, there are no worlds in $W - \mathcal{K}_1^M(w)$. Thus, $N_1 p$ is vacuously true. Intuitively, there just aren't enough "impossible" worlds in this case; the set of conceivable worlds is not independent of the model. To deal with this problem, we focus attention on one particular model, the *canonical model*, which intuitively has "enough" worlds. To define this formally, we first need the notion of a maximal consistent set.

Definition 3.1: A set Γ of basic formulas is a *maximal consistent set (of basic formulas)* iff Γ is K45_n-consistent and no proper superset of Γ is K45_n-consistent. ∎

Every maximal consistent set is satisfiable in some model. The canonical model is one where *every* maximal consistent set is satisfiable at some state.

Definition 3.2: The *canonical model* (for K45_n) $M^c = (W^c, \pi^c, \mathcal{K}_1^c, \ldots, \mathcal{K}_n^c)$ is defined as follows:

- $W^c = \{w \mid w$ is a maximal consistent set of basic formulas wrt $\text{K45}_n\}$
- for all primitive propositions p and $w \in W^c$, $\pi(w)(p) =$ **true** iff $p \in w$
- $(w, w') \in \mathcal{K}_i^c$ iff $w/L_i \subseteq w'$, where $w/L_i = \{\alpha \mid L_i\alpha \in w\}$.

Validity in the canonical-model approach is defined with respect to the canonical model only. More precisely, a formula α is said to be *valid in the canonical-model approach*, denoted $\models^c \alpha$, iff for all worlds w in the canonical model we have $(M^c, w) \models \alpha$.

We now want to argue that, for an appropriate notion of "possibility" and "conceivability", this semantics satisfies the first two of Levesque's properties. What then is a conceivable world? Intuitively, it is an objective state of affairs from agent i's point of view, which does not include i's beliefs. In the single-agent case, this is simply a truth assignment. In the multi-agent case, things are more complicated, since beliefs of other agents are also part of i's objective world. One way of characterizing a state of affairs

from i's point of view is by the set of i-objective formulas that are true at a particular world. For technical reasons, in this section we restrict even further to the i-objective *basic* formulas—that is, those formulas that do not mention any of the modal operators N_j, $j = 1, \ldots, n$—that are true. If we assume that the basic formulas determine all the other formulas, which can be shown to be true in the single-agent case, and under this semantics for the multi-agent case, then it is arguably reasonable to restrict to basic formulas. However, as we shall see in Section 4, it is not clear that this restriction is appropriate, although we make it for now.

Given a situation (M, w), let $obj_i(M, w)$ consist of all the i-objective basic formulas that are true at (M, w). We take $obj_i(M, w)$ to be i's state at (M, w). Notice that $obj_i(M, w)$ is a maximal consistent set of i-objective basic formulas. For ease of exposition, we say *i-set* from now on rather than "maximal set of i-objective basic formulas". Thus, the set of conceivable states for agent i is the set of all i-sets. Notice that the set of conceivable states is independent of the model. It is easy to show that this is a generalization of the single-agent case, since in the single-agent case the i-objective basic formulas are just the propositional formulas, and an i-set can be identified with a truth assignment. Given a situation (M, w), define $Obj_i(M, w) = \{obj_i(M, w') \mid w' \in \mathcal{K}_i^M(w)\}$. Thus, $Obj_i(M, w)$ is the set of i-sets that agent i considers possible in situation (M, w).

With these definitions, we can show that the first two of Levesque's properties hold in the canonical model. The first property says that at all worlds w' considered in evaluating a formula of the form $N_i\varphi$ at a world w, the set of possible states—that is, the set $\{obj_i(M^c, w'') \mid w'' \notin \mathcal{K}_i^c(w')\}$—is the same for all $w' \in \mathcal{K}_i^c(w)$. This is easy to see, since the only worlds w' we consider are those such that $\mathcal{K}_i^c(w') = \mathcal{K}_i^c(w)$. The second property says that the union of the set of states associated with the worlds used in computing $L_i\varphi$ at w and the set of states associated with the worlds used in computing $N_i\varphi$ at w should consist of all conceivable states. To show this, we must show that for every world w in the canonical model, the set $\{obj_i(M^c, w') \mid w' \approx_i w\}$ consists of all i-sets. This follows from

Theorem 3.3: *Let $w \in W^c$. Then for every i-set Γ there is exactly one world w^* such that $obj_i(M^c, w^*) = \Gamma$ and $w \approx_i w^*$.*

What about the third property? This says that every subset of i-sets arises as the set of i-sets associated with the worlds that i considers possible in some situation; that is, for every set S of i-sets, there should be some situation (M^c, w) such that $S = Obj_i(M^c, w)$. As we now show, this property does *not* hold in the canonical model. We do this by showing that the set of i-sets associated with the worlds considered possible in any situation in the canonical model all have a particular property we call *limit closure*.[6]

Definition 3.4: We say that an i-set Γ is a *limit* of a set S of i-sets if for every finite

[6]This turns out to be closely related to the limit closure property discussed in (Fagin, Geanakoplos, Halpern, and Vardi 1992); a detailed comparison would take us too far afield here though.

subset Δ of Γ, there is a set $\Gamma' \in S$ such that $\Delta \subset \Gamma'$. A set S of i-sets is *limit-closed* if every limit of S is in S. ∎

Lemma 3.5: *For every world w in M^c, the set $Obj_i(M^c, w)$ is limit-closed.*

Since there are clearly sets of i-sets that are not limit closed, it follows that this semantics does not satisfy the third property. One consequence of this is a result already proved by Lakemeyer (1993).

Lemma 3.6: (Lakemeyer 1993) *Let p be a primitive proposition. Then $\models^c \neg O_i \neg O_j p$.*

It may seem unreasonable that $\neg O_i \neg O_j p$ should be valid in the canonical-model approach. Why should it be impossible for i to know only that j does not only know p? After all, j can (truthfully) tell i that it is not the case that all he (j) knows is p. We return to this issue in Sections 4 and 5. For now, we focus on a proof theory for this semantics. For the proof theory, which we call AX, we consider the exact analogue of Levesque's axiomatization, now using subscripted modal operators, and replacing **A5** by

$\mathbf{A5}_n$. $N_i\alpha \Rightarrow \neg L_i \alpha$ if $\neg\alpha$ is a K45$_n$-consistent i-objective basic formula.

It is not hard to show that these axioms are sound. We write $\vdash \alpha$ for α *is provable in AX.*

Theorem 3.7: (Lakemeyer 1993) *For all α in $\mathcal{ONL}_n$, if $\vdash \alpha$ then $\models^c \alpha$.*

We show in Section 4 that this axiomatization is incomplete. In fact, the formula $\neg O_i \neg O_j p$ is not provable. Intuitively, part of the problem here is that $\mathbf{A5}_n$ is restricted to basic formulas. For completeness, we would need an analogue of $\mathbf{A5}_n$ for arbitrary formulas. However, we obtain completeness for a restricted language, which we call $\mathcal{ONL}_n^-$.

Definition 3.8: $\mathcal{ONL}_n^-$ consists of all formulas α in $\mathcal{ONL}_n$ such that, in α, no N_j may occur within the scope of an N_i or L_i for $i \neq j$. ∎

For example, while $N_i L_i \neg N_i p$ and $N_i(L_j p \vee N_i \neg p)$ are in $\mathcal{ONL}_n^-$, $N_i N_j p$ and $N_i L_j N_i p$ are not for distinct i and j.

Theorem 3.9: (Lakemeyer 1993) *For all $\alpha \in \mathcal{ONL}_n^-$, $\models^c \alpha$ iff $\vdash \alpha$.*

4 The i-Set Approach

As we have seen, the canonical-model semantics has some attractive features, particularly when restricted to the language $\mathcal{ONL}_n^-$. On the other hand, it has what may be considered undesirable features, such as making $\neg O_i \neg O_j p$ valid. In this section, we consider

a second approach to giving semantics to N_i, developed in (Halpern 1993). As we shall see, it too has its good points and bad points.

In the i-set approach, we maintain the intuition that the set of conceivable states for each agent i can be identified with the set of i-sets. We no longer restrict attention to the canonical model though; we consider all Kripke structures.

We define a new semantics $\models'$ as follows: all the clauses of $\models'$ are identical to the corresponding clauses for $\models$, except that for N_i. In this case, we have

$$(M,w) \models' N_i\varphi \text{ iff } (M',w') \models' \varphi \text{ for all situations } (M',w') \text{ such that } Obj_i(M,w) = Obj_i(M',w') \text{ and } obj_i(M',w') \notin Obj_i(M,w).$$

Notice that $\models$ and $\models'$ agree for basic formulas; in general, as we shall see, they differ. We remark that this definition is equivalent to the one given in (Halpern 1993), except there, rather than i-sets, *i-objective trees* were considered. We did not want to go through the overhead of introducing i-objective trees here, since it follows from results in (Halpern 1993; Halpern 1994) that i-sets are equivalent to i-objective trees: every i-set uniquely determines an i-objective tree and vice versa. Thus, the two definitions are essentially equivalent. Notice that to decide if $N_i\varphi$ holds in (M,w), we consider all situations that agree with (M,w) on the set of possible states, hence this semantics satisfies the first of the three properties we isolated in the single-agent case. It is also clear that the i-sets considered in evaluating the truth of $N_i\varphi$ are precisely those not considered in evaluating the truth of $L_i\varphi$; hence we satisfy the second property. Finally, as we now show, for every set S of i-sets, there is a situation (M,w) such that $Obj_i(M,w) = S$.

Proposition 4.1: *For each set S of i-sets, there is a situation (M,w) such that $Obj_i(M,w) = S$.*

How does this semantics compare to the canonical model semantics? First of all, it is easy to see that the axioms are sound. We write $\models' \varphi$ if $(M,w) \models' \varphi$ for every situation (M,w). Then we have

Theorem 4.2: (Halpern 1993) *For all $\alpha \in \mathcal{ONL}_n$, if $\vdash \alpha$ then $\models' \alpha$.*

Moreover, we again get completeness for the sublanguage $\mathcal{ONL}_n^-$.

Theorem 4.3: (Halpern 1993) *For all $\alpha \in \mathcal{ONL}_n^-$, $\vdash \alpha$ iff $\models' \alpha$.*

This result shows that $\models'$ shares many of the nice features of $\models$. Unfortunately, our axiomatization is not complete for the full language, for neither $\models$ nor $\models'$. Since the axiomatization is sound for both $\models$ and $\models'$, to prove incompleteness, it suffices to provide a formula which is satisfiable with respect to $\models'$ and not $\models$, and another formula which is satisfiable with respect to $\models$ and not $\models'$. As is shown in (Halpern 1993), $\neg O_i \neg O_j p$ is satisfiable with respect to $\models'$ and (by Lemma 3.6) not with respect to $\models$. On the other

hand, it is easy to see that $L_j false \wedge N_j \neg O_i \neg O_j p$ is satisfiable with respect to $\models$ (in fact, it is equivalent to $L_j false$); as we show in the full paper, it is not satisfiable with respect to $\models'$.

The fact that neither $\models$ nor $\models'$ is complete with respect to the axiomatization described earlier is not necessarily bad. We may be able to find a natural complete axiomatization. However, as we suggested above, the fact that $\neg O_j \neg O_i p$ is valid under the $\models$ semantics suggests that this semantics does not quite satisfy our intuitions with regards to only-knowing for formulas in $\mathcal{ONL}_n - \mathcal{ONL}_n^-$. As we now show, $\models'$ also has its problems. We might hope that if φ is a satisfiable i-objective formula, then $N_i\varphi \Rightarrow \neg L_i\varphi$ would be valid under the $\models'$ semantics. Unfortunately, it is not.

Lemma 4.4: *The formula $N_i \neg O_j p \wedge L_i \neg O_j p$ is satisfiable under the $\models'$ semantics.*

Lemma 4.4 shows that although the semantics has the three properties we claimed were appropriate, N_i and L_i still do not always interact in what seems to be the appropriate way. Intuitively, the problem here is that there is more to i's view of a world than just the i-objective basic formulas that are true there. We should really identify i's view of a situation (M, w) with the set of *all* i-objective formulas that are true there. In the canonical-model approach, the i-objective basic formulas that are true at a world can be shown to determine all the i-objective formulas that are true at that world. This is not true at all situations under the i-set approach. We consider a different approach in the next section that attempts to address these problems.

5 What Properties Should Only Knowing Have?

Up to now, we have provided two semantics for only knowing. While both have properties we view as desirable, they also have properties that seem somewhat undesirable. This leads to an obvious question: What properties should only knowing have? Roughly speaking, we would like to have the multi-agent version of Levesque's axioms, and no more. Of course, the problem here is axiom $\mathbf{A5}_n$. It is not so clear what the multi-agent version of that should be. The problem is one of circularity: We would like to be able to say that $N_i\varphi \Rightarrow \neg L_i\varphi$ should hold for any consistent i-objective formula. The problem is that in order to say what the consistent formulas are, we need to define the axiom system. In particular, we have to make precise what this axiom should be.

To deal with this problem, we extend the language so that we can explicitly talk about satisfiability and validity in the language. We add a modal operator *Val* to the language. The formula *Val(φ)* should be read "φ is valid". Of course, its dual *Sat(φ)*, defined as $\neg$ *Val($\neg\varphi$)*, should be read "φ is satisfiable". With this operator in the language, we can replace $\mathbf{A5}_n$ with

$\mathbf{A5}'_n$. *Sat*$(\neg\alpha) \Rightarrow (N_i\alpha \Rightarrow \neg L_i\alpha)$ if α is i-objective.

In addition, we have the following rules for reasoning about validity and satisfiability:

V1. $(Val(\varphi) \wedge Val(\varphi \Rightarrow \psi)) \Rightarrow Val(\psi)$

V2. $Sat(\varphi)$, if φ is a satisfiable propositional formula[7]

V3. $(Sat(\alpha \wedge \beta_1) \wedge \ldots \wedge Sat(\alpha \wedge \beta_k) \wedge Sat(\gamma \wedge \delta_1) \wedge \ldots \wedge Sat(\gamma \wedge \delta_m) \wedge Val(\alpha \vee \gamma)) \Rightarrow$
$Sat(L_i\alpha \wedge \neg L_i\neg\beta_1 \wedge \ldots \wedge \neg L_i\neg\beta_k \wedge N_i\gamma \wedge \neg N_i\neg\delta_1 \wedge \ldots \wedge \neg N_i\neg\delta_m)$,
if $\alpha, \beta_1, \ldots, \beta_k, \gamma, \delta_1, \ldots, \delta_m$ are i-objective formulas

V4. $(Sat(\alpha) \wedge Sat(\beta)) \Rightarrow Sat(\alpha \wedge \beta)$ if α is i-objective and β is i-subjective

$\mathbf{Nec}_V$. From φ infer $Val(\varphi)$.

Axiom **V1** and the rule $\mathbf{Nec}_V$ make *Val* what is called a *normal* modal operator. In fact, it can be shown to satisfy all the axioms of S5. The interesting clauses are clearly **V2–V4**, which capture the intuitive properties of validity and satisfiability.

Let AX′ consist of the axioms for $\mathcal{ONL}$ given earlier together with **V1–V4** and $\mathbf{Nec}_V$, except that $\mathbf{A5}_n$ is replaced by $\mathbf{A5}'_n$. Provability in AX′ is denoted by $\vdash_{\mathrm{AX}'}$. AX′ is the axiom system that provides what we claim is the desired generalization of Levesque's axioms to the multi-agent case. In particular, $\mathbf{A5}'_n$ is the appropriate generalization of **A5**. The question is, of course, whether there is a semantics for which this is a complete axiomatization. We now provide one, in the spirit of the canonical-model construction of Section 3, except that, in the spirit of the extended situations of Section 2, we do not attempt to make the set of worlds used for evaluating L_i and N_i disjoint.[8]

Let $\mathcal{ONL}_n^+$ be the extension of $\mathcal{ONL}_n$ to include the modal operator *Val*. For the remainder of this section, when we say "consistent", we mean consistent with the axiom system AX′. We define the *extended canonical model* $M^e = (W^e, \pi^e, \mathcal{K}_1^e, \ldots, \mathcal{K}_n^e, \mathcal{N}_1^e, \ldots, \mathcal{N}_n^e)$, as follows:

- W^e consist of the maximal consistent sets of formulas in $\mathcal{ONL}_n$.
- for all primitive propositions p and $w \in W^e$, we have $\pi(w)(p) = true$ iff $p \in w$.
- $(w, w') \in \mathcal{K}_i^e$ iff $w/L_i \in w'$.
- $(w', w') \in N_i^e$ iff $w/N_i \in w'$.

In this canonical model, the semantics for N_i is defined in terms of the $\mathcal{N}_i^e$ relation:

$$(M^e, w) \models N_i\alpha \text{ if } (M^e, w') \models \alpha \text{ for all } w' \text{ such that } (w, w') \in \mathcal{N}_i^e.$$

[7] We can replace this by the simpler $Sat(p'_1 \wedge \ldots \wedge p'_k)$, where p'_i is a literal—either a primitive proposition or its negation—and $p'_1 \wedge \ldots \wedge p'_k$ is consistent.

[8] In the full paper, we present a variant of the canonical model construction that does make them disjoint.

We define the *Val* operator so that it corresponds to validity in the extended canonical model:

$$(M^e, w) \models \mathit{Val}(\alpha) \text{ if } (M^e, w') \models \alpha \text{ for all worlds } w' \text{ in } M^e.$$

Using standard modal logic techniques, we can now prove.

Theorem 5.1: *M^e is a $K45_n$ structure (that is, $\mathcal{K}_i^e$ and $\mathcal{N}_i^e$ are Euclidean and transitive). Moreover, $\mathcal{K}_i^e \cup \mathcal{N}_i^e = W^e$, for $i = 1, \ldots, n$, and for each world $w \in W^e$, we have $(M^e, w) \models \alpha$ iff $\alpha \in w$.*

Suppose we define $\models^e \alpha$ iff $(M^e, w) \models \alpha$ for all worlds $w \in W^e$. It immediately follows from Theorem 5.1 that

Corollary 5.2: *$\models^e \alpha$ iff $\vdash_{\mathrm{AX}'} \alpha$.*

Thus, AX′ is a sound and complete axiomatization of $\mathcal{ONL}_n^+$ with respect to the $\models^e$ semantics.

How does this semantics compare to our earlier two? Clearly, they differ. It is easy to see that the formula $O_i \neg O_j p$, which was not satisfiable under $\models^c$, is satisfiable under $\models^e$. In addition, the formula $N_i \neg O_j p \wedge L_i \neg O_j p$, which is satisfiable under $\models'$, is not satisfiable under $\models^e$. In both cases, it seems that the behavior of $\models^e$ is more appropriate. On the other hand, all three semantics agree in the case where our intuitions are strongest, $\mathcal{ONL}_n^-$. Since the axiom system AX characterizes how our earlier two semantics deal with $\mathcal{ONL}_n^-$, this is shown by the following result.

Theorem 5.3: *If $\varphi \in \mathcal{ONL}_n^-$, then $\vdash \varphi$ iff $\vdash_{\mathrm{AX}'} \varphi$.*

Thus, we maintain all the benefits of the earlier semantics with this approach. Moreover, the validity problem for this logic is no harder than that for $\mathrm{K45}_n$ alone. It is PSPACE-complete.

Theorem 5.4: *The problem of deciding if $\vdash_{\mathrm{AX}'} \varphi$ is PSPACE-complete.*

To what extent do the three properties we have been focusing on hold under the $\models^e$ semantics? Suppose we take the conceivable states from i's point of view to be the maximal consistent sets of i-objective formulas with respect to AX′, or equivalently, the set of i-objective formulas true at some world in M^e. Let $obj_i^e(M^e, w)$ consist of all the i-objective formulas true at world w in the extended canonical model (under the $\models^e$ semantics), and let $Obj_i^e(M^e, w) = \{obj_i^e(M^e, w') \mid w' \in \mathcal{K}_i^e(w)\}$. It is easy to see that the first two properties we isolated hold under this interpretation of conceivable state. However, it is quite possible that the "possible states" at a world (M^c, w), that is, $Obj_i^e(M^c, w)$, and the "impossible states", that is, $\{obj_i^e(M^c, w') \mid w' \approx_i w, w \notin \mathcal{K}_i^e(w)\}$ are not disjoint.

Interestingly, this semantics does not satisfy the third property we isolated. Not all subsets of conceivable states arises as the set of possible states at some situation (M^e, w). A proof analogous to that of Lemma 3.5 shows that $Obj_i^e(M, w)$ is always limit-closed. Although we do get limit closure, we avoid problems by having in a precise sense "enough" possibilities.

6 Conclusion

We have provided three semantics for multi-agent only-knowing. All agree on the subset $\mathcal{ONL}_n^-$, but they differ on formulas involving nested N_i's. Although a case can be made that the $\models^e$ semantics comes closest to capturing our intuitions for "knowing at most", our intuitions beyond $\mathcal{ONL}_n^-$ are not well grounded. It would certainly help to have more compelling semantics corresponding to AX'.

On the other hand, it can be argued that semantics does not play quite as crucial a role when dealing with knowing at most as in other cases. The reason is that the structures we must deal with, in general, have uncountably many worlds. For example, whichever of the three semantics we use, there must be uncountably many worlds i-accessible from a situation (M, w) satisfying $O_i p$, at least one for every i-set that includes p. To the extent that we are interested in proof theory, the proof theory associated with $\models^e$, characterized by the axiom system AX', seems quite natural. The fact that the validity problem is no harder in this setting than that for $K45_n$ adds further support to its usefulness.

References

Ben-David, S. and Y. Gafni (1989). All we believe fails in impossible worlds. Manuscript.

Chellas, B. F. (1980). *Modal Logic.* Cambridge, U.K.: Cambridge University Press.

Fagin, R., J. Geanakoplos, J. Y. Halpern, and M. Y. Vardi (1992). The expressive power of the hierarchical approach to modeling knowledge and common knowledge. In Y. Moses (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pp. 229–244. Morgan Kaufmann.

Fagin, R., J. Y. Halpern, and M. Y. Vardi (1992). What can machines know? On the properties of knowledge in distributed systems. *Journal of the ACM 39*(2), 328–376.

Halpern, J. Y. (1993). Reasoning about only knowing with many agents. In *Proc. National Conference on Artificial Intelligence (AAAI '93)*, pp. 655–661.

Halpern, J. Y. (1994). A theory of knowledge and ignorance for many agents. Research Report RJ 9894, IBM. To appear, *Journal of Logic and Computation.*

Halpern, J. Y. and G. Lakemeyer (1995). Levesque's axiomatization of only knowing is incomplete. *Artificial Intelligence 74*(2), 381–387.

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.

Humberstone, I. L. (1986). A more discriminating approach to modal logic. *Journal of Symbolic Logic 51*(2), 503–504. (Abstract only.) There is also an expanded, but unpublished, manuscript.

Lakemeyer, G. (1993). All they know: a study in multi-agent autoepestemic reasoning. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, pp. 376–381.

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