All They Know: A Study in Multi-Agent Autoepistemic Reasoning

PRELIMINARY REPORT

Gerhard Lakemeyer
Institute of Computer Science III
University of Bonn
Romerstr. 164
5300 Bonn 1, Germany
gerhard@cs.uni-bonn.de

Abstract

With few exceptions the study of nonmonotonic reasoning has been confined to the single-agent case. However, it has been recognized that intelligent agents often need to reason about other agents and their ability to reason nonmonotonically. In this paper we present a formalization of multi-agent autoepistemic reasoning, which naturally extends earlier work by Levesque. In particular, we propose an n-agent modal belief logic, which allows us to express that a formula (or finite set of them) is *all* an agent knows, which may include beliefs about what other agents believe. The paper presents a formal semantics of the logic in the possible-world framework. We provide an axiomatization, which is complete for a large fragment of the logic and sufficient to characterize interesting forms of multi-agent autoepistemic reasoning. We also extend the stable set and stable expansion ideas of single-agent autoepistemic logic to the multi-agent case.

1 Introduction

While the study of nonmonotonic reasoning formalisms has been at the forefront of foundational research in knowledge representation for quite some time, work in this area has concentrated on the single-agent case with only few exceptions.

This focus on single agents is somewhat surprising, since there is little doubt that agents, who have been invested with nonmonotonic reasoning mechanisms, should be able to reason about other agents and their ability to reason nonmonotonically as well. For example, if we assume the common default that birds normally fly and if Jill tells Jack that she has just bought a bird, then Jill should be able to infer that Jack thinks that her bird flies. Multi-agent nonmonotonic reasoning is also crucial when agents need to coordinate their activity. For example, assume I promised a friend (who is always on time) to meet him at a restaurant at 7PM. If I leave my house knowing that I will not make it there by 7PM, I will probably not change my plans and still go to the restaurant. After all, I know that my friend has no reason to believe that I am not on my way to meet him and that he will therefore wait for me. Note that I reason about my friends default assumption to wait in this case. Other examples from areas like planning and temporal projection can be found in [Mor90].

One of the main formalisms of nonmonotonic reasoning is *autoepistemic logic* (e.g. [Moo85]). The basic idea is that the beliefs of agents are closed under *perfect introspection,* that is, they know[1] what they know and do not know. Nonmonotonic reasoning comes about in this framework in that agents can draw inferences on the basis of their own ignorance. The following example by Moore illustrates this feature: I can reasonably conclude that I have no older brother simply because I do not *know* of any older brother of mine. A particular formalization of autoepistemic reasoning is due to Levesque [Lev90], who proposes a logic of *only-knowing {OL),* which is a classical modal logic extended by a new modality to express that a formula is *all* an agent believes. An advantage of this approach is that, rather than having to appeal to non-standard inference rules as in Moore's original formulation, *OL* captures autoepistemic reasoning using only the classical notions of logical consequence and theoremhood.

In this paper, we propose a propositional multi-agent extension of *OL.* We provide a formal semantics within the possible-world framework and a proof theory, which is complete for a large fragment of the logic. The new logic also leads us to natural extensions of notions like *stable sets* and *stable expansions,* which were originally developed for the single-agent case.

General multi-agent nonmonotonic reasoning formalisms have received very little attention until recently.[2] A notable exception is work by Morgenstern and Guerreiro [Mor90, MG92], who consider both multi-agent autoepistemic reasoning and multi-agent circumscription theories. On the autoepistemic side, they propose multi-agent versions of stable sets, which al-

[1] While we are concerned with belief and, in particular, allow agents to have false beliefs, we nevertheless use the terms knowledge and belief interchangeably.

[2] There has also been work applying nonmonotonic theories to special multi-agent settings such as speech acts (e.g. [Per87, AK88]). Unlike our work, these approaches are not concerned with general purpose multi-agent nonmonotonic reasoning.

low agents to reason about other agents' nonmonotonic inferences. In contrast to our work, however, these stable sets are not justified by an independent semantic account. Recently, and independently of our work, Halpern [Hal93] also extended Levesque's logic *OL* to the multi-agent case. While Halpern's logic and ours share many (but not all) properties, the respective model theories are quite different. In particular, while our approach remains within classical possible-world semantics, Halpern uses concepts very much related to the so-called *knowledge structures* of [FHV91]. Using the same technique, Halpern also extends the notion of only-knowing proposed in [HM84] to the multi-agent case. There, however, agents are not capable of reasoning about other agent's nonmonotonic inferences because only-knowing is used only as a meta-logical concept.

The rest of the paper is organized as follows. Section 2 defines the logic *OLn,* extending Levesque's logic of only-knowing to many agents. Besides a formal semantics we also provide a proof theory, which is complete for a large fragment of *0Ln.* Furthermore, we look at the properties of formulas in *0Ln* which uniquely determine the beliefs of an agent and are thus of particular interest to knowledge representation. Section 3 considers examples of multi-agent autoepistemic reasoning as modeled by *0Ln* Section 4 shows how *0Ln* yields natural multi-agent versions of *stable sets* and *stable expansions.* Finally, we summarize the results of the paper and point to some future work in Section 5.

2 The Logic *0Ln*

After introducing the syntax of the logic, we define the semantics in two stages. First we describe that part of the semantics that does not deal with only-knowing. In fact, this is just an ordinary possible-world semantics for n agents with perfect introspection. Then we introduce the necessary extensions that give us the semantics of only-knowing. Finally, we present a proof theoretic account, which is complete for a large fragment of the logic, and discuss properties of the logic which are important in the context of knowledge representation.

2.1 Syntax

Definition 1 *The Language $\mathcal{OL}_n$*
The primitives of $\mathcal{OL}_n$ consist of a countably infinite set of atomic propositions (or atoms), the connectives $\vee$, $\neg$, *and the modal operators* $\mathbf{L}_i$ *and* $\mathbf{O}_i$ *for* $1 \leq i \leq n$. *(Agents are referred to as* $1, 2, \ldots, n$.*) Formulas are formed in the usual way from these primitives.*[3] $\mathbf{L}_i\alpha$ *should be read as "the agent i believes* α*" and* $\mathbf{O}_i\alpha$ *as "*α *is all agent i believes." A formula* α *is called* **basic** *iff there are no occurrences of* $\mathbf{O}_i$ $(1 \leq i \leq n)$ *in* α.

Definition 2 *A modal operator occurs at* **depth** n *of a formula* α *iff it occurs within the scope of exactly* n *modal operators.*

[3] We will freely use other connectives like $\wedge$, $\supset$ and $\equiv$, which should be understood as syntactic abbreviations of the usual kind.

For example, given $\alpha = p \wedge \mathbf{L}_1\mathbf{L}_2(\mathbf{L}_3 q \vee \neg\mathbf{O}_2 r)$, $\mathbf{L}_1$ occurs at depth 0, $\mathbf{L}_2$ at depth 1, and $\mathbf{L}_3$ and $\mathbf{O}_2$ occur both at depth 2.

Definition 3 *A formula* α *is called* **i-objective** *(for* $i = 1, \ldots n$*) iff every modal operator at depth 0 is of the form* $\mathbf{O}_j$ *or* $\mathbf{L}_j$ *with* $i \neq j$.

In other words, i-objective formulas talk about the external world from agent i's point of view, which includes beliefs of other agents but not his own. For example, $(p \vee \mathbf{L}_2 q) \wedge \neg\mathbf{O}_3\mathbf{L}_1 p$ is 1-objective, but $(p \vee \mathbf{L}_2 q) \wedge \neg\mathbf{L}_1\mathbf{O}_3 p$ is not.

Definition 4 *A formula* α *is called* **i-subjective** *iff every modality at depth 0 is of the form* $\mathbf{L}_i$ *or* $\mathbf{O}_i$ *and every atom occurs within the scope of a modal operator.*

That is, i-subjective formulas talk only about what agent i believes. For example, $\mathbf{L}_1 p \wedge \neg\mathbf{O}_1 p$ is 1-subjective, $\mathbf{L}_1 p \wedge q$ and $\mathbf{L}_1 p \wedge \mathbf{L}_2 q$ are not.

2.2 The Semantics of Basic Formulas

Basic formulas are given a standard possible-world semantics [Kri63, Hin62, Hin71], which the reader is assumed to be familiar with.[4] Roughly, a *possible-world model* consists of *worlds*, which determine the truth of atomic propositions, and binary *accessibility* relations between worlds. An agent's beliefs at a given world w are determined by what is true in all those worlds that are accessible to the agent from w. Since we are concerned with agents who possess *perfect introspection*, we restrict the accessibility relations in the usual way, that is, they are both transitive and Euclidean.[5] The resulting logic is called $K45_n$ or *weak* $S5_n$.[6]

Definition 5 *A $K45_n$-Model*
$M = (W, \pi, R_1, \ldots, R_n)$ *is called a* $K45_n$**-model** *(or simply* **model***) iff*

1. W *is a set (of worlds).*
2. π *is a mapping from the set of atoms into* 2^W.
3. $R_i \subseteq W \times W$ *for* $1 \leq i \leq n$.
4. R_i *is transitive and Euclidean for* $1 \leq i \leq n$.

Given a model $M = (W, \pi, R_1, \ldots, R_n)$ and a world $w \in W$, the possible-world semantics of basic formulas is defined as follows: Let p be an atom and α and β arbitrary basic formulas.

$w \models p \iff w \in \pi(p)$
$w \models \neg\alpha \iff w \not\models \alpha$
$w \models \alpha \vee \beta \iff w \models \alpha$ or $w \models \beta$
$w \models \mathbf{L}_i\alpha \iff$ for all w', if wR_iw' then $w' \models \alpha$

A set of basic formulas Γ is *satisfiable* iff there is a model $M = (W, \pi, R_1, \ldots, R_n)$ and $w \in W$ such that $w \models \gamma$ for all $\gamma \in \Gamma$.

In the single-agent case, Levesque made use of the fact that the semantics of $K45_1$ has a much simpler formulation. There we can assume a fixed set of worlds, which is the set of all truth assignments. A model then consists

[4] See [HC84, HM92] for an introduction.

[5] R_i is Euclidean iff $\forall w, w', w''$, if wR_iw' and wR_iw'', then $w'R_iw''$.

[6] The subscript n indicates that we are concerned with the n-agent case.

of a world w, which corresponds intuitively to the real world, and a set of worlds W', which determine the beliefs of the agent. There is no need for an explicit accessibility relation, since the worlds in M are globally accessible from every world and a sentence is believed just in case it is true at all worlds in W. Unfortunately, such a simple model does not extend to the multi-agent case and we are forced to a more complicated semantics with explicit accessibility relations as defined above.7 For this reason, the extension of Levesque's logic OL to many agents turns out to be a non-trivial exercise.

2.3 The Canonical Model

It is well known that, as far as basic formulas are concerned, it suffices to look at just one, the so-called canonical model [HC84, HM92]. This canonical model will be used later on to define the semantics of only-knowing.

The central idea behind canonical models are maximally consistent sets.

Definition 6 Maximally consistent sets
Given any proof theory of K45n and the usual notion of theoremhood and consistency, a set of basic formulas T is called maximally consistent iff T is consistent and for every basic α, either α or $\neg\alpha$ is contained in T.

The canonical K45n-model Mc has as worlds precisely all the maximally consistent sets and a world w' is inaccessible from w just in case all of i's beliefs at w are included in w'.

Definition 7 The Canonical K45n-Model Mc
The canonical model Mc = (Wc, Tl, R1, . . . , Rn) is a Kripke structure such that

1. $W_c = \{w \mid w \textit{ is a maximally consistent set}\}$.
2. *For all atoms p and* $w \in W_c$, $w \in \pi(p)$ *iff* $p \in w$.
3. wR_iw' *iff for all formulas* $\mathbf{L}_i\alpha$, *if* $\mathbf{L}_i\alpha \in w$ *then* $\alpha \in w'$.

The following (well known) theorem tells us that nothing is lost from a logical point of view if we confine our attention to the canonical model.

Theorem 1 Mc is a $K45_n$-model and for every set of basic formulas T, T is satisfiable iff it is satisfiable in Mc.

2.4 The Semantics of All They Know

Given this classical possible-world framework, what does it mean for an agent i to only-know, say, an atom p at some world w in a model Ml Certainly, i should believe p, that is, all worlds that are i-accessible from w should make p true. Furthermore, i should believe as little else as possible apart from p. For example, i should neither believe q nor believe that j believes p etc. Minimizing knowledge using possible worlds simply means maximizing the number of accessible worlds. Thus, in our example, there should be an accessible world where q is false and another one where j does not believe p and so on. It should be clear that in order for w to satisfy only-knowing α this way, the model M must have a huge supply of worlds that are accessible from w. While not essential for the definition of only-knowing, it turns out to be very convenient to simply restrict our attention to models that are guaranteed to contain a sufficient supply of worlds.[8] In fact, we will consider just one, namely the *canonical model* of $K45_n$. Let us call the set of all formulas that are true at some world w in some model of $K45_n$ a *world state*. The canonical model has the nice property that it contains precisely one world for every possible world state, since world states are just maximally consistent sets.

With that agent i is said to only-know a formula α at some world w (in the canonical model) just in case α is believed and any world w' which satisfies α and from which the same worlds are i-accessible as from w is itself i-accessible from w. We now turn to the formal definitions.

Definition 8 *Given a model* $M = (W, \pi, R_1, \ldots, R_n)$ *and worlds* w *and* w' *in* W, *we say that* w *and* w' *are* **i-equivalent** ($w \approx_i w'$) *iff for all worlds* $w^* \in W$, wR_iw^* *iff* $w'R_iw^*$.

Given an arbitrary formula α of $\mathcal{OL}_n$, a world w in a model M, let

$$w \models \mathbf{O}_i\alpha \iff \text{for all } w' \text{ s.t. } w \approx_i w', wR_iw' \text{ iff } w' \models \alpha.$$

A formula α of $\mathcal{OL}_n$ is a logical consequence of a set of formulas Γ iff for all worlds w in the canonical model M_c, if $w \models \gamma$ for all $\gamma \in \Gamma$, then $w \models \alpha$. As usual, we say that α is valid ($\models \alpha$) iff $\{\} \models \alpha$. A formula α is satisfiable iff $\neg\alpha$ is not valid.

Note that Theorem 1 guarantees that OL_n restricted to basic formulas is still simply $K45_n$.

A superficial comparison of the rules for $\mathbf{O}_i$ and $\mathbf{L}_i$ suggests that they differ at two places. For one the $\mathbf{O}_i$-rule quantifies over i-equivalent worlds only. For another, the "if ... then" of the $\mathbf{L}_i$-rule is replaced by an "iff." While the latter is significant, it can be easily shown that we can restrict the $\mathbf{L}_i$-rule to quantify over i-equivalent worlds as well.

Lemma 2.1
For any $K45_n$*-model* M *and world* w *in* M,

$$w \models \mathbf{L}_i\alpha \iff \text{for all } w' \text{ s.t. } w \approx_i w', \text{if } wR_iw' \text{ then } w' \models \alpha$$

2.5 A Proof Theory

In order to obtain a proof theory for a large fragment of OL_n, we apply the same idea Levesque used to axiomatize his logic OL. The idea is to add a dual operator $\mathbf{N}_i$ for every $\mathbf{L}_i$ to the language $\mathcal{OL}_n$, which we refer to as $\mathcal{ONL}_n$. While $\mathbf{L}_i\alpha$ can be read as "agent i knows *at least* α," $\mathbf{N}_i\alpha$ should be read as "agent i knows *at most* that α is false." $\mathbf{O}_i\alpha$ now becomes simply an abbreviation for $\mathbf{L}_i\alpha \wedge \mathbf{N}_i\neg\alpha$, that is, agent i believes only α iff he believes at least and at most α. We extend the notion of an *i-subjective* or *i-objective* formula to cover the $\mathbf{N}_i$-operators as well. For example, $\neg\mathbf{L}_1p \wedge \mathbf{N}_1\neg\mathbf{L}_2p$ is both 1-subjective and 2-objective.

7In essence, if we have more than 1 agent and a global set of worlds for each agent, the agents would be mutually introspective, which is not what we want.

[8] In Levesque's OL, this is automatically given since worlds are drawn from the fixed set of all truth assignments.

The semantics of $\mathbf{N}_i$ is as follows. $\mathbf{N}_i\alpha$ is true at a world w iff α is true at all i-equivalent worlds of w which are *not* accessible from w. In the formalization it is convenient to add new accessibility relations $\overline{R}_i$ for just these non-accessible worlds to the canonical $K45_n$-model. This way, the $\mathbf{N}_i$ behave just like ordinary $K45$ belief operators.

Definition 9 *The Extended Canonical Model*
Given the canonical model $M_c = (W_c, \pi, R_1, \ldots, R_n)$, let the extended canonical model be

$$M_{ec} = (W_c, \pi, R_1, \ldots, R_n, \overline{R}_1, \ldots, \overline{R}_n),$$

where for all $w, w' \in W_c$, $w\overline{R}_i w'$ iff $w \approx_i w'$ and $w \not R_i w'$.

Lemma 2.2 *The $\overline{R}_i$ are transitive and Euclidean.*

Given the extended canonical model M_{ec} and a world $w \in W_c$, the semantic rule for $\mathbf{N}_i$ is simply:

$$w \models \mathbf{N}_i\alpha \iff \text{for all } w', \text{ if } w\overline{R}_i w' \text{ then } w' \models \alpha.$$

Notions like logical consequence and validity for this extended logic are defined as for OL_n with M_c replaced by M_{ec}.

Given Lemma 2.2, it is obvious that the $\mathbf{N}_i$ have all the properties of a $K45_n$-operator. Moreover, if we view $\mathbf{L}_i$ and $\mathbf{N}_i$ as two different agents, then the $\mathbf{L}_i$-agent knows exactly which worlds the $\mathbf{N}_i$-agent can see and vice versa. In other words, the $\mathbf{L}_i$-agent and $\mathbf{N}_i$-agent are mutually introspective. For example $\neg\mathbf{L}_i p \supset \mathbf{N}_i\neg\mathbf{L}_i p$ is valid. Finally, note that for every equivalence class of worlds W_i under $\approx_i$, R_i and $\overline{R}_i$ cover all worlds in W_i exhaustively, that is, taken together, the two relations form a complete subgraph over W_i. This feature is reflected in valid formulas of the form $\mathbf{N}_i\alpha \supset \neg\mathbf{L}_i\alpha$ for falsifiable α.

Axioms:

A1	Axioms of propositional logic
A2	$\mathbf{L}_i(\alpha \supset \beta) \supset (\mathbf{L}_i\alpha \supset \mathbf{L}_i\beta)$
A3	$\mathbf{N}_i(\alpha \supset \beta) \supset (\mathbf{N}_i\alpha \supset \mathbf{N}_i\beta)$
A4	$\sigma \supset \mathbf{L}_i\sigma \wedge \mathbf{N}_i\sigma$ for all i-subjective σ
A5	$\mathbf{N}_i\alpha \supset \neg\mathbf{L}_i\alpha$ for all basic i-obj. α falsif. in $K45_n$
A6	$\mathbf{O}_i\alpha \equiv (\mathbf{L}_i\alpha \wedge \mathbf{N}_i\neg\alpha)$ for all α

Inference Rules:

MP	From α and $\alpha \supset \beta$ infer β.
Nec	From α infer $\mathbf{L}_i\alpha$ and $\mathbf{N}_i\alpha$.

Note that **A2–A4** imply that the $\mathbf{L}_i$ and $\mathbf{N}_i$ have all the properties of regular $K45_n$-operators. Also, **A4** not only gives us the regular introspection axioms of $K45_n$ but also cross axoms (mutual introspection) such as $\neg\mathbf{L}_i\alpha \supset \mathbf{N}_i\neg\mathbf{L}_i\alpha$. **A5** provides the crucial link between $\mathbf{L}_i$ and $\mathbf{N}_i$ capturing the relationship between "knowing at least" and "knowing at most."[9]

Theorem 2 *Soundness*
Given the usual definition of theoremhood ($\vdash$) with respect to the above proof theory, for all α in $\mathcal{ONL}_n$, if $\vdash\alpha$ then $\models\alpha$.

[9] Levesque, in his axiomatization of OL, uses a special case of this axiom. In his case, α ranges merely over the falsifiable formulas of classical propositional logic (no modalities).

Notice that axiom **A5** assumes that α ranges only over *basic* i-objective formulas. We need this restriction in order to appeal to falsifiability in the existing logic $K45_n$.[10] For the axiomatization to be complete, such a restriction essentially requires that arbitrary formulas $\mathbf{N}_i\beta$ or $\mathbf{L}_i\beta$ are reducible to equivalent forms $\mathbf{N}_i\beta^*$ and $\mathbf{L}_i\beta^*$, where β^* is a basic formula, that is a formula without any $\mathbf{N}_j$'s. That, however, is not true in general. For example, $\mathbf{L}_i\mathbf{N}_j p$ and $\mathbf{N}_i\mathbf{L}_j\mathbf{N}_i p$ are not reducible for distinct i and j. However, if we rule out such cases, the proof theory is indeed complete for the restricted language, which we call $\mathcal{ONL}_n^-$.

Definition 10 *The Language $\mathcal{ONL}_n^-$*
α is a formula of $\mathcal{ONL}_n^-$ iff α is a formula of $\mathcal{ONL}_n$ and, after replacing every occurrence of $\mathbf{O}_i\beta$ within α by its definition $\mathbf{L}_i\beta \wedge \mathbf{N}_i\neg\beta$, no $\mathbf{N}_j$ may occur within the scope of an $\mathbf{N}_i$ or $\mathbf{L}_i$ for $i \neq j$.

For example, while $\mathbf{N}_i\mathbf{L}_i\neg\mathbf{N}_i p$ and $\mathbf{N}_i(\mathbf{L}_j p \vee \mathbf{N}_i\neg p)$ are in $\mathcal{ONL}_n^-$, $\mathbf{N}_i\mathbf{N}_j p$ and $\mathbf{N}_i\mathbf{L}_j\mathbf{N}_i p$ are not for distinct i and j.

Theorem 3 *Completeness For $\mathcal{ONL}_n^-$*
For all $\alpha \in \mathcal{ONL}_n^-$, if $\models\alpha$ then $\vdash\alpha$.

Proof: The proof is an adaptation of Levesque's completeness proof for OL [Lev90]. ■

(Halpern recently proved that the above proof theory is indeed *not* complete for all of OL_n.)

Given Theorem 2 and 3, it is not hard to prove that OL_1 is equivalent to the propositional version of Levesque's logic OL because Levesque's proof theory is a special case of ours.

Theorem 4 *OL_1 and OL are equivalent*
A formula α in $\mathcal{OL}_1$ is a theorem of OL_1 iff α is a theorem of OL.

We introduced the operator $\mathbf{N}_i$ mainly for the purpose of obtaining a proof theory for OL_n. Except for a formal derivation using the axioms in Section 3, we will no longer be concerned with the $\mathbf{N}_i$-operators in the rest of the paper and go back to the original OL_n and its language $\mathcal{OL}_n$.

2.6 i-Determinate Sentences

We now turn to a class of formulas which are of particular interest to knowledge representation, since they determine, when only-known, precisely what an agent believes and does not believe.

Definition 11 *i-Determinate Formulas*[11]
A formula $\alpha \in \mathcal{OL}_n$ is **i-determinate** *iff*

1. *there is a world w in M_c such that $w \models \mathbf{O}_i\alpha$ and*
2. *for all worlds w', if $w' \models \mathbf{O}_i\alpha$ then $w \approx_i w'$.*

The following theorem demonstrates that i-determinate formulas indeed deserve their name.

Theorem 5 *For all $\alpha \in \mathcal{OL}_n$, α is i-determinate iff for all β, exactly one of $\mathbf{O}_i\alpha \supset \mathbf{L}_i\beta$ and $\mathbf{O}_i\alpha \supset \neg\mathbf{L}_i\beta$ is valid.*

[10] Note that this peculiar axiom schema is recursive since falsifiability in propositional $K45_n$ is decidable.

[11] In the single-agent case, the analogous concept of *determinate* formulas was defined in [Lev90].

What are examples of determinate formulas? In the single-agent case, it has been shown that all objective formulas (no modalities at all) are determinate. Not surprisingly, objective formulas are also i-determinate. In fact, this result can be generalized to include all basic i-objective formulas.

Theorem 6
All basic i-objective formulas are i-*determinate.*

Other examples of i-determinate formulas, which are not i-objective, include $\neg \mathbf{L}_i \neg p \supset p$ and $\neg \mathbf{L}_i \mathbf{L}_j p \supset \neg \mathbf{L}_j p$, which allow agents to reason nonmonotonically and are discussed in more detail in Section 3.

So far we have only considered basic i-determinate sentences. Does the above result extend to non-basic i-objective formulas as well? The answer, surprisingly, is: sometimes but not always! For example, it is not hard to show that the formula $\mathbf{O}_j p$ (where *p* is an atom) is also i-determinate, that is, the beliefs of agent *i* are uniquely determined if all *i* knows is that all *j* knows is *p*. However, the formula $\neg \mathbf{O}_j p$ is not i-determinate because $\models \neg \mathbf{O}_i \neg \mathbf{O}_j p$.[12] In other words, it is impossible for i to only-know that *j* does not only-know p. Intuitively, for *i* to know that *j* does not only-know p, *i* needs to have some evidence in terms of a basic belief or non-belief of *j*. It is because of properties like $\models \neg \mathbf{O}_i \neg \mathbf{O}_j p$ that our axiomatization is not complete for all of $\mathcal{ON}_n$

3 Multi-agent Nonmonotonic Reasoning

While our axiomatization is complete only for a subset of OL_n, it is nevertheless strong enough to model interesting cases of multi-agent nonmonotonic reasoning. Here are two examples:

1. Let *p* be agent i's secret and suppose *i* makes the following assumption: unless 1 know that *j* knows my secret assume that *j* does not know it. We can prove in OL_n that if this assumption is all *i* believes then he indeed believes that *j* does not know his secret. Formally $\vdash \mathbf{O}_i(\neg \mathbf{L}_i \mathbf{L}_j p \supset \neg \mathbf{L}_j p) \supset \mathbf{L}_i \neg \mathbf{L}_j p$.[13]

A formal derivation of this theorem of OL_n can be obtained as follows. Let $\alpha = \neg \mathbf{L}_i \mathbf{L}_j p \supset \neg \mathbf{L}_j p$. 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 $K45_n$ is used is without further analysis.

1	$\mathbf{O}_i \alpha \supset \mathbf{L}_i \alpha$	A6;*PL*
2	$\mathbf{O}_i \alpha \supset \mathbf{N}_i \neg \alpha$	A6;*PL*
3	$(\mathbf{L}_i \alpha \wedge \neg \mathbf{L}_i \mathbf{L}_j p) \supset \mathbf{L}_i \neg \mathbf{L}_j p$	$K45_n$
4	$\mathbf{N}_i \neg \alpha \supset (\mathbf{N}_i \neg \mathbf{L}_i \mathbf{L}_j p \wedge \mathbf{N}_i \mathbf{L}_j p)$	$K45_n$
5	$\mathbf{N}_i \mathbf{L}_j p \supset \neg \mathbf{L}_i \mathbf{L}_j p$	A5
6	$\mathbf{O}_i \alpha \supset \neg \mathbf{L}_i \mathbf{L}_j p$	2;4;5;*PL*
7	$\mathbf{O}_i \alpha \supset \mathbf{L}_i \neg \mathbf{L}_j p$	1;3;6;*PL*

To see that i's beliefs may evolve nonmonotonically given that *i* only-knows 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. Formally $\vdash \mathbf{O}_i(\mathbf{L}_j p \wedge (\neg \mathbf{L}_i \mathbf{L}_j p \supset \neg \mathbf{L}_j p)) \supset (\neg \mathbf{L}_i \neg \mathbf{L}_j p \wedge \mathbf{L}_i \mathbf{L}_j p)$. Notice that, while OL_n is itself a regular monotonic logic, the nonmonotonicity of agent *i*'s beliefs is hidden within the $\mathbf{O}_i$-operator.

Finally, $\neg \mathbf{L}_i \mathbf{L}_j p \supset \neg \mathbf{L}_j p$ is also i-determinate. In particular, $\mathbf{O}_i(\neg \mathbf{L}_i \mathbf{L}_j p \supset \neg \mathbf{L}_j p) \equiv \mathbf{O}_i \neg \mathbf{L}_j p$ is valid.

2. Now let *p* stand for the old "Tweety flies." As expected, we obtain $\vdash \mathbf{O}_j(\neg \mathbf{L}_j \neg p \supset p) \supset \mathbf{L}_j p$. (Not surprisingly, $\neg \mathbf{L}_j \neg p \supset p$ is j-determinate.) By the rule of necessitation and the distribution axiom for belief we immediately get $\vdash \mathbf{L}_i \mathbf{O}_j(\neg \mathbf{L}_j \neg p \supset p) \supset \mathbf{L}_i \mathbf{L}_j p$. In other words, *i* is able to reason about *j*'s ability to reason nonmonotonically, essentially by simulating *j*'s reasoning.

It should be pointed out that an assumption like *i* knows that *all j* knows is α is certainly unrealistic in general and is related to what Morgenstern calls *arrogance* [Mor90]. It would be much more reasonable if we could say that *i* knows that α is all *j* knows about *some relevant subject*, say Tweety. In fact, we have proposed such a notion of *only-knowing-about* as an extension of OL_n in [Lak93]. There we also show that only-knowing is often a good approximation of *only-knowing-about* in the sense that, if we restrict ourselves to beliefs that are only about the subject matter in question,[14] then the beliefs that follow from only-knowing-about are the same as those that follow from only-knowing.

4 i-Stable Sets and i-Stable Expansions

Single-agent autoepistemic logic was originally developed using the concepts of *stable sets* [Sta80] and *stable expansions* [Moo85]. Here we define natural n-agent extensions of these notions and show how they relate to OL_n. For the purposes of this paper, we confine ourselves to *basic* formulas only.

Definition 12 *i-Epistemic State*
A set of basic formulas Γ *is called an i-epistemic state iff there is a world* w *in* M_e *such that for all basic* γ, $w \models \mathbf{L}_i \gamma$ *iff* $\gamma \in \Gamma$.

Definition 13 *i-Stable Sets*
Let Γ *be a set of basic formulas.* Γ *is called i-stable iff*

1. Γ *contains all valid formulas of* $K45_n$.
2. *If* $\alpha \in \Gamma$ *and* $\alpha \supset \beta \in \Gamma$ *then* $\beta \in \Gamma$.
3. *If* $\alpha \in \Gamma$ *then* $\mathbf{L}_i \alpha \in \Gamma$.
4. *If* $\alpha \notin \Gamma$ *then* $\neg \mathbf{L}_i \alpha \in \Gamma$.

Note that the only difference between i-stable sets and the original definition of stable sets lies in condition 1. While stable sets are only required to contain the tautologies of regular propositional logic, i-stable sets must contain all $K45_n$-valid formulas.

The next theorem proves that i-stable sets and i-epistemic states are equivalent notions. In other words, the definition of i-stability falls out naturally given the semantics of OL_n.

[12] In contrast, $\mathbf{O}_i \neg \mathbf{O}_j p$ is satisfiable in Halpern's logic.

[13] Note that if we replace $\mathbf{L}_j p$ by p we obtain regular single-agent autoepistemic reasoning.

[14] The subject matter is defined as a set of atomic propositions π. The restriction says that the atoms occurring within a belief must be contained in π.

Theorem 7
Let Γ be a set of basic formulas. Γ is i-stable iff Γ is an i-epistemic state.

Halpern arrived at the same notion of i-stability independently [Hal93]. He also considered definitions based on logics other than $K45_n$.

Given a set of basic formulas Γ, let $\overline{\Gamma} = \{\gamma \mid \gamma \text{ is basic and } \gamma \notin \Gamma\}$, $\mathbf{L}_i\Gamma = \{\mathbf{L}_i\gamma \mid \text{for all } \gamma \in \Gamma\}$, and $\neg\mathbf{L}_i\overline{\Gamma} = \{\neg\mathbf{L}_i\gamma \mid \text{for all } \gamma \in \overline{\Gamma}\}$.

Definition 14 *i-Stable Expansion*
Let A be a set of basic formulas and let $\models_{K45}$ denote logical consequence in $K45_n$. Γ is called an i-stable expansion of A iff $\Gamma = \{\text{basic } \gamma \mid A \cup \mathbf{L}_i\Gamma \cup \neg\mathbf{L}_i\overline{\Gamma} \models_{K45} \gamma\}$.

The definition of i-stable expansions looks exactly like Moore's definition of stable expansions except that we use logical consequence in $K45_n$ instead of logical consequence in propositional logic. Using the stronger $K45_n$ is necessary in the multi-agent case, since an agent knows that other agents are also perfectly introspective. For example, if agent i believes $\neg\mathbf{L}_j p$ for a different agent j then he also believes $\mathbf{L}_j\neg\mathbf{L}_j p$.

Finally, the following theorem demonstrates that the i-stable expansions of a formula α correspond precisely to the different i-epistemic states of agent i who only-knows α.

Theorem 8 *Only-Knowing and i-Stable Expansions*
Let α be basic and $w \in W_c$ with $\Gamma = \{\text{basic } \alpha \mid w \models \mathbf{L}_i\alpha\}$. Then $w \models \mathbf{O}_i\alpha$ iff Γ is an i-stable expansion of $\{\alpha\}$.

An analogous result was obtained by Halpern [Hal93].

Corollary 4.1 *Basic i-objective formulas have a unique i-stable expansion.*

5 Conclusion

We proposed a multi-agent logic of only-knowing that extends earlier work by Levesque regarding the single-agent case. Our logic gives a semantic and proof theoretic account of autoepistemic reasoning for many knowers. Notions like stable set and stable expansion fall out as natural extensions of single-agent autoepistemic logic.

As for future work, it would be interesting to obtain a complete axiomatization for all of OL. Also, as noted earlier, there are subtle differences between OL_n and Halpern's logic. Halpern showed that every valid sentence in his logic is also valid in OL_n and that the valid sentences of both logics coincide when restricted to $\mathcal{ONL}_n^-$. However, there are sentences such as $\neg\mathbf{O}_i\neg\mathbf{O}_j p$ that are valid in OL_n but not in Halpern's case. For a better comparison of the two approaches it would be interesting to see which modifications are necessary to obtain identical logics. Finally, a more expressive first-order language should be considered to make OL_n more applicable in real world domains. We conjecture that an approach as in [Lev90] could be adapted for this purpose without great difficulty.

Acknowledgements

I would like to thank Joe Halpern and Hector Levesque for fruitful discussions on this subject.

References

[AK88] Appelt, D. and Konolige, K., A Practical Nonmonotonic Theory of Reasoning about Speech Acts, in Proc. of the 26th Conf of the ACL, 1988.

[FHV91] A Model-Theoretic Analysis of Knowledge, Journal of the ACM 91(2), 1991, pp. 382-428.

[Hal93] Halpern, J. Y., Reasoning about only knowing with many agents, in Proc. of the 11th National Conference on Artificial Intelligence (AAAI-93).

[HM84] Halpern, J. Y. and Moses, Y. O., Towards a Theory of Knowledge and Ignorance: Preliminary Report, in Proceedings of The Non-Monotonic Workshop, New Paltz, NY, 1984, pp.125-143.

[HM92] Halpern, J. Y. and Moses, Y. O., A Guide to Completeness and Complexity for Modal Logics of Knowledge and Belief, Artificial Intelligence 54, 1992, pp. 319-379.

[HC84] Hughes, G. E. and Cresswell, M. J., A Companion to Modal Logic, Methuen & Co., London, 1984.

[Hin62] Hintikka, J., Knowledge and Belief: An Introduction to the Logic of the Two Notions, Cornell University Press, 1962.

[Hin71] Hintikka, J., Semantics for Propositional Attitudes, in L. Linsky (ed.), Reference and Modality, Oxford University Press, Oxford, 1971.

[Kri63] Kripke, S. A., Semantical Considerations on Modal Logic, Acta Philosophica Fennica 16, 1963, pp. 83-94.

[Lak93] All They Know About, to appear in: Proc. of the 11th National Conference on Artificial Intelligence (AAAI-93), Washington DC, 1993.

[Lev90] Levesque, H. J., All I Know: A Study in Autoepistemic Logic, Artificial Intelligence, North Holland, 42, 1990, pp. 263-309.

[Moo85] Moore, R., Semantical Considerations on Nonmonotonic Logic, Artificial Intelligence 25, 1985, pp. 75-94.

[Mor90] Morgenstern, L., A Theory of Multiple Agent Nonmonotonic Reasoning, in Proc. of AAAI-90, 1990, pp. 538-544.

[MG92] Morgenstern, L. and Guerreiro, R., Epistemic Logics for Multiple Agent Nonmonotonic Reasoning!, Symposium on Formal Reasoning about Beliefs, Intentions, and Actions, Austin, TX, 1992.

[Per87] Perrault, R., An Application of Default Logic to Speech Act Theory, in Proc. of the Symposium on Intentions and Plans in Communication and Discourse, Monterey, 1987.

[Sta80] Stalnaker, R. C, A Note on Nonmonotonic Modal Logic, Department of Philosophy, Cornell University, 1980.