Multi-Agent Only Knowing

Joseph Y. Halpern
IBM Research Division
Almaden Research Center, Dept. K53-B2
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

December 11, 1996

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 simplify Levesque's completeness proof, and 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, he hoped 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 only with 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.

Recently, each of us (Halpern 1993; Lakemeyer 1993a) independently attempted to extend Levesque's logic to the multi-agent case. Although there are a number of similarities in the 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 provide an alternative semantics. The use of the alternative semantics leads to a simplification of Levesque's completeness proof. 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. In Section 5, we consider our new approach. In Section 6, we show how the logic can be used, and discuss its relationship to Moore's *autoepistemic logic* (Moore 1985). Levesque showed that the single-agent version of his logic of only knowing was closely connected to autoepistemic logic. We extend his result to the multi-agent case. We conclude in Section 7 with some discussion of only knowing.

2 Levesque's Logic of Only Knowing

We begin by reconsidering Levesque's definition. Let Φ be a set of primitive propositions. Let $\mathcal{ONL}(\Phi)$ be a propositional modal language formed by starting with the primitive propositions in Φ, and closing off under the classical operators $\neg$ and $\vee$ and two modalities, L and N. We omit the Φ whenever it is clear from context or not relevant to the

[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.

discussion. 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.

$(W, w) \models p$ 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. This is why we should think of $L\alpha$ as saying that the agent knows *at least* α.

It is well-known that this logic is characterized by the axiom system K45. For convenience, we describe K45 here:

Axioms:

P. All instances of axioms of propositional logic
K. $(L\varphi \wedge L(\varphi \Rightarrow \psi)) \Rightarrow L\psi$
4. $L\varphi \Rightarrow LL\varphi$
5. $\neg L\varphi \Rightarrow L\neg L\varphi$

Inference Rules:

R1. From φ and $\varphi \Rightarrow \psi$ infer ψ
R2. From φ infer $L\varphi$

The axioms **4** and **5** are called the *positive introspection axiom* and *negative introspection axiom*, respectively. They are appropriate for agents that are sufficiently introspective so that they know what they know and do not know.

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

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

(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 shall 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

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

Let us call a formula *basic* if there are no occurrences of the operator N; thus, basic formulas may mention at most the operator L.

As it stands, the semantics has the somewhat odd property that there are situations that agree on all basic beliefs yet disagree on what is only believed. As pointed out by Levesque (1990), the problem is that there are far too many sets of worlds than there are basic belief sets. In order to find a perfect match between the sets of basic beliefs an agent may hold and sets of worlds, Levesque introduces what he calls *maximal* sets of worlds. In essence, a maximal set is the largest set in the sense that adding any other world to it would change the agent's basic beliefs. Furthermore, every set of worlds can be extended to a unique maximal set of worlds. It is well known that in the logic K45, an agent's beliefs are completely determined by his beliefs about objective formulas (see, for example, (Halpern and Moses 1984) for a proof). Thus, we define a maximal set as follows:

Definition 2.1: If W is a set of worlds, let

$$W^+ = \{w \mid \text{for all objective formulas } \varphi, \text{ if } (W, w) \models L\varphi \text{ then } (W, w) \models \varphi\}.$$

W is called *maximal* iff $W = W^+$. ∎

Levesque defines validity and satisfiability with respect to *maximal* sets only. In particular, a formula α is valid iff for every maximal set of worlds W and every world $w \in W$, we have $(W, w) \models \alpha$.

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 have all the properties of K45-operators. 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 2.2: (Levesque 1990) *If Φ is infinite, then Levesque's axiomatization is sound and complete for the language $\mathcal{ONL}(\Phi)$ with respect to Levesque's semantics.*

As we shall see, the assumption that there are infinitely many primitive propositions in Φ is crucial for Levesque's completeness result. Extra axioms are required if Φ is finite. In addition, 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 the soundness of Axiom **A5**, it suffices that the sets considered for L and N cover all conceivable worlds; 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 $\models^x$ that 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.

[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).

Theorem 2.3: *For all Φ, Levesque's axiomatization is sound and complete for the language $\mathcal{ONL}(\Phi)$ with respect to $\models^x$.*

Proof: We omit the soundness proof, which is straightforward. Note that for axiom **A5** to be sound it suffices that W_L and W_N together cover all worlds. In particular, it does not matter whether or not the two sets overlap.

To prove completeness, we use the notion of a *maximal consistent* set. Given an arbitrary axiom system AX, we say that a formula φ is *consistent with respect to* AX if it is not the case that $AX \vdash \neg\varphi$, where, as usual, we use $\vdash$ to denote provability. A finite set of formulas $\varphi_1, \ldots, \varphi_n$ is consistent with respect to AX if the conjunction $\varphi_1 \wedge \ldots \wedge \varphi_n$ is consistent with respect to AX. An infinite set of formulas is consistent with respect to AX if every finite subset of its formulas is consistent with respect to AX. Finally, given a set F of formulas, a *maximal consistent* subset of F is a subset F' of F which is consistent with respect to AX such that any superset of F' is not consistent with respect to AX.

In the following, provability, consistency, and maximal consistency all refer to Levesque's axiom system unless stated otherwise. To prove completeness we show that every consistent formula is satisfiable with respect to $\models^x$, using a standard canonical model construction (Halpern and Moses 1992; Hughes and Cresswell 1968).

Let Γ_0 be the set of all maximal consistent sets of formulas in $\mathcal{ONL}(\Phi)$. For $\theta \in \Gamma_0$, define $\theta/L = \{\alpha \mid L\alpha \in \theta\}$ and $\theta/N = \{\alpha \mid N\alpha \in \theta\}$. We then define

- $\Gamma_L^\theta = \{\theta' \in \Gamma_0 \mid \theta/L \subseteq \theta'\}$,
- $\Gamma_N^\theta = \{\theta' \in \Gamma_0 \mid \theta/N \subseteq \theta'\}$.

If we view maximal consistent sets as worlds, then Γ_L^θ and Γ_N^θ represent the worlds accessible from θ for L and N, respectively. The following lemma reflects the fact that L and N are both fully and mutually introspective (axiom **A4**).

Lemma 2.4: *If $\theta' \in \Gamma_L^\theta \cup \Gamma_N^\theta$, then $\Gamma_L^{\theta'} = \Gamma_L^\theta$ and $\Gamma_N^{\theta'} = \Gamma_N^\theta$.*

Proof: We prove the lemma for $\theta' \in \Gamma_L^\theta$. The case $\theta' \in \Gamma_N^\theta$ is completely symmetric. To show that $\Gamma_L^\theta = \Gamma_L^{\theta'}$, it clearly suffices to show that $\theta/L = \theta'/L$. Let $\alpha \in \theta/L$. Then $L\alpha \in \theta$ and also $LL\alpha \in \theta$ by axiom **A4**. Thus $L\alpha \in \theta'$ (since $\theta' \in \Gamma_L^\theta$ implies that $\theta/L \subseteq \theta'$) and, hence, $\alpha \in \theta'/L$.

For the converse, let $\alpha \in \theta'/L$. Thus, $L\alpha \in \theta'$. Assume that $\alpha \notin \theta/L$. Then $\neg L\alpha \in \theta$ (since θ is a maximal consistent set) and, therefore, $L\neg L\alpha \in \theta$, from which $\neg L\alpha \in \theta'$ follows, a contradiction.

The proof that $\Gamma_N^{\theta'} = \Gamma_N^\theta$ proceeds the same way, that is, we show that $\theta/N = \theta'/N$. Let $\alpha \in \theta/N$. Then $N\alpha \in \theta$ and also $LN\alpha \in \theta$ by axiom **A4**. Hence $N\alpha \in \theta'$, so $\alpha \in \theta'/N$.

For the converse, let $\alpha \in \theta'/N$. Thus, $N\alpha \in \theta'$. Assume that $\alpha \notin \theta/N$. Then $\neg N\alpha \in \theta$ and also $L\neg N\alpha \in \theta$, from which $\neg N\alpha \in \theta'$ follows, a contradiction. ∎

In traditional completeness proofs using maximal consistent sets (see, for example, (Halpern and Moses 1992; Hughes and Cresswell 1968)), a situation is constructed whose worlds consists of all maximal consistent sets. Here, we must be a little more careful.

We say that a maximal consistent set θ contains a truth assignment w if for all atomic formulas p, we have $w \models p$ iff $p \in \theta$. Clearly a maximal consistent set θ contains exactly one world; we denote this world by w_θ. For $\theta \in \Gamma_0$, let $W_L^\theta = \{w_{\theta'} \mid \theta' \in \Gamma_L^\theta\}$ and $W_N^\theta = \{w_{\theta'} \mid \theta' \in \Gamma_N^\theta\}$.

Lemma 2.5:

(a) $(W_L^\theta, W_N^\theta, w_\theta)$ *is an extended situation.*

(b) For all α*, we have* $\alpha \in \theta$ *iff* $(W_L^\theta, W_N^\theta, w_\theta) \models^x \alpha$.

Proof: For part (a), to show that $(W_L^\theta, W_N^\theta, w_\theta)$ is an extended situation, we must show that $W_L^\theta \cup W_N^\theta$ consists of all truth assignments. By way of contradiction, suppose there is a truth assignment w not in $W_L^\theta \cup W_N^\theta$. Let $F_w = \{p \in \Phi \mid w \models p\} \cup \{\neg p \mid p \in \Phi,\ w \models \neg p\}$. $F_w \cup \theta/L$ cannot be consistent, for otherwise there would be some $\theta' \in \Gamma_L^\theta$ that contains F_w, which would mean that $w \in W_L^\theta$. Similarly $F_w \cup \theta/N$ cannot be consistent. Thus, there must be formulas $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ such that φ_1 and φ_2 are both conjunctions of a finite number of formulas in F_w, φ_3 is the conjunction of a finite number of formulas in θ/L, and φ_4 is the conjunction of a finite number of formulas in θ/N, and both $\varphi_1 \wedge \varphi_3$ and $\varphi_2 \wedge \varphi_4$ are inconsistent. Thus, we have $\vdash \varphi_3 \Rightarrow \neg\varphi_1$ and $\vdash \varphi_4 \Rightarrow \neg\varphi_2$. Using standard modal reasoning (**A2**, **A3**, and **Nec**), we have $\vdash L\varphi_3 \Rightarrow L\neg\varphi_1$ and $N\varphi_4 \Rightarrow N\neg\varphi_2$. Since $L\psi \in \theta$ for each conjunct ψ of φ_3, standard modal reasoning shows that $L\varphi_3 \in \theta$. Similarly, we have $N\varphi_4 \in \theta$. Since θ is a maximal consistent set, both $L\neg\varphi_1$ and $N\neg\varphi_2$ are in θ. Since $\vdash L\neg\varphi_1 \Rightarrow L(\neg\varphi_1 \vee \neg\varphi_2)$ and $\vdash N\neg\varphi_2 \Rightarrow N(\neg\varphi_1 \vee \neg\varphi_2)$, it follows that both $L(\neg\varphi_1 \vee \neg\varphi_2)$ and $N(\neg\varphi_1 \vee \neg\varphi_2)$ are in θ. But this contradicts **A5**, since $\varphi_1 \wedge \varphi_2$ is a propositionally consistent objective formula.

For part (b), the proof proceeds by induction on the structure of α. The statement holds trivially for atomic propositions, conjunctions, and negations. In the case of $L\alpha$, we proceed by the following chain of equivalences:

$$
\begin{array}{ll}
 & L\alpha \in \theta \\
\text{iff} & \text{for all } \theta' \in \Gamma_L^\theta, \text{ we have } \alpha \in \theta' \\
\text{iff} & \text{for all } \theta' \in \Gamma_L^\theta, \text{ we have } (W_L^{\theta'}, W_N^{\theta'}, w_{\theta'}) \models^x \alpha \text{ (using the induction hypothesis)} \\
\text{iff} & \text{for all } w_{\theta'} \in W_L^\theta, \text{ we have } (W_L^\theta, W_N^\theta, w_{\theta'}) \models^x \alpha \text{ (by Lemma 2.4)} \\
\text{iff} & (W_L^\theta, W_N^\theta, w_\theta) \models^x L\alpha.
\end{array}
$$

The case $N\alpha$ is completely symmetric. ∎

The completeness result now follows easily. Let α be a consistent formula and θ a maximal consistent set of formulas containing α. By Lemma 2.5, $(W_L^\theta, W_N^\theta, w_\theta) \models^x \alpha$. ∎

Levesque considered only maximal sets in his definition of validity. In fact, this restriction has no effect on the notion validity.

Corollary 2.6: *A formula $\alpha \in \mathcal{ONL}(\Phi)$ is valid iff $(W, w) \models \alpha$ for all situations (W, w) (including nonmaximal W).*

Proof: If Φ is finite, it is easy to check that $W^+ = W$ for all sets W, so the result is trivially true if Φ is finite. So suppose Φ is infinite. Notice that each situation (W, w) corresponds to an extended situation (W_L, W_N, w), where $W_L = W$ and W_N is the complement of W. Let us call such an (W_L, W_N, w) an *extended complementary situation.* Theorems 2.2 and 2.3 together imply the valid sentences obtained when considering all extended situations remain the same when we restrict ourselves to complementary situations with maximal W_L. The corollary then follows from the fact that the set of all extended situations properly includes the set of all extended complementary situations, which in turn includes the set of all extended complementary situations with maximal W_L. ∎

As Theorem 2.3 shows, for the $\models^x$ semantics, Levesque's axioms are sound and complete for all sets Φ of primitive propositions. On the other hand, as we said earlier, Levesque's completeness proof (with respect to his semantics) depends crucially on the fact that Φ is infinite. If Φ is finite, Levesque's axioms are still sound with respect to his semantics, but they are no longer complete. For example, if $\Phi = \{p\}$, then $\neg L \neg p \Rightarrow N \neg p$ would be valid under $\models$; this does not follow from the axioms given above. In fact, for each finite set Φ of primitive propositions, we can find a new axiom scheme that, taken together with the previous axioms, gives a complete axiomatization for $\mathcal{ONL}(\Phi)$ for Levesque's semantics if Φ is finite.[4] The new axiom, which subsumes axiom **A5**, allows us to reduce formulas involving N formulas involving only L.

Note that worlds, which are truth assignments to the primitive propositions Φ, are themselves finite if Φ is finite. Hence we can identify a world w with the conjunction of all literals over Φ that 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 $W_{\alpha,\Phi}$ be the set of all worlds (over the primitive propositions Φ) that satisfy α. The axiom system AX_Φ is then obtained from Levesque's system by replacing **A5** by the following axiom:

A5$_\Phi$. $N\alpha \equiv \bigwedge_{w \in W_{\neg\alpha,\Phi}} \neg L \neg w$ if $\neg\alpha$ is a propositionally consistent 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 this axiomatically. Completeness is also very easy to establish. Levesque (1990)

[4] This was also the situation for 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).

showed that in his system, even without **A5**, every formula is provably equivalent to one without nested modalities. With $\mathbf{A5}_\Phi$, we then obtain an equivalent formula that does not mention N. In other words, given a formula consistent with respect to AX_Φ, a satisfying model can be constructed with the usual technique for $K45$ alone.

Theorem 2.7: *AX_Φ is sound and complete for the language $\mathcal{ONL}(\Phi)$ with respect to Levesque's semantics, if Φ is finite.*

3 The Canonical-Model Approach

How do we extend our intuitions about only knowing to the multi-agent case? First we extend the language $\mathcal{ONL}(\Phi)$ to the case of many agents. That is, we now consider a language $\mathcal{ONL}_n(\Phi)$, 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. In the remainder of the paper, we omit the Φ, just writing $\mathcal{ONL}$ and $\mathcal{ONL}_n$, since the set of primitive propositions does not play a significant role. By analogy with the single-agent case, we call a formula *basic* if it does not mention any of the operators N_i ($i = 1, \ldots, n$) and *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 $L_2 p$ or even $L_2 L_1 p$ 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, $q \wedge N_2 L_1 p$ is 1-objective, but $L_1 p$ and $q \wedge L_1 p$ 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 $\mathrm{K45}_n$-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

[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.

(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). In this approach, the semantics of N_i is given as follows:

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

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

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

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. Its worlds consist of all the maximal consistent subsets of basic formulas. (Recall that maximal consistent sets were defined in the proof of Theorem 2.3.) Thus, in some sense, the canonical model has as many worlds as possible.

Definition 3.1: 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:

- $W^c = \{w \mid w \text{ is a maximal consistent set of basic formulas wrt K45}_n\}$
- for all primitive propositions p and $w \in W^c$, $\pi(w)(p) = \textbf{true}$ 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 $M^c \models \alpha$, that is, if 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.

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.

To prove this, we need some preliminary results and definitions.

Definition 3.2: 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). Define $subj_i(M, w) = \{\text{basic } L_i\alpha \mid (M, w)\models L_i\alpha\} \cup \{\text{basic } \neg L_i\alpha \mid (M, w)\not\models L_i\alpha\}$. Thus, $subj_i(M, w)$ characterizes i's basic beliefs. ∎

Lemma 3.3: *Let w and w' be worlds in M^c. Then $w \approx_i w'$ iff agent i has the same basic beliefs at w and w', that is, $subj_i(M^c, w) = subj_i(M^c, w')$.*

Proof: The "only if" direction is immediate because w and w' are assumed to have the same $\mathcal{K}_i$-accessible worlds. To prove the "if" direction, suppose that $subj_i(M^c, w) = subj_i(M^c, w')$ but $w \not\approx_i w'$. Without loss of generality, there is a world $w^* \in \mathcal{K}_i^c(w) - \mathcal{K}_i^c(w')$. By the definition of the canonical model, there must be a basic formula $L_i\alpha \in w'$ such that $\alpha \notin w^*$. By assumption, $L_i\alpha \in w$, contradicting the assumption that $w^* \in \mathcal{K}_i^c(w)$. ∎

Lemma 3.4: *Suppose Γ consists only of i-objective formulas, Σ consists only of i-subjective basic formulas, and Γ and Σ are both $K45_n$-consistent. Then $\Gamma \cup \Sigma$ is $K45_n$-consistent.*

Proof: This follows immediately from part (c) of Proposition 4.1 below. ∎

We can now prove that the set of conceivable states for agent i is the same at all worlds of the canonical model. This follows from the following result.

Theorem 3.5: *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 $w \approx_i w^*$.*

Proof: Let $\Sigma = subj_i(M^c, w)$. Since Γ consists of i-objective basic formulas only, Σ consists of i-subjective formulas, and Γ and Σ are both $K45_n$-consistent, by Lemma 3.4, $\Gamma \cup \Sigma$ is $K45_n$-consistent. Let w^* be a maximal consistent set that contains $\Gamma \cup \Sigma$. Since w and w^* agree on Σ, $w \approx_i w^*$ by Lemma 3.3. The uniqueness of w^* follows by a simple induction argument. ∎

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 3.6: We say that an i-set Γ is a *limit* of a set S of i-sets if, for every finite 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. ∎

[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.

Lemma 3.7: *For every world w in M^c, the set $Obj_i(M^c, w)$ is limit closed.*

Proof: Let w be a world in the canonical model and let Γ be an i-set which is a limit of $Obj_i(M^c, w)$. We want to show that $\Gamma \in Obj_i(M^c, w)$. Let $\Sigma = \{\text{basic } \beta \mid L_i\beta \in w\}$. We claim that $\Gamma \cup \Sigma$ is $\mathrm{K45}_n$-consistent. For suppose not. Then there must be a finite subset $\Delta \subseteq \Gamma$ such that $\Delta \cup \Sigma$ is inconsistent. Since Γ is a limit of $Obj_i(M^c, w)$, there must be some world w' in $Obj_i(M^c, w)$ such that $(M^c, w') \models \Delta$. By construction of the canonical model, since $w' \in \mathcal{K}_i^c(w)$, we must have that $(M^c, w') \models \Sigma$. Thus $\Delta \cup \Sigma$ is consistent, contradicting our assumption.

Since $\Gamma \cup \Sigma$ is consistent, there is a world w^* in the canonical model such that $(M^c, w^*) \models \Gamma \cup \Sigma$. Clearly $\Gamma = obj_i(M^c, w^*)$. Moreover, by construction, we must have $w^* \in \mathcal{K}_i^c(w)$. Thus, $\Gamma \in Obj_i(M^c, w)$, as desired. ∎

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 in (Lakemeyer 1993a), which we reprove here, using an approach that will be useful for later results.

Proposition 3.8: (Lakemeyer 1993a) *If $p \in \Phi$ and $i \neq j$, then $\models^c \neg O_i \neg O_j p$.*

Proof: We first need a definition and a lemma. We say that a basic formula ψ is *(K45)-independent* of a basic formula φ if neither $\vdash_{\mathrm{K45}_n} \varphi \Rightarrow \psi$ nor $\vdash_{\mathrm{K45}_n} \varphi \Rightarrow \neg\psi$ hold.

Lemma 3.9: *If $n \geq 2$ (i.e., there are at least two agents) and $\varphi_1, \ldots, \varphi_m$ are consistent basic i-objective formulas, then there exists a basic i-objective formula ψ of the form $L_j\psi'$ which is independent of each of $\varphi_1, \ldots, \varphi_m$.*

Proof: Define the *depth* of a formula φ, denoted $d(\varphi)$, inductively:

- $d(p) = 0$ for a primitive proposition p
- $d(\varphi \wedge \psi) = \max(d(\varphi), d(\psi))$
- $d(\neg\varphi) = d(\varphi)$
- $d(K_i\varphi) = 1 + (\varphi)$

Suppose that $\varphi_1, \ldots, \varphi_m$ are i-objective formulas such that $\max(d(\varphi_1(, \ldots, d(\varphi_k)) = K$. Let p be an arbitrary primitive proposition, and suppose $j \neq i$. (Such a j exists, since we are assuming $n \geq 2$.) Let ψ be the formula $(L_jL_i)^{K+1}p$, where by $(L_jL_i)^{K+1}$ we mean $K + 1$ occurrences of L_jL_i. Standard model theoretic arguments show that ψ is independent of $\varphi_1, \ldots, \varphi_m$. Very briefly: By results of (Halpern and Moses 1992), we know that φ_i is satisfiable in a treelike structure of depth at most $d(\varphi_i)$, for $i = 1, \ldots, m$.

It is easy to see that this can be extended to two structures, one of which satisfies ψ, and the other of which satisfies $\neg\psi$. Hence, ψ is independent of φ_i.∎

Continuing with the proof of Proposition 3.8, suppose by way of contradiction that $(M^c, w) \models O_i \neg O_j p$. Let $\overline{w}$ be a world such that $(M^c, \overline{w}) \models O_j p$, and let $\Gamma = obj_i(M^c, \overline{w})$. We claim that Γ is a limit of $Obj_i(M^c, w)$. To see this, consider any finite subset Δ of Γ. Let ψ be an i-objective basic formula of the form $L_j\psi'$ which is independent of both the conjunction of the formulas in Δ and of p. The existence of such a formula follows from Lemma 3.9. Let $\Sigma = subj_i(M^c, w)$. By Lemma 3.4, $\Sigma \cup \Delta \cup \{L_j\psi'\}$ is consistent. Thus, there is some world $w' \in W^c$ such that $(M^c, w') \models \Sigma \cup \Delta \cup \{L_j\psi'\}$. By Lemma 3.4 again, there is some world w'' satisfying $p \wedge \neg L_j\psi'$ such that $w'' \approx_i w'$. Since $(M^c, w') \models L_j\psi'$, we cannot have $w'' \in \mathcal{K}_j^c(w')$. It follows that $(M^c, w') \models \neg N_j \neg p$, and hence $(M^c, w') \models \neg O_j p$. Moreover, since w' and w agree on all i-subjective formulas, the canonical model construction guarantees that $w' \approx_i w$. Since $(M^c, w) \models O_i \neg O_j p$, it must be the case that for all u, we have $u \in \mathcal{K}_i^c(w)$ iff $(M^c, u) \models \neg O_j p$. In particular, this means that $w' \in \mathcal{K}_i^c(w)$. Thus, $obj_i(M^c, w') \in Obj_i(M^c, w)$. Moreover, by construction, $\Delta \subseteq obj_i(M^c, w')$. Since Δ was chosen arbitrarily, it follows that Γ is a limit of $Obj_i(M^c, w)$. By Lemma 3.7, $\Gamma \in Obj_i(M^c, w)$. Thus, there is some world $v \in \mathcal{K}_i^c(w)$ such that $obj_i(M^c, v) = \Gamma$. It is a simple property of the canonical-model approach that two worlds that agree on all basic beliefs of an agent also agree on what the agent only believes. Hence, since $\overline{w}$ and v agree on j's basic beliefs, it follows that $(M^c, v) \models O_j p$, contradicting the assumption that $(M^c, w) \models O_i \neg O_j p$. ∎

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.

3.1 A Proof Theory

We now consider an axiomatization for the language. The following axiomatization is exactly like Levesque's except that axiom **A5** now requires K45$_n$-consistency instead of merely propositional consistency. For ease of exposition, we use the same names for the axioms as we did in the single-agent case with a subscript n to emphasize that we are looking at the multi-agent version.

Axioms:

- **A1**$_n$. Axioms of propositional logic
- **A2**$_n$. $L_i(\alpha \Rightarrow \beta) \Rightarrow (L_i\alpha \Rightarrow L_i\beta)$
- **A3**$_n$. $N_i(\alpha \Rightarrow \beta) \Rightarrow (N_i\alpha \Rightarrow N_i\beta)$
- **A4**$_n$. $\sigma \Rightarrow L_i\sigma \wedge N_i\sigma$ if σ is an i-subjective formula
- **A5**$_n$. $N_i\alpha \Rightarrow \neg L_i\alpha$ if $\neg\alpha$ is a K45$_n$-consistent i-objective basic formula

Inference Rules:

MP$_n$. From α and $\alpha \Rightarrow \beta$ infer β
Nec$_n$. From α infer $L_i\alpha$ and $N_i\alpha$

Notice that **A5**$_n$ assumes that α ranges only over *basic* i-objective formulas. We need this restriction in order to appeal to satisfiability in the existing logic K45$_n$.[7] To get a more general version of **A5**$_n$, that applies to arbitrary formulas, we will need to appeal to consistency within the logic that the axioms are meant to characterize. We return to this issue in Section 5. It is not hard to show that these axioms are sound.

Theorem 3.10: (Lakemeyer 1993a) *For all* α *in* $\mathcal{ONL}_n$, *if* $\vdash \alpha$ *then* $\models^c \alpha$.

Proof: The proof proceeds by the usual induction on the length of a derivation. Here we show only the soundness of **A5**$_n$. Suppose α is a basic i-objective formula such that $\neg\alpha$ is K45$_n$-consistent. Thus, there is an i-set containing $\neg\alpha$. By Theorem 3.5, it follows that for each world $w \in W^c$, there is a world $w' \approx_i w$ such that $(M^c, w') \models \neg\alpha$. If $w' \in \mathcal{K}_i^c(w)$, it follows that $(M^c, w) \models \neg L_i\alpha$. If $w' \notin \mathcal{K}_i^c(w)$, then $(M^c, w) \models \neg N_i\alpha$. Thus, $(M^c, w) \models \neg L_i\alpha \vee \neg N_i\alpha$; equivalently, $(M^c, w) \models N_i\alpha \Rightarrow \neg L_i\alpha$. It follows that $\models N_i\alpha \Rightarrow \neg L_i\alpha$. ∎

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 **A5**$_n$ is restricted to basic formulas. For completeness, we would need an analogue of **A5**$_n$ for arbitrary formulas. However, we obtain completeness for a restricted language, which we call $\mathcal{ONL}_n^-$.

Definition 3.11: $\mathcal{ONL}_n^-$ 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, $N_iL_i\neg N_ip$ and $N_i(L_jp \vee N_i\neg p)$ are in $\mathcal{ONL}_n^-$, and N_iN_jp and $N_iL_jN_ip$ are not, for distinct i and j.

To prove completeness for the sublanguage $\mathcal{ONL}_n^-$, we need a preliminary lemma, which describes a normal form for formulas.

Lemma 3.12: *Every formula* α *in* $\mathcal{ONL}_n$ *is provably equivalent to a disjunction of formulas of the following form:*

$$\sigma \wedge L_1\varphi_{10} \wedge \neg L_1\varphi_{11} \wedge \ldots \wedge \neg L_1\varphi_{1m_1} \wedge \ldots \wedge L_n\varphi_{n0} \wedge \neg L_n\varphi_{n1} \wedge \ldots \wedge \neg L_n\varphi_{nm_n} \wedge$$
$$N_1\psi_{10} \wedge \neg N_1\psi_{11} \wedge \ldots \wedge \neg N_1\psi_{1k_1} \wedge \ldots \wedge N_n\psi_{n0} \wedge \neg N_n\psi_{n1} \wedge \ldots \wedge \neg N_n\psi_{nk_n},$$

where σ *is a propositional formula and* φ_{ij} *and* ψ_{ij} *are all i-objective formulas. Moreover, if* α *in* $\mathcal{ONL}_n^-$, *we can assume that* φ_{ij} *and* ψ_{ij} *are i-objective basic formulas.*

[7] Note that this peculiar axiom schema is recursive since satisfiability in propositional K45$_n$ is decidable (Halpern and Moses 1992).

Proof: We proceed by induction on the structure of φ. The only nontrivial cases are if φ is of the form $L_i\varphi'$ or $N_i\varphi'$. If φ is of the form $L_i\varphi'$, then, since $\varphi \in \mathcal{ONL}_n^-$, N_j does not appear in φ' for $j \neq i$. We use the inductive hypothesis to get φ' into the normal form described in the lemma. Notice that N_j does not appear in the normal form for $j \neq i$. We now use the the following equivalences to get $L_i\varphi'$ into the normal form:

- $L_i(\psi \wedge \psi') \Leftrightarrow (L_i\psi \wedge L_i\psi')$
- $L_i(\psi \vee L_i\psi') \Leftrightarrow (L_i\psi \vee L_i\psi')$
- $L_i(\psi \vee \neg L_i\psi') \Leftrightarrow (L_i\psi \vee \neg L_i\psi')$
- $L_iL_i\psi \Leftrightarrow L_i\psi$
- $(\neg L_i \mathit{false} \wedge L_i\neg L_i\psi) \Leftrightarrow \neg L_i\psi$
- $L_iN_i\psi \Leftrightarrow N_i\psi$
- $L_i\neg N_i\psi \Leftrightarrow \neg N_i\psi$.

The first five of these equivalences are standard K45$_n$ properties; the last two are instances of axiom $\mathbf{A4}_n$. Similar arguments work in the case that φ is of the form $N_i\varphi'$. We leave the straightforward details to the reader. ∎

Lemma 3.13: *If S_j is a set of consistent j-sets, $j = 1, \ldots, n$, and σ is a consistent propositional formula, then there is a K45$_n$ situation (M, w) such that $(M, w) \models \sigma$ and $Obj_j(M, w) = S_j$.*

Proof: This follows immediately from part (b) of Proposition 4.1 below. ∎

Lemma 3.14: *If φ and ψ are i-objective basic formulas such that $L_i\varphi \wedge N_i\psi$ is consistent, then $\varphi \vee \psi$ is valid.*

Proof: Suppose that $\neg\varphi \wedge \neg\psi$ is consistent. Then, by Axiom $\mathbf{A5}_n$, $N_i(\varphi \vee \psi) \Rightarrow \neg L_i(\varphi \vee \psi)$ is provable. It follows that $N_i\psi \Rightarrow \neg L_i\varphi$ is provable, contradicting the consistency of $L_i\varphi \wedge N_i\psi$. ∎

Lemma 3.15: *If w, w' are worlds in the canonical model such that $w \approx_i w'$ and $w' \notin \mathcal{K}_i^c(w)$, then there is an i-objective basic formula φ such that $L_i\varphi \in w$ and $\varphi \notin w'$.*

Proof: By the construction of the canonical model, we know that if $w' \notin \mathcal{K}_i^c(w)$, then there is some basic formula φ such that $L_i\varphi \in w$ and $\varphi \notin w'$. From Lemma 3.12, it follows that we can assume without loss of generality that φ is in the normal form described by that lemma. Let A consist of all subformulas of φ that are of the form $L_i\psi$ and do not appear in the scope of any other modal operator. Let φ_A be $\bigwedge_{\psi \in A \cap w} \psi \wedge \bigwedge_{\psi \in A - w} \neg\psi$. It is easy to see that $\varphi_A \in w$ (since each of its conjuncts is). Since $w \approx_i w'$, it follows that $\varphi_A \in w'$. Let φ' be the result of replacing each subformula $L_i\psi$ of φ that is in A by either *true* or *false*, depending on whether $L_i\psi \in w$. By construction, φ' is i-objective. It is easy to see that $\varphi_A \Rightarrow (\varphi \Leftrightarrow \varphi')$ is provable. It follows that $\varphi' \notin w'$. It also follows that $L_i\varphi_A \Rightarrow (L_i\varphi \Leftrightarrow L_i\varphi')$ is provable. Since φ_A is an i-subjective formula, $\varphi_A \Rightarrow L_i\varphi_A$ is provable. Hence, $L_i\varphi_A \in w$. Since $L_i\varphi \in w$, it follows that $L_i\varphi' \in w$. This gives us the desired result. ∎

Theorem 3.16: (Lakemeyer 1993a) *For all $\alpha \in \mathcal{ONL}_n^-$, if $\models^c \alpha$ then $\vdash \alpha$.*

Proof: As usual, it suffices to show that if the formula $\alpha \in \mathcal{ONL}_n^-$ is consistent, then it is satisfiable in the canonical model. Without loss of generality, we can assume that α is in the normal form described in Lemma 3.12:

$$\begin{array}{l}\sigma \wedge L_1\varphi_{10} \wedge \neg L_1\varphi_{11} \wedge \ldots \wedge \neg L_1\varphi_{1m_1} \wedge \ldots \wedge L_n\varphi_{n0} \wedge \neg L_n\varphi_{n1} \wedge \ldots \wedge \neg L_n\varphi_{nm_n} \wedge \\ N_1\psi_{10} \wedge \neg N_1\psi_{11} \wedge \ldots \wedge \neg N_1\psi_{1k_1} \wedge \ldots \wedge N_n\psi_{n0} \wedge \neg N_n\psi_{n1} \wedge \ldots \wedge \neg N_n\psi_{nk_n}.\end{array}$$

Moreover, since $\alpha \in \mathcal{ONL}_n^-$, we can assume that φ_{ij} and ψ_{ij} are i-objective basic formulas. Let A_i consist of all the consistent formulas of the form $\varphi_{i0} \wedge \psi_{i0} \wedge \neg\varphi_{ij}$ or $\varphi_{i0} \wedge \psi_{i0} \wedge \neg\psi_{ij}$, $j \geq 1$. Let ξ_i be a formula that is independent of all the formulas in A_i; such a formula exists by Lemma 3.9. Let S_i consist of all i-sets containing $\varphi_{i0} \wedge (\neg\psi_{i0} \vee (\psi_{i0} \wedge \xi_i))$. By Lemma 3.13, there is a K45$_n$ structure (M, w) such that $Obj_i(M, w) = S_i$, $i = 1, \ldots, n$, and $(M, w) \models \sigma$. Thus, there must be a world w^* in the canonical model such that $w^* = \{\text{basic } \varphi' \mid (M, w) \models \varphi'\}$. We claim that $(M^c, w^*) \models \alpha$.

To see this, let α' be the formula $\sigma \wedge L_1\varphi_{10} \wedge \neg L_1\varphi_{11} \wedge \ldots \wedge \neg L_1\varphi_{1m_1} \wedge \ldots \wedge L_n\varphi_{n0} \wedge \neg L_n\varphi_{n1} \wedge \ldots \wedge \neg L_n\varphi_{nm_n}$. We first show that $(M, w) \models \alpha'$. By construction, we have that $(M, w) \models \sigma$. Furthermore, by definition, each world $w' \in \mathcal{K}_i^M(w)$ satisfies φ_{i0}, so we have that $(M, w) \models L_i\varphi_{i0}$. Since $L_i\varphi_{i0} \wedge \neg L_i\varphi_{ij}$ is consistent for each $j \geq 1$, it must be the case that $\varphi_{i0} \wedge \neg\varphi_{ij}$ is consistent. Thus, one of $\varphi_{i0} \wedge \neg\psi_{i0} \wedge \neg\varphi_{ij}$ or $\varphi_{i0} \wedge \psi_{i0} \wedge \neg\varphi_{ij}$ is consistent. If the latter is consistent, then by the choice of ξ, $\varphi_{i0} \wedge \psi_{i0} \wedge \xi_i \wedge \neg\varphi_{ij}$ must be consistent as well. Since S_i consists of all i-sets containing $\varphi_{i0} \wedge (\neg\psi_{i0} \vee (\psi_{i0} \wedge \xi_i))$, it follows that there must be an i-set in S_i containing $\neg\varphi_{ij}$. It follows that $(M, w) \models \neg L_i\varphi_{ij}$, for $j \geq 1$. Thus, we have shown that $(M, w) \models \alpha'$. Since (M, w) and (M^c, w^*) agree on basic formulas, it follows that $(M^c, w^*) \models \alpha'$.

Next, we show that $(M^c, w^*) \models N_1\psi_{10} \wedge \ldots \wedge N_n\psi_{n0}$. To this end, suppose that $w' \approx_i w^*$ and $w' \notin \mathcal{K}_i^c(w^*)$. By Lemma 3.15, there must be some i-objective basic formula φ' such that $L_i\varphi' \in w^*$ and $\neg\varphi' \in w'$. Since $L_i\varphi' \in w^*$, it follows that $(M, w) \models L_i\varphi'$, and hence φ' is in every i-set in S_i. It follows that $obj_i(w') \notin S_i$. Now, one of the following four formulas must be in $obj_i(w')$: (1) $\varphi_{i0} \wedge \psi_{i0}$, (2) $\varphi_{i0} \wedge \neg\psi_{i0}$, (3) $\neg\varphi_{i0} \wedge \psi_{i0}$, (4) $\neg\varphi_{i0} \wedge \neg\psi_{i0}$. Since $L_i\varphi_{i0} \wedge N_i\psi_{i0}$ is consistent, it cannot be (4), by Lemma 3.14. It cannot be (2), for otherwise w' would be in S_i. Thus, it must be (1) or (3), so $\psi_{i0} \in obj_i(w')$. Since this is true for all w' such that $w' \approx_i w^*$ and $w' \notin \mathcal{K}_i^c(w^*)$, it follows that $(M^c, w^*) \models N_i\psi_{i0}$, for $i = 1, \ldots, n$.

Finally, we must show that $(M^*, w^*) \models \neg N_1\psi_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k_i$. Clearly $\psi_{i0} \wedge \neg\psi_{ij}$ is consistent, for otherwise $N_i\psi_{i0} \wedge \neg N_i\psi_{ij}$ would be inconsistent. Thus, at least one of (1) $\psi_{i0} \wedge \neg\psi_{ij} \wedge \neg\varphi_{i0}$ or (2) $\psi_{i0} \wedge \neg\psi_{ij} \wedge \varphi_{i0}$ is consistent. In case (2), by choice of ξ_i, the formula $\psi_{i0} \wedge \neg\psi_{ij} \wedge \varphi_{i0} \wedge \neg\xi_i$ is consistent. Let β be $\psi_{i0} \wedge \neg\psi_{ij} \wedge \neg\varphi_{i0}$ if it is consistent, and $\psi_{i0} \wedge \neg\psi_{ij} \wedge \varphi_{i0} \wedge \neg\xi_i$ otherwise. By construction, β is consistent. By Lemma 3.4, there is a situation (M', v) such that $subj_i(M', v) = subj_i(M^*, w^*)$ and $(M', v) \models \beta$. There is a world w'' in the canonical model which agrees with (M', v) on the basic formulas. By construction, we have $w'' \approx_i w^*$. Moreover, since $(M^c, w) \models L_i(\varphi_{i0} \wedge (\neg\psi_{i0} \vee (\psi_{i0} \wedge \xi)))$, it follows that $(M^c, w^*) \models L_i\neg\beta$. Since $(M^c, w') \models \beta$, we have that $w' \notin \mathcal{K}_i^c(w^*)$. Moreover, since $(M^c, w') \models \neg\psi_{ij}$, it follows that $(M^c, w^*) \models \neg N_i\psi_{ij}$, as desired. This completes the proof. ∎

3.2 Discussion

As we have shown, the canonical-model semantics for N_i has some attractive features, in particular when restricted to the language $\mathcal{ONL}_n^-$. It is for this sublanguage that we have a nice proof-theoretic characterization. There is some evidence, however, that the semantics may not have the behavior we desire when we move beyond $\mathcal{ONL}_n^-$. For one thing, the formula $\neg O_i\neg O_j p$ is valid in the canonical model: it is impossible that all i knows is that it is not the case that all j knows is p. While it is certainly consistent for $\neg O_i\neg O_j p$ to hold, it seems reasonable to have a semantics that allows $O_i\neg O_j p$ to be hold as well. As we have seen, the validity of $\neg O_i\neg O_j p$ is intimately connected with the fact that the canonical-model semantics does not have the third property of Levesque's semantics in the single-agent case: not all subsets of conceivable states are possible. In the next section, we discuss a different approach to giving semantics to only knowing—essentially that taken in (Halpern 1993)—that has all three of Levesque's properties. It agrees with the canonical-model approach on formulas in $\mathcal{ONL}_n^-$, but makes $\neg O_i\neg O_j p$ satisfiable. Unfortunately, as we shall see, it too suffers from problems.

4 The i-Set Approach

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 that 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.

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 $Obj_i(M,w) = S$.

In fact, we prove an even stronger result. Let $obj_i^+(M,w)$ consist of all i-objective formulas (not necessarily just i-objective basic formulas) true at (M,w) (with respect to $\models'$), and let $Obj_i^+(M,w) = \{obj_i^+(M,w') \mid w' \in \mathcal{K}_i^M(w)\}$.

Proposition 4.1: *Let Γ be a satisfiable set of i-objective formulas, let S_i be a set of maximal satisfiable sets of i-objective formulas, $i = 1,\ldots,n$, let Σ be a satisfiable set of i-subjective formulas, and let σ be a propositional formula. Then*

(a) there exists a situation (M_1,w_1) such that $\Gamma \subseteq obj_i^+(M_1,w_1)$ and $S_i = Obj_i^+(M_1,w_1)$.

(b) there exists a situation (M_2,w_2) such that $(M_2,w_2) \models \sigma$ and $Obj_j^+(M_2,w_2) = S_j$, $j = 1,\ldots,n$.

(c) there exists a situation (M_3,w_3) such that $(M_3,w_3) \models \Gamma \wedge \Sigma$.

Proof: For part (a), we first show that, given an arbitrary situation (M,w), we can construct a situation (M^*,w^*) such that $obj_i^+(M^*,w^*) = obj_i^+(M,w)$ and there are no worlds i-accessible from w^*. The idea is to have M^* be the result of adding w^* to the worlds in M, where w^* is just like w except that it has no i-accessible worlds and w^* is not accessible from any world. More formally, if $M = (W,\pi,\mathcal{K}_1,\ldots,\mathcal{K}_n)$, we take $M^* = (W^*,\pi^*,\mathcal{K}_1^*,\ldots,\mathcal{K}_n^*)$, where $W^* = W \cup \{w^*\}$, $\pi^*(w') = \pi(w')$ for $w' \in W$, $\pi^*(w^*) = \pi(w)$, $\mathcal{K}_j^* = \mathcal{K}_j \cup \{(w^*,w') \mid w' \in \mathcal{K}_j(w)\}$ for $j \neq i$, and $\mathcal{K}_i^* = \mathcal{K}_i$. It is easy to see that $\mathcal{K}_j^*$ is Euclidean and transitive. By construction, there are no worlds i-accessible from w^* and $(w',w^*) \notin \mathcal{K}_j^*$ for all w' and all j. Moreover, if ψ is an i-objective formula, we have $(M^*,w^*) \models' \psi$ iff $(M,w) \models' \psi$, since for $j \neq i$, we have $\mathcal{K}_j(w^*) = \mathcal{K}_j(w)$. In particular, this means that $obj_i^+(M^*,w^*) = obj_i^+(M,w)$.

For each $\Delta \in S' = S_i \cup \{\Gamma\}$, there is a situation $(M^\Delta, w^\Delta) \models' \Delta$, where $M^\Delta = (W^\Delta, \pi^\Delta, \mathcal{K}_1^\Delta, \ldots, \mathcal{K}_n^\Delta)$. By the argument above, we can assume without loss of generality that there are no worlds i-accessible from w^Δ and w^Δ is not accessible from any world. We define $M_1 = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ by taking W to be the union of all the worlds in W^Δ, $\Delta \in S'$. (We can assume without loss of generality that these are disjoint sets of worlds.) We define π so that $\pi|_{W^\Delta} = \pi^\Delta$. We define $\mathcal{K}_j = \cup_{\Delta \in S'} \mathcal{K}_j^\Delta$ for $j \neq i$, and $\mathcal{K}_i$ to be the least transitive, Euclidean set containing $\cup_{\Delta \in S'} \mathcal{K}_i^\Delta \cup \{(w^\Gamma, w^\Delta) \mid \Delta \in S\}$. It is easy to check that $obj_i^+(M_1, w^\Delta) = obj_i^+(M^\Delta, w^\Delta)$ (although this depends on the fact that w_Δ is not j-accessible from any world for $j \neq i$). Thus, $Obj_i^+(M_1, w^\Gamma) = S_i$ and $obj_i^+(M_1, w^\Gamma) \supseteq \Gamma$. Thus, we can take $w_1 = w^\Gamma$, completing the proof of part (a).

To summarize the construction of part (a), we start with an arbitrary situation (M, w) satisfying Γ, convert it to a situation satisfying $L_i \mathit{false} \wedge \Gamma$, essentially by modifying the i-accessibility relation at w so that there are no worlds i-accessible from w and w is not accessible from any world, and then again modifying the i-accessibility relation at w so that we get a structure (M_1, w_1) such that $Obj_i^+(M_i, w_i) = S_i$. Note that in doing this construction, we did not change the propositional formulas true at w, nor did we change the worlds that were j-accessible from w for $j \neq i$. Thus, starting with a situation that satisfies a propositional formula σ, we can repeat this construction for each i in turn, for $i = 1, \ldots, n$. The resulting situation is (M_2, w_2), and it clearly has the desired properties.

For part (c), suppose $(M', w') \models \Sigma$, and let $Obj_i^+(M', w') = S_i$. By part (a), there is a situation (M_3, w_3) such that $Obj_i^+(M_3, w_3) = S_i$ and $(M_3, w_3) \models \Gamma$. Since the set of subjective formulas true at a situation (M, w) is completely determined by $Obj_i^+(M, w)$, and $Obj_i^+(M', w') = Obj_i^+(M_3, w_3)$, it follows that $(M_3, w_3) \models \Sigma$ as well. ∎

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 the following result.

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

Proof: As usual, the proof is by induction on the length of a derivation. All that needs to be done is to show that all the axioms are sound. Again, this is straightforward. The proof in the case of $\mathbf{A5}_n$ proceeds just as that in the proof of Theorem 3.10, using the fact that this semantics satisfies Levesque's second property. ∎

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 $\models' \alpha$.*

Proof: As usual, it suffices to show that if α is consistent with the axioms, then α is satisfiable under the $\models'$ semantics. From Theorem 3.16, we know that α is satisfiable in the canonical model under the $\models$ semantics. Thus, it suffices to show that for all formulas $\alpha \in \mathcal{ONL}_n^-$, we have $(M^c, w) \models \alpha$ iff $(M^c, w) \models' \alpha$. By Lemma 3.12, it suffices to consider formulas α in normal form. We proceed by induction on the structure of formulas. The only nontrivial case obtains if α is of the form $N_i\alpha'$. Since α is in normal form, we can assume that α' is basic. Suppose $(M^c, w) \models' N_i\alpha'$. To show that $(M^c, w) \models N_i\alpha'$, we must show that if $w' \approx_i w$ and $w' \notin \mathcal{K}_i^c(w)$, then $(M^c, w') \models \alpha$. By definition, if $w' \approx_i w$, then $Obj_i(M^c, w') = Obj_i(M^c, w)$. Moreover, we must have $obj_i(M^c, w') \notin Obj_i(M^c, w)$, for otherwise we would have $w' \in \mathcal{K}_i^c(w)$. Hence, we must have $(M^c, w') \models' \alpha'$. By the induction hypothesis, we have $(M^c, w') \models \alpha'$. Thus, $(M^c, w) \models N_i\alpha'$, as desired.

For the converse, suppose that $(M^c, w) \models N_i\alpha'$. We want to show that $(M^c, w) \models' N_i\alpha'$. Suppose that (M', w') is such that $Obj_i(M', w') = Obj_i(M^c, w)$ and $obj_i(M', w') \notin Obj_i(M^c, w)$. We must show that $(M', w') \models' \alpha$. It is easy to see that for every situation (M, w) and basic formula φ, we have that $(M, w) \models L_i\varphi$ iff $(M, w) \models' L_i\varphi$ iff φ is in every set in $Obj_i(M, w)$. Thus, it follows that $subj_i(M', w') = subj_i(M^c, w)$. There must be a world w'' in M^c such that (M^c, w'') agrees with (M', w') on all basic formulas according to the $\models$ semantics. Since $subj_i(M^c, w'') = subj_i(M^c, w)$, it follows from Lemma 3.3 that $w \approx_i w''$. Since $obj_i(M', w') \notin Obj_i(M^c, w)$ and $obj_i(M', w') = obj_i(M^c, w'')$, it follows that $w'' \notin \mathcal{K}_i^c(w)$. Since $(M^c, w) \models N_i\alpha'$, we must have that $(M^c, w'') \models \alpha'$. And since (M^c, w'') and (M', w') agree on basic formulas, it follows that $(M', w') \models \alpha'$. Finally, since $\models$ and $\models'$ agree for basic formulas, we have $(M', w') \models' \alpha'$. This completes the proof that $(M, w) \models N_i\alpha'$. ∎

Although our axiomatization is complete for $\mathcal{ONL}_n^-$, as we now show, it 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 Proposition 4.4, $O_i\neg O_jp$ is satisfiable with respect to $\models'$ and (by Proposition 3.8) not with respect to $\models$. On the other hand, it is easy to see that $L_j\mathit{false} \wedge N_j\neg O_i\neg O_jp$ is satisfiable with respect to $\models$ (in fact, it is equivalent to $L_j\mathit{false}$); as shown in Proposition 4.5, it is not satisfiable with respect to $\models'$.

Proposition 4.4: *$O_i\neg O_jp$ is satisfiable under the $\models'$ semantics.*

Proof: Let $S = \{obj_i^+(M, w) \mid (M, w) \models' \neg O_jp\}$. By Proposition 4.1, there is a situation (M^*, w^*) such that $Obj_i^+(M^*, w^*) = S$. We claim that $(M^*, w^*) \models' O_i\neg O_jp$. Clearly $(M^*, w^*) \models' L_i\neg O_jp$, since $\neg O_jp$ is true at all worlds i-accessible from w^*. To see that $(M^*, w^*) \models' N_iO_jp$, suppose that $Obj_i(M, w) = Obj_i(M^*, w^*)$ and $obj_i(M, w) \notin Obj_i(M^*, w^*)$. We want to show that $(M, w) \models' O_jp$. Suppose that $(M, w) \models' \neg O_jp$. By definition, $obj_i^+(M, w) \in S$, so there is some world $w' \in \mathcal{K}_i^{M^*}(w^*)$ such that $obj_i^+(M^*, w') = obj_i^+(M, w)$. In particular, this means that $obj_i(M^*, w') = obj_i(M, w)$. But this contradicts the assumpton that $obj_i(M, w) \notin Obj_i(M^*, w^*)$. Thus, $(M, w) \models O_jp$ as desired, and $(M^*, w^*) \models O_i\neg O_jp$. ∎

Proposition 4.5: *There is a formula β such that $\models' \beta$ but $\neg\beta$ is satisfiable under the $\models$ semantics.*

Proof: We first show that if φ is an i-objective formula that is satisfiable under the $\models'$ semantics, then $\models' L_i \mathit{false} \Rightarrow \neg N_i \neg\varphi$. For suppose that φ is satisfiable in a situation (M, w). By Proposition 4.1, there is a situation (M^*, w^*) such that $obj_i^+(M^*, w^*) = obj_i^+(M, w)$ and $Obj_i^+(M^*, w^*) = \emptyset$. This means that $(M^*, w^*) \models' \varphi \wedge L_i \mathit{false}$. Now let (M', w') be any situation satisfying $L_i \mathit{false}$. Then $Obj_i(M', w') = Obj_i(M^*, w^*) = \emptyset$, and $obj_i(M^*, w^*) \notin Obj_i(M', w')$. It follows that $(M', w') \models' \neg N_i \neg\varphi$. Thus, we have shown that $\models' L_i \mathit{false} \Rightarrow \neg N_i \neg\varphi$. Since, as we showed in Proposition 4.4, the formula $O_j \neg O_i p$ is satisfiable, this means that $\models' L_i \mathit{false} \Rightarrow \neg N_i \neg O_j \neg O_i p$. On the other hand, since $O_j \neg O_i p$ is not satisfiable with respect to $\models$, as we showed in Proposition 3.8, neither is $\neg N_i \neg O_j \neg O_i p$, and hence $L_i \mathit{false} \Rightarrow \neg N_i \neg O_j \neg O_i p$ is not valid under the $\models$ semantics. Indeed, $L_i \mathit{false} \wedge N_i \neg O_j \neg O_i p$ is equivalent to $L_i \mathit{false}$ under the $\models$ semantics. ∎

We can now show that our axiom system is incomplete for the full language with respect to both the $\models$ and $\models'$ semantics.

Theorem 4.6: *There exist formulas α and β in $\mathcal{ONL}_n$ such that $\not\vdash \alpha$ and $\models \alpha$, and $\not\vdash \beta$ and $\models' \beta$.*

Proof: By Propositions 3.8 and 4.4, we have that $\models \neg O_j \neg O_i p$, but $\not\models' \neg O_j \neg O_i p$. Since $\vdash$ is sound with respect to $\models'$, we cannot have $\vdash \neg O_j \neg O_i p$ (for otherwise we would have $\models' \neg O_j \neg O_i p$). Thus, we can take α to be $\neg O_j \neg O_i p$. A similar argument shows we can take β to be $L_i \mathit{false} \Rightarrow \neg N_i \neg O_j O_i p$. ∎

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

Proof: First we show that for any situation (M, w) that satisfies $O_j p$, there exists another situation (M', w') such that (M, w) and (M', w') agree on all basic formulas, but $(M', w') \models' \neg O_j p$. We can construct (M', w') as follows: Choose a particular set $\Gamma \in Obj_j(M, w)$. It easily follows from Proposition 4.1 that there is a situation (M', w') such that $Obj_j(M', w') = Obj_j(M, w) - \{\Gamma\}$ and $obj_j(M', w') = obj_j(M, w)$. We now show that for any basic formula φ, we have $(M, w) \models' \varphi$ iff $(M', w') \models' \varphi$. If φ is a j-objective formula, this is immediate from the construction. Thus, it suffices to deal with the case that φ is of the form $L_j\varphi'$. By Lemma 3.12, we can assume without loss of generality that φ' is j-objective. Suppose that φ' is a consistent j-objective formula. In this case, it is almost immediate from the definitions that if $(M, w) \models' L_j\varphi'$ then

$(M', w') \models' L_j\varphi'$. For the converse, suppose that $(M, w) \models' \neg L_j\varphi'$. Then there is some world $w'' \in \mathcal{K}_j^M(w)$ such that $(M, w'') \models \neg\varphi'$. Since $(M, w) \models' L_j p$, we have that $(M, w'') \models' p \wedge \neg\varphi'$. Let ψ be a j-objective basic formula that is independent of $p \wedge \neg\varphi'$. Let φ'' be $\varphi' \wedge p \wedge \neg\psi$ if $\psi \in \Gamma$, and $\varphi' \wedge p \wedge \psi$ otherwise. Since ψ is independent of $p \wedge \neg\varphi'$, it follows that φ'' is consistent. Let Δ be any j-set containing φ''. It must be the case that $\Delta \in \mathit{Obj}_j(M, w)$, for if not, let (M^*, w^*) be a situation such that $\mathit{Obj}_j(M^*, w^*) = \mathit{Obj}_j(M, w)$ and $\mathit{obj}_j(M^*, w^*) = \Delta$ (such a situation exists by Proposition 4.1). Then $(M^*, w^*) \models p$, contradicting the assumption that $(M, w) \models N_j\neg p$. By construction, $\Delta \neq \Gamma$. Thus, there is some world $v \in \mathcal{K}_j^{M'}(w')$ such that $\mathit{obj}_j(M', v) = \Delta$. It follows that $(M', v) \models' \neg\varphi'$, so $(M', w') \models' \neg L_j\varphi'$. Thus, (M', w') agrees with (M, w) on all basic formulas. However, since $\Gamma \notin \mathit{Obj}_j(M, w)$, it follows that $(M', w') \models \neg N_j\neg p$, and hence that $(M', w') \models \neg O_j p$.

Let $S = \{\mathit{obj}_i^+(M, w) \mid (M, w) \models' \neg O_j p\}$. By Proposition 4.1, there is a situation (M^*, w^*) such that $\mathit{Obj}_i^+(M^*, w^*) = S$. Clearly $(M^*, w^*) \models' L_i\neg O_j p$. We now show that $(M^*, w^*) \models' N_i\neg O_j p$ as well. For suppose that (M, w) is a situation such that $\mathit{Obj}_i(M, w) = \mathit{Obj}_i(M^*, w^*)$ and $\mathit{obj}_i(M, w) \notin \mathit{Obj}_i(M^*, w^*)$. Moreover, suppose, by way of contradiction, that $(M, w) \models' O_j p$. By the arguments above, it follows that there is a situation (M', w') such that $(M', w') \models' \neg O_j p$ and (M, w) and (M', w') agree on all basic formulas. By construction, $\mathit{obj}_i^+(M', w') \in S = \mathit{Obj}_i^+(M^*, w^*)$, so $\mathit{obj}_i(M', w') \in \mathit{Obj}_i(M^*, w^*)$. Since $\mathit{obj}_i(M, w) = \mathit{obj}_i(M', w')$, we must also have $\mathit{obj}_i(M, w) \in \mathit{Obj}_i(M^*, w^*)$, contradicting the choice of (M, w). Thus, $(M, w) \models \neg O_j p$, as desired, and so $(M^*, w^*) \models L_i\neg O_j p \wedge N_i\neg O_j p$. ∎

Lemma 4.7 shows that although the i-set 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.

Indeed, it is no longer true that the i-set approach has the second of the three properties once we take i's view of (M, w) to be $\mathit{obj}_i^+(M, w)$. For consider the situation (M^*, w^*) constructed in the proof of Lemma 4.7. As the proof of that lemma shows, $\{\mathit{obj}_i^+(M^*, w) \mid w \in \mathcal{K}_i^{M^*}(w^*)\} \cup \{\mathit{obj}_i^+(M', w') \mid \mathit{obj}_i(M', w') \notin \mathit{Obj}_i(M^*, w^*)\}$ does not include all maximal sets of i-objective formulas. In particular, it does not include those maximal sets that satisfy $O_j p$. We consider a different approach in the next section that attempts to address this problem.

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 $Val(\varphi)$ should be read "φ is valid". Of course, its dual $Sat(\varphi)$, 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[8]

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.

If we restrict to basic formulas, then **V3** simplifies to $(Sat(\alpha \wedge \beta_1) \wedge \ldots \wedge Sat(\alpha \wedge \beta_k)) \Rightarrow Sat(L_i\alpha \wedge \neg L_i\neg\beta_1 \wedge \ldots \wedge \neg L_i\neg\beta_k)$ (we can take $\gamma, \delta_1, \ldots, \delta_m$ to be *true* to get this). The soundness of this axiom (interpreting *Sat* as satisfiability) follows using much the same arguments as those in the proof of Proposition 4.1. The soundness of **V4** if we restrict to basic formulas follows from Lemma 3.4. More interestingly, it follows from the completeness proof given below that these axioms completely characterize satisfiability in

[8] 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.

$K45_n$; together with the $K45_n$ axioms, they provide a sound and complete language for the language augmented with the *Val* operator.

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 $\mathbf{A5}'_n$. 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.

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*, denoted $M^e = (W^e, \pi^e, \mathcal{K}_1^e, \ldots, \mathcal{K}_n^e, \mathcal{N}_1^e, \ldots, \mathcal{N}_n^e)$, as follows:

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

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

$$\begin{array}{lll} (M^e, w) \models L_i\alpha & \text{if} & (M^e, w') \models \alpha \text{ for all } w' \text{ such that } (w, w') \in \mathcal{K}_i^e. \\ (M^e, w) \models N_i\alpha & \text{if} & (M^e, w') \models \alpha \text{ for all } w' \text{ such that } (w, w') \in \mathcal{N}_i^e. \end{array}$$

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

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

We now want to show that every formula in Γ is satisfied at a world in the extended canonical model. To do this, we need one preliminary result, showing that *Val* and *Sat* really correspond to provability and consistency in this framework.

Proposition 5.1: *For every formula $\varphi \in \mathcal{ONL}_n$, if φ is provable then so is* $\mathrm{Val}(\varphi)$*, while if φ is not provable, then* $\neg\mathrm{Val}(\varphi)$ *is provable.*

Proof: By $\mathbf{Nec}_V$, it is clear that if φ is provable, so is $\mathit{Val}(\varphi)$. Thus, it remains to show that if φ is not provable, then $\neg\mathit{Val}(\varphi)$ is. Using **V1**, it is easy to see that $\neg\mathit{Val}(\varphi)$ is provably equivalent to $\mathit{Sat}(\neg\varphi)$, so it suffices to show that if φ is not provable—i.e., if $\neg\varphi$ is consistent—then $\mathit{Sat}(\neg\varphi)$ is provable. We prove by induction on φ that if φ is consistent, then $\mathit{Sat}(\varphi)$ is provable.

If φ is propositional, the result is immediate from **V2**. For the general case, we first use Lemma 3.12 to restrict attention to formulas in the canonical form specified by the lemma. Using standard modal reasoning (**V1** and $\mathbf{Nec}_V$) it is easy to show that $\vdash$ *Sat*$(\varphi \vee \psi) \Leftrightarrow ($*Sat*$(\varphi) \vee$ *Sat*$(\psi))$. Thus, it suffices to restrict attention to a conjunction in the form specified by the lemma. It is easy to see that if the conjunction is consistent, then each conjunct must be consistent. Using **V4**, it is easy to see that we can restrict attention to i-subjective formulas. By applying Lemma 3.12, we can assume without loss of generality that we are dealing with a consistent formula φ of the form $L_i\alpha \wedge \neg L_i\neg\beta_1 \wedge \ldots \wedge \neg L_i \neg \beta_k \wedge N_i\gamma \wedge \neg N_i \neg \delta_1 \wedge \ldots \wedge \neg N_i \neg \delta_m$, where $\alpha, \beta_1, \ldots, \beta_k, \gamma, \delta_1, \ldots, \delta_m$ are all i-objective. We can also assume that each of $\alpha \wedge \beta_i$, $i = 1, \ldots, k$ and $\gamma \wedge \delta_j$, $j = 1, \ldots, m$ are consistent, for otherwise we could easily show that φ is not consistent. Finally, we can show that $\alpha \vee \gamma$ must be provable, for if not, by applying $\mathbf{A5}'_n$, we can again show that φ is not consistent. We now apply the induction hypothesis to prove the result. ∎

Corollary 5.2: *Each formula in $\mathcal{ONL}_n^+$ is provably equivalent to a formula in $\mathcal{ONL}_n$.*

Proof: We proceed by induction on the structure of formulas. The only nontrivial case is for formulas of the form *Val*(φ). By the induction hypothesis, φ is provably equivalent to a formula $\varphi' \in \mathcal{ONL}_n$. By straightforward modal reasoning using **V1** and $\mathbf{Nec}_V$, we can show that *Val*(φ) is provably equivalent to *Val*(φ'). By Lemma 5.1, *Val*(φ') is provably equivalent to either *true* or *false*, depending on whether φ' is provable. ∎

Using standard modal logic techniques, we can now prove the following result.

Theorem 5.3: *M^e is a $K45_n$ structure (that is, $\mathcal{K}_i^e$ and $\mathcal{N}_i^e$ are Euclidean and transitive). Moreover, for each world $w \in W^e$, we have $(M^e, w) \models \alpha$ iff $\alpha \in w$.*

Proof: We leave it to the reader to check that the definition of $\mathcal{K}_i^e$ guarantees that M^e is a K45$_n$ structure. Given Corollary 5.2, which allows us to restrict attention to $\alpha \in \mathcal{ONL}_n$, the proof that $(M^e, w) \models \alpha$ iff $\alpha \in w$ is completely straight forward and follows the same lines as the usual proofs dealing with canonical models (see, for example, (Chellas 1980; Halpern and Moses 1992)). ∎

We say that α is *e-valid*, denoted $\models^e \alpha$, if $M^e \models \alpha$, that is, if $(M^e, w) \models \alpha$ for all worlds $w \in W^e$. The following result is immediate from Theorem 5.3.

Corollary 5.4: *$\models^e \alpha$ iff $AX' \vdash \alpha$.*

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

Proof: It is easy to see that each axiom of AX is sound in AX$'$. It follows that AX $\vdash \varphi$ implies AX$' \vdash \varphi$. For the converse, it suffices to show that if $\varphi \in \mathcal{ONL}_n^-$ is consistent with AX, then it is also consistent with AX$'$, i.e., that *Sat*(φ) holds. We show this by induction on the structure of φ, much in the same way we proved Proposition 5.1. We can assume without loss of generality that φ is a conjunction in the normal form described Lemma 3.12. It is easy to see that if we can deal with the case that φ is an i-subjective formula, then we can deal with arbitrary φ by repeated applications of **V4** followed by an application of **V2**. Thus, suppose that φ is an i-subjective formula which is consistent with AX. We can assume that φ is of the form $L_i\alpha \wedge \neg L_i \neg \beta_1 \wedge \ldots \wedge \neg L_i \neg \beta_k \wedge N_i \gamma \wedge \neg N_i \neg \delta_1 \wedge \ldots \wedge \neg N_i \neg \delta_m$. We must have that $\alpha \wedge \beta_j$ is consistent for $j = 1, \ldots, k$, and that $\gamma \wedge \delta_l$ is AX-consistent for $l = 1, \ldots, m$, for otherwise φ would not be AX-consistent. Similarly, by Lemma 3.14, we must have that $\alpha \vee \gamma$ is K45$_n$-provable, otherwise φ would not be AX-consistent. We can now apply **V3** and the inductive hypothesis to show that α is AX$'$-consistent. ∎

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 K45$_n$ alone. It is PSPACE-complete.

Theorem 5.6: *The problem of deciding if $AX' \vdash \varphi$ is PSPACE-complete.*

Proof: PSPACE hardness follows from the PSPACE hardness of K45$_n$ (Halpern and Moses 1992).[9] Wo sketch the proof of the upper bound. First of all, observe that it suffices to deal with the case that φ is in $\mathcal{ONL}_n$, since we can then apply the arguments of Corollary 5.2 to remove all occurrences of *Val* from inside out. We consider the dual problem of consistency. Thus, we want to check if *Sat*(α) holds. The first step is to convert α to the normal form of Lemma 3.12. Observe that α is consistent iff at least one of the disjuncts is consistent. Although the conversion to normal form may result in exponentially many disjuncts, each one is no longer than α. Thus, we deal with them one by one, without ever writing down the full disjunction. It suffices to show that we can decide if each disjunct is consistent in polynomial space, since we can then erase all the work and start over for the next disjunct (with a little space necessary for bookkeeping).

[9]The result in (Halpern and Moses 1992) is proved only for KD45$_n$, but the same proof applies to K45$_n$.

We now proceed much as in the proof of Proposition 5.1. By applying **V4** repeatedly and then **V2** (as in the previous theorem), it suffices to deal with i-subjective formulas. We then apply **V3** to get simpler formulas, and repeat the procedure. We remark that this gives another PSPACE decision procedure for K45$_n$, quite different from that presented in (Halpern and Moses 1992). ∎

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 arise as the set of possible states at some situation (M^e, w). A proof analogous to that of Lemma 3.7 shows that $Obj_i^e(M, w)$ is always limit closed. Although we do get limit closure, roughly speaking, we avoid problems by having in a precise sense "enough" possibilities.

6 Multi-Agent Nonmonotonic Reasoning

In this section, we demonstrate that the logic developed in Section 5 captures multi-agent autoepistemic reasoning in a reasonable way. We do this in two ways. First we show by example that the logic can be used to derive some reasonable nonmonotonic inferences in a multi-agent context. We then show that the logic can be used to extend the definitions of stable sets and stable expansions originally developed for single agent autoepistemic logic to the multi-agent setting.

6.1 Formal Derivations of Nonmonotonic Inferences

In this section, we provide two examples of how the logic can be used for nonmonotonic reasoning.

Example 6.1: Let p be agent i's secret and suppose i makes the following assumption: unless I know that j knows my secret assume that j does not know it. We can prove that if this assumption is all i believes then he indeed believes that j does not know his secret. Formally, we can show

$$\vdash O_i(\neg L_i L_j p \Rightarrow \neg L_j p) \Rightarrow L_i \neg L_j p.^{10}$$

[10] Note that if we replace $L_j p$ by p we obtain regular single-agent autoepistemic reasoning.

A formal derivation of this theorem can be obtained as follows. Let $\alpha = \neg L_i L_j p \Rightarrow \neg L_j p$. The justifications in the following derivation indicate which axioms or previous derivations have been used to derive the current line. PL or K45$_n$ indicate that reasoning in either standard propositional logic or K45$_n$, which are subsumed by AX$'$, is used without further analysis.

1.	$O_i\alpha \Rightarrow L_i\alpha$	PL
2.	$O_i\alpha \Rightarrow N_i\neg\alpha$	PL
3.	$(L_i\alpha \wedge \neg L_i L_j p) \Rightarrow L_i \neg L_j p$	K45$_n$
4.	$N_i\neg\alpha \Rightarrow (N_i \neg L_i L_j p \wedge N_i L_j p)$	K45$_n$
5.	$\mathit{Sat}(p)$	**V2**
6.	$\mathit{Sat}(p) \Rightarrow \mathit{Sat}(\neg L_j p)$	**V3**
7.	$\mathit{Sat}(\neg L_j p)$	**V1**, PL
8.	$\mathit{Sat}(\neg L_j p) \Rightarrow (N_i L_j p \Rightarrow \neg L_i L_j p)$	$\mathbf{A5}'_n$
9.	$N_i L_j p \Rightarrow \neg L_i L_j p$	*PL*
10.	$O_i\alpha \Rightarrow \neg L_i L_j p$	2; 4; 9; PL
11.	$O_i\alpha \Rightarrow L_i \neg L_j p$	1; 3; 10; PL

To see that i's beliefs may evolve nonmonotonically given that i knows only α, assume that i finds out that j has found out about the secret. Then i's belief that j does not believe the secret will be retracted. In fact, i will believe that j does believe the secret. Formally, we can show

$$\vdash O_i(L_j p \wedge \alpha) \Rightarrow L_i L_j p.$$

Notice that the logic itself is a regular monotonic logic; the nonmonotonicity of agent i's beliefs is hidden within the O_i-operator.

All the formulas that appear in the proof above are in $\mathcal{ONL}_n^-$. Thus, we could have used the somewhat simpler proof theory of Section 3.1. To obtain an example where we need the full power of AX$'$, simply replace $L_j p$ by $O_j p$, that is i now uses the default that unless he knows that j only knows p, then he assumes that j does not only know p. In other words, i (prudently) makes rather cautious assumptions about i's epistemic state and assumes that i usually knows more than just p. The proof is very similar to the one above. The only difference is that we now have to establish that $\mathit{Sat}(\neg O_j p)$ is provable, which is straightforward. ∎

Example 6.2: Now let p stand for "Tweety flies".[11] We want to show that if j knows that all i knows about Tweety is that by default it flies, then j knows that i believes that Tweety flies. As before, we capture the fact that all i believes is that, by default, Tweety flies, by saying that all i believes is that, unless i believes that Tweety does not fly, then Tweety flies. Thus, we want to show

$$\vdash L_j O_i(\neg L_i \neg p \Rightarrow p) \Rightarrow L_j L_i p.$$

[11] Regular readers of papers on nonmonotonic logic will no doubt be gratified to see Tweety's reappearance.

We proceed as follows:

1.	$O_i(\neg L_i \neg p \Rightarrow p) \Rightarrow L_i p$	as above, with $L_j p$ replaced by $\neg p$
2.	$L_j(O_i(\neg L_i \neg p \Rightarrow p) \Rightarrow L_i p)$	1; $\mathbf{Nec}_n$
3.	$L_j O_i(\neg L_i \neg p \Rightarrow p) \Rightarrow L_j L_i p$	2; K45$_n$

In a sense, i is able to reason about j's ability to reason nonmonotonically essentially by simulating j's reasoning pattern. ∎

A situation where i knows that *all* j knows is α seems hardly attainable in practice, since an agent usually has at best incomplete information about another agent's beliefs. It would seem much more reasonable if we could say that i knows that α is all j knows about *some relevant subject*, say Tweety. This issue is dealt with in (Lakemeyer 1993b), where the canonical-model approach is extended to allow statements of the kind that all agent i knows about x is y. It is shown that the forms of nonmonotonic reasoning just described, when restricted to $\mathcal{ONL}_n^-$, go through just as well with the weaker notion of only knowing about.

6.2 i-Stable Sets and i-Stable Expansions

Single-agent autoepistemic logic was developed by Moore (1985) using the concepts of *stable sets* and *stable expansions*. Levesque proved that there is a close relationship between stable sets and only-knowing in the single-agent case. Here we prove an analogous relationship for the multi-agent case. We first need to define a multi-agent analogue of stable sets.

In the single-agent case, it is well known that a stable set is a complete set of formulas that agent i could know in some situation; that is, a set S is stable if and only if there is a situation (W, w) such that $S = \{\alpha \mid (W, w) \models L\alpha\}$. This is the intuition that we want to extend to the multi-agent setting, where the underlying language is now $\mathcal{ONL}_n$. First we define logical consequence in the extended canonical model in the usual way: If Γ is a set of formulas, we write $M^e \models \Gamma$ if $M^e \models \gamma'$ for each formula $\gamma' \in \Gamma$. We say that γ is an *e-consequence* of Γ, and write $\Gamma \models^e \gamma$, exactly if $M^e \models \Gamma$ implies $M^e \models \gamma$.

Definition 6.3: Let Γ be a set of formulas in $\mathcal{ONL}_n$. Γ is called *i-stable* iff

(a) if $\Gamma \models^e \gamma$ then $\gamma \in \Gamma$,

(b) if $\alpha \in \Gamma$ then $L_i\alpha \in \Gamma$,

(c) if $\alpha \notin \Gamma$ then $\neg L_i\alpha \in \Gamma$. ∎

Note that the only difference between i-stable sets and the original definition of stable sets is in condition 1, which requires i-stable sets to be closed under e-consequence instead of tautological consequence (i.e., logical consequence in propositional logic) as in the single-agent case. Using e-consequence rather than tautological consequence makes no difference in the single-agent case; in the multi-agent case it does. Intuitively, we want to allow the agents to use e-consequence here to capture the intuition that it is common knowledge that all agents are perfect reasoners under the extended canonical model semantics. For example, if agent i believes $\neg L_j p$ for a different agent j, then we want him to also believe $N_j \neg L_j p$. To do this, we need to close off under e-consequence.

The next theorem shows that i-stable sets do indeed satisfy the intuitive requirement. We define an *i-epistemic* state to be a set Γ of $\mathcal{ONL}_n$-formulas such that for some situation (M^e, w) in the extended canonical model, $\Gamma = \{\alpha \in \mathcal{ONL}_n \mid (M^e, w) \models L_i\alpha\}$; in this case, we say that Γ is the i-epistemic situation corresponding to (M^e, w). For a set of $\mathcal{ONL}_n$-formulas Γ, let $\overline{\Gamma} = \{\gamma \mid \gamma \in \mathcal{ONL}_n \text{ and } \gamma \notin \Gamma\}$, $L_i\Gamma = \{L_i\gamma \mid \gamma \in \Gamma\}$, and $\neg L_i\overline{\Gamma} = \{\neg L_i\gamma \mid \gamma \in \overline{\Gamma}\}$.

Theorem 6.4: *Let Γ be a set of $\mathcal{ONL}_n$-formulas. Γ is i-stable iff Γ is an i-epistemic state.*

Proof: It is straightforward to show that every i-epistemic state is i-stable. To show the converse, let Γ be i-stable. We need to show that it is also an i-epistemic state. Certainly $L_i\Gamma$ is consistent. If Γ contains all $\mathcal{ONL}_n$-formulas, that is, if agent i is inconsistent, then let (M^e, w) be a situation where $\mathcal{K}_i^e(w) = \emptyset$; such a situation clearly exists. Then Γ is the i-epistemic situation corresponding to (M^e, w). If Γ is a proper subset of the $\mathcal{ONL}_n$-formulas, then Γ must be consistent. (If Γ were inconsistent, by the first property of stable sets, Γ would contain all formulas.) In particular, this means that $p \wedge \neg p \notin \Gamma$ for a primitive proposition p. The second property of stable sets guarantees that $L_i\Gamma \subseteq \Gamma$, while the third guarantees that $\neg L_i(p \wedge \neg p) \in \Gamma$. Since Γ is consistent, so is $L_i\Gamma \cup \{\neg L_i(p \wedge \neg p)\}$. W^e consists of all the maximal consistent sets (with respect to AX$'$); thus, there must be some $w^* \in W^e$ that contains $L_i\Gamma \cup \{\neg L_i(p \wedge \neg p)\}$. We claim that Γ is the i-epistemic state corresponding to (M^e, w^*). Thus, we must show that $\varphi \in \Gamma$ iff $(M^e, w^*) \models L_i\varphi$. To see this, first suppose that $\varphi \in \Gamma$. Thus, $L_i\varphi \in L_i\Gamma$. By construction, w^* contains $L_i\Gamma$. By Theorem 5.3, we have that $(M^e, w^*) \models L_i\varphi$. On the other hand, if $\varphi \notin \Gamma$, then, since Γ is i-stable, we have that $\neg L_i\varphi \in \Gamma$. By the previous argument, it follows that $(M^e, w^*) \models L_i\neg L_i\varphi$. From $\mathbf{A4}_n$, it follows that $(M^e, w^*) \models \neg L_i\varphi$, and hence $(M^e, w^*) \not\models L_i\varphi$. This proves the claim. ∎

Moore (1985) defined the notion of a *stable expansion* of a set A of formulas in the single-agent case. Intuitively, a stable expansion of A is a stable set containing A all of whose formulas can be justified, given A and the formulas believed in that stable set. Further discussion and justification of the notion of stable expansion can be found in (Halpern 1994; Moore 1985). Rather than discussing this here, we go directly to our multi-agent generalization of Moore's notion.

Definition 6.5: Let A be a set of $\mathcal{ONL}_n$-formulas. Γ is called an *i-stable expansion of* A iff $\Gamma = \{\gamma \in \mathcal{ONL}_n \mid A \cup L_i\Gamma \cup \neg L_i\overline{\Gamma} \models^e \gamma\}$. ∎

It is easy to see that an i-stable expansion is an i-stable set. The definition of i-stable expansions looks exactly like Moore's definition of stable expansions except that we again use e-consequence instead of tautological consequence. As in the case of stable sets, this is necessary to capture the fact that it is common knowledge that all agents can do reasoning under the extended canonical model semantics.

We now generalize a result of Levesque's (1990) (who proved it for the single-agent case), showing that the i-stable expansions of a formula α correspond precisely to the different situations where i only knows α. We first need a lemma.

Lemma 6.6:

Let (M^e, w) be a situation with $\Sigma = \{L_i\gamma \mid (M^e, w) \models L_i\gamma\} \cup \{\neg L_i\gamma \mid (M^e, w) \models \neg L_i\gamma\}$. For any α,

there is an i-objective formula α^* such that

(a) $\Sigma \models^e \alpha \Leftrightarrow \alpha^*$

(b) $(M^e, w) \models (L_i\alpha \Leftrightarrow L_i\alpha^*) \wedge (N_i\alpha \Leftrightarrow N_i\alpha^*)$.

Proof: Given a formula φ, we say that a subformula $L_i\psi$ of φ occurs at *top level* if it is not in the scope of any modal operators. Let α^* be the result of replacing each top-level subformula of α of the form $L_i\gamma$ (resp., $N_i\gamma$) by *true* if $\Sigma \models^e L_i\gamma$ (resp., $\Sigma \models^e N_i\gamma$) and by *false* otherwise. Clearly α^* is i-objective. Moreover, a trivial argument by induction on the structure of α shows that $\Sigma \models \alpha \Leftrightarrow \alpha^*$: If α is a primitive proposition then $\alpha^* = \alpha$; if α is of the form $\alpha_1 \wedge \alpha_2$ or $\neg\alpha'$, then the result follows easily by the induction hypothesis; if α is of the form $L_j\beta$ for $j \neq i$, then $\alpha = \alpha^*$, since α has no top-level subformulas of the form $L_i\varphi$; finally, if α is of the form $L_i\beta$, then α^* is either *true* or *false*, depending on whether $L_i\beta$ is in Σ. Since either $L_i\beta$ or $\neg L_i\beta$ must be in Σ, the result is immediate in this case too. Part (b) follows immediately from part (a), since if $w' \approx_i w$, we must have $(M^e, w') \models \Sigma$, so $(M^e, w') \models \alpha \Leftrightarrow \alpha^*$. ∎

Theorem 6.7: *Let w be a world in the extended canonical model and let Γ be the i-epistemic state corresponding to (M^e, w). Then, for every $\mathcal{ONL}_n$-formula α, we have that $(M^e, w) \models O_i\alpha$ iff (a) Γ is an i-stable expansion of $\{\alpha\}$ and (b) $\mathcal{K}_i^e(w)$ and $\mathcal{N}_i^e(w)$ are disjoint.*

Proof: Let $\Sigma = L_i\Gamma \cup \neg L_i\overline{\Gamma}$. To prove the "only if" direction, suppose $(M^e, w) \models O_i\alpha$. The disjointness of $\mathcal{K}_i^e(w)$ and $\mathcal{N}_i^e(w)$ follows immediately from the fact that $(M^e, w) \models L_i\alpha \wedge N_i\neg\alpha$. To prove that Γ is an i-stable expansion of $\{\alpha\}$, it suffices to show that for all $\mathcal{ONL}_n$-formulas β, we have $(M^e, w) \models L_i\beta$ iff $\{\alpha\} \cup \Sigma \models^e \beta$.

First suppose that $\{\alpha\} \cup \Sigma \models^e \beta$. Since $(M^e, w) \models \{O_i\alpha\} \cup \Sigma$, and every formula in Σ is of the form $L_i\gamma$ or $\neg L_i\gamma$, it easily follows that $(M^e, w) \models L_i\alpha$ and $(M^e, w) \models L_i\Sigma$.

Hence, $(M^e, w') \models \alpha$ and $(M^e, w') \models \Sigma$ for every $w' \in \mathcal{K}_i^e(w)$. It follows that $(M^e, w') \models \beta$ and, therefore, $(M^e, w) \models L_i\beta$.

For the converse, suppose that $(M^e, w) \models L_i\beta$. We want to show that $\{\alpha\} \cup \Sigma \models^e \beta$. By Lemma 6.6, we can assume without loss of generality that β is i-objective. (For if not, we can replace β by an i-objective β^* such that $\Sigma \models^e \beta \Leftrightarrow \beta^*$ and $(M^e, w) \models L_i\beta \Leftrightarrow L_i\beta^*$, prove the result for β^*, and conclude that it holds for β as well.) To show that $\{\alpha\} \cup \Sigma \models^e \beta$, we must show that for all worlds w' such that $(M^e, w') \models \{\alpha\} \cup \Sigma$, we have $(M^e, w') \models \beta$. By Lemma 6.6, there is an i-objective formula α^* such that $(M^e, w) \models O\alpha \Leftrightarrow O\alpha^*$ and $\Sigma \models^e \alpha \Leftrightarrow \alpha^*$. Thus, $(M^e, w) \models N_i\alpha^* \wedge L_i\beta$. By the arguments of Lemma 3.14 (which apply without change to the extended canonical model semantics), it must be the case that $\models^e \alpha^* \Rightarrow \beta$. Since $\Sigma \models^e \alpha \Leftrightarrow \alpha^*$, it follows that $\{\alpha\} \cup \Sigma \models^e \beta$. Since $(M^e, w') \models \{\alpha\} \cup \Sigma$ by assumption, we have that $(M^e, w') \models \beta$, as desired.

To prove the "if" direction of the theorem, suppose that $\mathcal{K}_i^e(w)$ and $\mathcal{N}_i^e(w)$ are disjoint and that for all $\mathcal{ONL}_n$-formulas β, we have $(M^e, w) \models L_i\beta$ iff $\{\alpha\} \cup \Sigma \models^e \beta$. We need to show that $(M^e, w) \models O_i\alpha$, that is, $(M^e, w) \models L_i\alpha \wedge N_i\neg\alpha$.

Since $\{\alpha\} \cup \Sigma \models^e \alpha$, the fact that $(M^e, w) \models L_i\alpha$ follows immediately. To prove that $(M^e, w) \models N_i\neg\alpha$, let $w' \in \mathcal{N}_i^e(w)$ and assume, to the contrary, that $(M^e, w') \models \alpha$. Since $w' \in \mathcal{N}_i^e(w)$, it follows that $\Sigma \subseteq w'$. Hence $w/L_i \subseteq w'$, from which $w' \in \mathcal{K}_i^e(w)$ follows, contradicting the assumption that $\mathcal{K}_i^e(w)$ and $\mathcal{N}_i^e(w)$ are disjoint. ∎

7 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_ip, 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. Of course, as we suggested above, rather than only knowing, it seems more appropriate to reason about only knowing about a certain topic. Lakemeyer (1993b) provides a semantics for only knowing about, using the canonical-model approach. It would be interesting to see if this can also be done using the other approaches we have explored here.

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. San Francisco, Calif.: 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 (1984). Towards a theory of knowledge and ignorance. In *Proc. AAAI Workshop on Non-monotonic Logic*, pp. 125–143. Reprinted in K. R. Apt (Ed.), *Logics and Models of Concurrent Systems*, 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.

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

Lakemeyer, G. (1993b). All they know about. In *Proc. National Conference on Artificial Intelligence (AAAI '93)*, pp. 662–667.

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.