Only Knowing Meets Common Knowledge

Vaishak Belle*
Dept. of Computer Science
KU Leuven
Belgium
vaishak@cs.kuleuven.be

Gerhard Lakemeyer†
Dept. of Computer Science
RWTH Aachen University
Germany
gerhard@cs.rwth-aachen.de

Abstract

Only knowing captures the intuitive notion that the beliefs of an agent are precisely those that follow from its knowledge base. While only knowing has a simple possible-world semantics in a single agent setting, the many agent case has turned out to be much more challenging. In a recent paper, we proposed an account which arguably extends only knowing to multiple agents in a natural way. However, the approach was limited in that the semantics cannot deal with infinitary notions such as common knowledge. In this work, we lift that serious limitation to obtain a first-order language with only knowing and common knowledge, allowing us to study the interaction between these notions for the very first time. By adding a simple form of public announcement, we then demonstrate how the muddy children puzzle can be cast in terms of logical implications given what is only known initially.

1 Introduction

When considering knowledge-based agents, it seems natural that the beliefs of the agent are those that follow from the assumption that its knowledge base is all that is known.[1] Levesque [1990] was among the first to capture this idea in the logic of *only knowing*, where a modality $\boldsymbol{O}$ is introduced in addition to the classical epistemic operator $\boldsymbol{K}$. For example, from $\boldsymbol{O}p$ it follows that $\neg\boldsymbol{K}q$, which is different from classical epistemic logic where $\boldsymbol{K}p$ does not rule out $\boldsymbol{K}(p \wedge q)$. Similarly, from $\boldsymbol{O}(P(a) \vee P(b))$ we get $\boldsymbol{K}(\exists x[P(x) \wedge \neg\boldsymbol{K}P(x)])$, which says that "if all I know is $P(a)$ or $P(b)$ then I know that there is an instance of P but not what." So, one of the advantages here is that a precise characterization of the *beliefs* and the *non-beliefs* can be given.

While single-agent only knowing has a particularly simple possible-world semantics, the many agent case, where agents may have beliefs about what other agents believe, has turned out to be much more challenging. Early approaches such as [Halpern, 1993; Lakemeyer, 1993; Halpern and Lakemeyer, 2001; Waaler, 2004; Waaler and Solhaug, 2005] either had undesirable properties or had to resort to proof-theoretic concepts such as canonical models built into the semantics. In a recent paper, we [Belle and Lakemeyer, 2010] came up with an account which arguably extends the semantics of only knowing to multiple agents in a natural way. Most significantly, this scheme is for a first-order language with equality and quantification, as in Levesque's proposal, but in stark contrast to approaches such as [Halpern and Lakemeyer, 2001], based on canonical models, that are propositional.[2]

In Levesque's single agent semantics of only knowing, a model, or an epistemic state, is simply a set of worlds e, which satisfies only knowing a sentence α just in case e is maximal, that is, adding any other world to e would lead to not believing α any more. (Worlds are simply truth assignments to atoms of a first-order language and believing is interpreted in the usual way as truth in all worlds in e.)

For many agents, in [Belle and Lakemeyer, 2010], we introduced *sets* of states of affairs as epistemic states, where each element consists of a world in the sense of Levesque together with sets of states of affairs for every other agent, representing their beliefs at this particular world. In order to avoid circularity, each model is limited to a finite depth, where depth refers to the maximal number of alternating nested beliefs considered at this model. The idea is roughly as follows. Suppose there are n agents. At level 1, an agent's epistemic state, called a 1-structure (or k-structure more generally), is simply a set of worlds; a $(k+1)$-structure is a set of tuples $(w, e_1, e_2, \ldots, e_n)$, where each e_i is a k-structure. In order to interpret a formula, one then picks a model consisting of a world and k-structures, one for each agent, such that k is at least as big as the depth of the formula in question. Since k-structures are sets, knowing and only knowing can be defined in a way similar to the single agent case except that interpreting a belief of another agent leads to "popping up" the k-structures of the respective agent. Of course, despite the depth limitation of each model, validity and satisfiability are well defined for all formulas (of arbitrary depth), since each formula has finite depth, and there are enough models of this (or greater) depth. Most significantly, this approach was shown to generalize the *features* of Levesque's logic as

*Partially funded by the FWO project on Data Cleaning and KU Leuven's GOA on Declarative Modeling for Mining and Learning.

†Also affiliated with the University of Toronto.

[1] In this paper, we do not distinguish between "knowledge" and "belief" and freely use the terms interchangeably.

[2] It is also not clear how such a proof-theory based semantics can be extended to the first-order case mainly because there cannot be a complete first-order axiomatization of only knowing even for a single agent [Halpern and Lakemeyer, 1995].

identified in [Halpern and Lakemeyer, 2001], and when restricted to the propositional case, the relationship to many related efforts, including [Halpern and Lakemeyer, 2001], was also established in our previous paper.

However, as we acknowledged in that work, there is a price to pay when using depth-limited models, since there is no single model which works for all formulas. While this is often not a problem, as knowledge bases are usually assumed to be finite collections of sentences, there is one important exception: common knowledge. This notion has been found useful in distributed systems [Moses *et al.*, 1986] as well as in artificial intelligence, game theory and philosophy [Fagin *et al.*, 1995]. A formula ϕ is common knowledge if everybody believes ϕ and everybody believes that everybody believes ϕ, and so on, ad infinitum. Thus, given the depth limitation of the our account, it is not possible to give meaning to such (useful) infinitary notions.

In this paper, we show how this limitation can be overcome. The idea is, roughly, to move from fixed k-structures to infinite sequences of k-structures for all k and all agents, where each member of a sequence for a particular agent is compatible with the beliefs of its predecessor and extends it to account for one additional level of nested beliefs. This construction is shown to be fully compatible with depth-limited models in terms of validity, and so it inherits the reasonableness of the previous account. The treatment is inspired by Fagin *et al.* [1991] and their idea of using infinite sequences of epistemic states for many agents and common knowledge. There are significant technical differences, however, which we will discuss in more detail later. Moreover, Fagin *et al.* do not consider only knowing and limit themselves to the propositional case. Thus, for the very first time, we are able to study the interaction between the notions of only knowing and common knowledge. Interestingly, by adding a simple form of public announcement to the logic, we are also able to demonstrate how the muddy children puzzle [Fagin *et al.*, 1995] can be cast in terms of logical implications given what is only known initially.

We structure the work by beginning with the new logic, discussing some of its properties and extensions before turning to the puzzle. We conclude after discussing related work.

2 The Logic $\mathcal{COL}_n$

Syntax

The non-modal fragment of $\mathcal{COL}_n$ consists of standard first-order logic with $=$ (that is, connectives $\{\wedge, \forall, \neg\}$, syntactic abbreviations $\{\exists, \equiv, \supset\}$, parentheses, period) and a countably infinite set of *standard names* $\mathcal{N}$. As we shall see, these standard names will serve as a fixed domain of discourse, permitting a substitutional interpretation for quantifiers. To keep matters simple, function symbols are not considered in this language. We call a predicate other than $=$, applied to first-order variables and standard names, an *atomic* formula. We write α^x_m to mean that the variable x is substituted in α by a standard name m. If all the variables in a formula α are substituted by standard names, then we call it a *ground* formula.

We let $I = \{1, \ldots, n\}$ be a set of agents, and let i range over this set. $\mathcal{COL}_n$ has three epistemic operators: $\boldsymbol{K}_i\alpha$ is to be read as "i knows α," $\boldsymbol{O}_i\alpha$ is to be read as "all that i knows is α," and $\boldsymbol{C}\alpha$ is to be read as "α is common knowledge among the agents in I." For convenience, we also use $\boldsymbol{E}\alpha$, to be read as "everybody knows α," as an abbreviation for $\boldsymbol{K}_1\alpha \wedge \ldots \wedge \boldsymbol{K}_n\alpha$. Letting $\boldsymbol{E}^1\alpha$ mean $\boldsymbol{E}\alpha$, $\boldsymbol{E}^{k+1}\alpha$ for $k \geqslant 1$ is used as an abbreviation for $\boldsymbol{E}(\boldsymbol{E}^k\alpha)$. A formula not mentioning modalities is called *objective*, and a formula where all predicate symbols appear within the scope of a modal operator is called *subjective.*

Semantics

The semantics is given in terms of possible worlds. Here, a world is simply a set of ground atoms, and let $\mathcal{W}$ denote the set of all possible worlds.

Epistemic states, which are used to define the meaning of subjective formulas, are described in two steps. First, we describe the beliefs of an agent using the notion of a k-structure.

Definition 1: Let $k > 0$. Then define $\mathcal{E}^k$ as follows:

- $\mathcal{E}^1 = 2^{\mathcal{W}}$;
- $\mathcal{E}^k = 2^{\mathcal{W} \times_n \mathcal{E}^{k-1}}$ for $k > 1$.

An element of $\mathcal{E}^k$ is referred to as a k-structure, denoted e^k. (When the context is clear, the superscript is dropped.)

As noted, $e \in \mathcal{E}^1$ for agent i, written e_i for clarity, is of the form $\{w', w'', \ldots\}$ and determines what i believes about the world, but has nothing interesting to say about i's views of other agents. By extension, $e_i \in \mathcal{E}^2$ is of the form $\{(w', e'_1, \ldots, e'_n), (w'', e''_1, \ldots, e''_n), \ldots\}$, which says that at w', i's beliefs about (say) agent n is given by e'_n but at w'', i believes that e''_n captures n's beliefs.

An *epistemic state*, roughly speaking, is used to determine the beliefs of all agents and at all levels.

Definition 2: An *epistemic state* f is a function of the form $I \times \mathbb{N} \rightarrow \bigcup_{k=1}^{\infty} \mathcal{E}^k$ such that for any i and $k \geqslant 1$, $f(i,k) \in \mathcal{E}^k$.

Such an epistemic state is reasonable only when an agent's beliefs are *consistent across all levels*. That is, each level extends the previous level by adding another layer of (nested) beliefs about other agents' beliefs, while keeping the set of worlds an agent considers the same. (This latter set is possibly duplicated to allow for situations where at a world, an agent's beliefs about what another knows differs; see the example below.) To formalize this idea we need the following notion of *i-compatibility* for subsequent k-structures.

Definition 3: We define $e \in \mathcal{E}^k$ and $e' \in \mathcal{E}^{k+1}$ as *i-compatible* inductively as follows:

- for $k = 1$:
  $$e = \{w \mid (w, e_1, \ldots, e_i, \ldots, e_n) \in e'\} \text{ and } e_i = e;$$
- for $k > 1$:
  - $\{w \mid (w, e_1, \ldots, e_i, \ldots, e_n) \in e\} =$ $\{w \mid (w, e'_1, \ldots, e'_i, \ldots, e'_n) \in e'\}$ and $e'_i = e$;
  - for every $(w, e'_1, \ldots, e'_i, \ldots, e'_n) \in e'$ there is a $(w, e_1, \ldots, e_i, \ldots, e_n) \in e$ such that e'_j and e_j are j-compatible for all $j \neq i$.

Definition 4: We say an epistemic state f is *proper* if for all i, for all $k > 0$, $f(i,k)$ and $f(i,k+1)$ are i-compatible.

Henceforth, epistemic states are implicitly assumed to be proper.

Example 5: Here is a simple illustration using three levels and two agents, from the first agent's perspective:

$$f(1,1) = \{w\}$$
$$f(1,2) = \{(w, f(1,1), \{w'\}),\ (w, f(1,1), \{w''\})\}$$
$$f(1,3) = \{(w, f(1,2), \{\ (w', \{w^*\}, \{w'\})\ \}),$$
$$\qquad (w, f(1,2), \{\ (w'', \{w^{**}\}, \{w''\})\ \})\}$$

At level 1, the first agent only considers w possible. At level 2, the second agent is believed to consider the worlds w' and w'' possible. (At this level, the first agent's view of himself is precisely given by w, and so is fully compatible with the first level.) At level 3, at the world w', the second agent believes that the first agent's knowledge is characterized by the world w^*. In other words, the first agent believes that the second agent considers the first agent's knowledge to be given by w^*. Analogously, at w'', the second agent's beliefs about what the first agent knows is given by w^{**}. (Here too, the agents' view of themselves is fully compatible with the previous level.)

We will also be interested in the *progression* of an epistemic state wrt some i and w, by means of which we can refer to the state of affairs as considered possible by i at w.

Definition 6: Given an epistemic state f, we define its *progression* wrt an agent index i and a world w as follows:

$$f_i^w = \{\text{proper } f' \mid \text{for all } k > 0: \text{ if } f'(1,k) = e_1, \ldots, f'(n,k) = e_n, \text{ then } (w, e_1, \ldots, e_n) \in f(i, k+1)\}.$$

Example 7: Using Example 5, and considering only the first level, suppose $f'(1,1) = f(1,1)$ and $f'(2,1) = \{w'\}$. Since $(w, f(1,1), \{w'\}) \in f(1,2)$, f' would belong in f_1^w.

Considering only the first and second levels, let $f'(1,1)$ and $f'(2,1)$ be as above, $f'(1,2) = f(1,2)$ and $f'(2,2) = \{(w', \{w^*\}, \{w'\})\}$. Since $(w, f(1,2), \{(w', \{w^*\}, \{w'\})\}) \in f(1,3)$, f' would belong in f_1^w.

(This can be seen to be analogous to the notion of an accessibility relation in Kripke structures [Fagin *et al.*, 1995].)

We now provide a semantics. By a model, we mean a pair (f, w), where f is a (proper) epistemic state and w is a world. The semantic rules are given inductively as follows:

- $f, w \models p$ iff $p \in w$;
- $f, w \models (m_1 = m_2)$ iff m_1 and m_2 are the same names;
- $f, w \models \neg\alpha$ iff $f, w \not\models \alpha$;
- $f, w \models \alpha \wedge \beta$ iff $f, w \models \alpha$ and $f, w \models \beta$;
- $f, w \models \forall x.\ \alpha$ iff $f, w \models \alpha_m^x$ for every standard name m;
- $f, w \models \boldsymbol{K}_i\alpha$ iff for all $w' \in f(i,1)$, for all $f' \in f_i^{w'}$, $f', w' \models \alpha$;
- $f, w \models \boldsymbol{C}\alpha$ iff $f, w \models \boldsymbol{E}^k\alpha$ for all $k \geqslant 1$;
- $f, w \models \boldsymbol{O}_i\alpha$ iff for all w' and f', $w' \in f(i,1)$ and $f' \in f_i^{w'}$ iff $f', w' \models \alpha$.

A formula $\alpha \in \mathcal{COL}_n$ is said to be *satisfiable* iff there is a proper epistemic state f and a world w such that $f, w \models \alpha$. Given any set Σ of sentences, we write $\Sigma \models \alpha$ (read: "Σ entails α") if for all proper f and w, if $f, w \models \beta$ for every $\beta \in \Sigma$, then $f, w \models \alpha$. We write $\models \alpha$ (read: "α is valid") to mean $\{\} \models \alpha$.

3 Properties

Knowledge and Introspection

As we model knowledge as a set of possible worlds [Kripke, 1963; Hintikka, 1962], it is perhaps not surprising that we obtain the usual properties of $K45_n$ [Fagin *et al.*, 1995]:

- $\models \boldsymbol{K}_i\alpha \wedge \boldsymbol{K}_i(\alpha \supset \beta) \supset \boldsymbol{K}_i\beta$
- $\models \boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\boldsymbol{K}_i\alpha$
- $\models \neg\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\neg\boldsymbol{K}_i\alpha$

The first property says that knowledge is closed under *modus ponens*. The second and third say that the agent knows its beliefs and non-beliefs.

Proof: We prove the case for positive introspection, and the others are analogous. Suppose $f, w \models \boldsymbol{K}_i\alpha$. By definition, for all $w' \in f(i,1)$ and all $f' \in f_i^{w'}$, $f', w' \models \alpha$. Suppose $f, w \models \neg\boldsymbol{K}_i\boldsymbol{K}_i\alpha$. Then there is some $w' \in f(i,1)$ and some $f' \in f_i^{w'}$ such that $f', w' \models \neg\boldsymbol{K}_i\alpha$, that is, some $w'' \in f'(i,1)$ and some $f'' \in f_i'^{w''}$ such that $f'', w'' \not\models \alpha$. However, by definition, $f(i,1) = f'(i,1)$, and so this is a contradiction. ∎

Knowledge and Validity

As a simple property of the semantics, valid sentences are always believed:

- If $\models \alpha$ then $\models \boldsymbol{K}_i\alpha$

This property, however, should be distinguished from the knowledge of true sentences:

- $\alpha \supset \boldsymbol{K}_i\alpha$ is falsifiable

These are analogous to the single agent case [Levesque and Lakemeyer, 2001, Section 4.2].

Knowledge and Barcan Formula

Owing to the fixed domain of discourse, we obtain the Barcan formula for knowledge, including its existential version:

- $\models \forall x\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i(\forall x\alpha)$
- $\models \exists x\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\exists x\alpha$

Proof: We show the case for $\forall$, the other being analogous. Suppose $f, w \models \forall x\boldsymbol{K}_i\alpha$. By definition, $f, w \models \boldsymbol{K}_i\alpha_m^x$ for all names m. By definition, again, for all $w' \in f(i,1)$ and all $f' \in f_i^{w'}$, $f', w' \models \alpha_m^x$. Then, $f', w' \models \forall x\alpha$. So, $f, w \models \boldsymbol{K}_i\forall x\alpha$. ∎

Common Knowledge

Since $\boldsymbol{C}$ is interpreted in terms of $\boldsymbol{K}_i$, it follows that all of the introspection properties also hold for $\boldsymbol{C}$ [Fagin *et al.*, 1995; 1991], as well as a version of the Barcan Formula:

- $\models \boldsymbol{C}\alpha \wedge \boldsymbol{C}(\alpha \supset \beta) \supset \boldsymbol{C}\beta$
- $\models \boldsymbol{C}\alpha \supset \boldsymbol{C}\boldsymbol{C}\alpha$
- $\models \neg\boldsymbol{C}\alpha \supset \boldsymbol{C}\neg\boldsymbol{C}\alpha$
- $\models \forall x\boldsymbol{C}\alpha \supset \boldsymbol{C}(\forall x\alpha)$

Proof: We show the case for positive introspection, and the others are analogous in relation to the proofs for $\boldsymbol{K}_i$. Suppose $f, w \models \boldsymbol{C}\alpha$. By definition, $f, w \models \boldsymbol{E}^k\alpha$ for $k \geqslant 1$. Suppose $f, w \not\models \boldsymbol{C}\boldsymbol{C}\alpha$. That is, $f, w \not\models \boldsymbol{E}^k(\boldsymbol{E}^l\alpha)$ for $k, l \geqslant 1$, and so, $f, w \not\models \boldsymbol{E}^{k+l}\alpha$, which is a contradiction. ∎

An inductive definition of $\boldsymbol{C}$ in terms of $\boldsymbol{E}$ also holds:

- $\models \boldsymbol{C}\alpha \supset \boldsymbol{E}(\alpha \wedge \boldsymbol{C}\alpha)$

Proof: Suppose $f, w \models \boldsymbol{C}\alpha$ but $f, w \not\models \boldsymbol{E}(\alpha \wedge \boldsymbol{C}\alpha)$. Clearly, $f, w \not\models \boldsymbol{E}\alpha$ is a contradiction, and so $f, w \not\models \boldsymbol{E}\boldsymbol{C}\alpha$. From the definition of $\boldsymbol{E}$ and $\boldsymbol{C}$ it follows that $f, w \not\models \boldsymbol{K}_i\boldsymbol{E}^k\alpha$ for some i and some $k \geqslant 1$ and hence $f, w \not\models \boldsymbol{E}^{k+1}\alpha$, which is also a contradiction. ∎

Let us reiterate that since knowledge need not be true, by extension, common knowledge need not be true:

- $\not\models \boldsymbol{K}_i\alpha \supset \alpha$
- $\not\models \boldsymbol{C}\alpha \supset \alpha$

Proof: Let f be any epistemic state such that every world $w \in f(i,1)$ is such that $p \in w$ for some proposition p. Let w^* be a world such that $p \notin w^*$. Clearly, $f, w^* \not\models \boldsymbol{K}_i p \supset p$ showing that $\not\models \boldsymbol{K}_i\alpha \supset \alpha$, and so $\boldsymbol{C}\alpha \supset \alpha$ cannot be valid. ∎

A weaker version in the context of knowledge, however, can be shown:

- $\models \boldsymbol{K}_i(\boldsymbol{C}\alpha \supset \alpha)$

Proof: Suppose $f, w \models \neg\boldsymbol{K}_i(\boldsymbol{C}\alpha \supset \alpha)$. Then for some $w' \in f(i,1)$ and $f' \in f_i^{w'}$, $f', w' \models \neg(\boldsymbol{C}\alpha \supset \alpha)$, that is, $f', w' \models \boldsymbol{C}\alpha \wedge \neg\alpha$. So, $f', w' \models \boldsymbol{K}_i\alpha \wedge \neg\alpha$. However note that, then, $f, w \models \neg\boldsymbol{K}_i\alpha$ and so $f, w \models \boldsymbol{K}_i\neg\boldsymbol{K}_i\alpha$. This means that $f', w' \models \neg\boldsymbol{K}_i\alpha$ as well, an inconsistency. ∎

Only Knowing and Common Knowledge

Here we discuss a few properties on the interaction between only knowing and common knowledge.

Let us begin by noting that, as in the single agent setting [Levesque, 1990], only knowing implies knowing:

- $\models \boldsymbol{O}_i\alpha \supset \boldsymbol{K}_i\alpha$

Proof: The semantic rule for $\boldsymbol{O}_i$ uses "iff" instead of the "if" for $\boldsymbol{K}_i$ to test that the pairs (f', w'), where $w' \in f(i,1)$ and $f' \in f_i^{w'}$, are models of α. So $\boldsymbol{O}_i\alpha$ implies $\boldsymbol{K}_i\alpha$. ∎

This does not extend to common knowledge, because $\boldsymbol{O}_i p$, where p is a proposition, means that i knows nothing about the beliefs of other agents:

- $\models \boldsymbol{O}_i p \supset \neg\boldsymbol{C}p$

Proof: Suppose $f, w \models \boldsymbol{O}_i p \wedge \boldsymbol{C}p$. It is easy to see that $f, w \models \boldsymbol{O}_i p \supset \neg\boldsymbol{K}_i\boldsymbol{K}_j p$ for $j \neq i$ that contradicts the truth of $\boldsymbol{C}p$ at (f, w). ∎

In general, however, $\not\models \boldsymbol{O}_i\alpha \supset \neg\boldsymbol{C}\alpha$. For example, suppose α is $p \wedge \boldsymbol{C}p$. As we shall see, $\boldsymbol{O}_i\alpha$ is satisfiable, and suppose $f, w \models \boldsymbol{O}_i\alpha$. Clearly, then, $f, w \models \boldsymbol{C}p$ and so $f, w \models \boldsymbol{C}(p \wedge \boldsymbol{C}p)$. Thus, $f, w \models \boldsymbol{O}_i\alpha \wedge \boldsymbol{C}\alpha$.

Since common knowledge implies having knowledge about what other agents believe, common knowledge about $\boldsymbol{O}_i p$ is impossible:

- $\models \neg\boldsymbol{C}\boldsymbol{O}_i p$

Proof: Suppose $f, w \models \boldsymbol{C}\boldsymbol{O}_i p$. Then, $f, w \models \boldsymbol{K}_i\boldsymbol{K}_j\boldsymbol{K}_i p$, letting $j \neq i$. By definition of $\boldsymbol{O}_i$, it is easy to show that $\boldsymbol{O}_i p \supset \neg\boldsymbol{K}_i\boldsymbol{K}_j\boldsymbol{K}_i p$ is valid. Because $f, w \models \boldsymbol{C}\boldsymbol{O}_i p$, we have $f, w \models \boldsymbol{K}_i\boldsymbol{O}_i p$, and so, $f, w \models \boldsymbol{K}_i\neg\boldsymbol{K}_i\boldsymbol{K}_j\boldsymbol{K}_i p$, which is a contradiction. ∎

Of course, then, common knowledge that p is not the only thing known is satisfiable:

- $\boldsymbol{C}\neg\boldsymbol{O}_i p$ is satisfiable

Proof: Let w any world such that $p \notin w$. Let $U^1 = \{w\}$, and inductively, $U^{k+1} = \{(w, U^k, \ldots, U^k)\}$. Let f be a proper epistemic structure such that $f(j,1) = U^1$ and $f(j,k) = U^k$ for all j. Then, $f, w' \models \boldsymbol{C}\neg\boldsymbol{O}_i p$ for any w'. ∎

A knowledge base can believe p and also believe that p is common knowledge:

- $\boldsymbol{O}_i(p \wedge \boldsymbol{C}p)$ is satisfiable

Proof: Let $U^1 = \{w \mid p \in w, w \in \mathcal{W}\}$, and inductively, let $U^{k+1} = \{(w, U^k, \ldots, U^k) \mid w \in U^1\}$. Then, let $f(i,k) = U^k$ for every k. It is easy to see $f, w \models \boldsymbol{O}_i(p \wedge \boldsymbol{C}p)$ for any w. ∎

It is worth remarking that one cannot be made to only know $\boldsymbol{C}p$ without actually knowing p:

- $\models \neg\boldsymbol{O}_i\boldsymbol{C}p$

Proof: Suppose $f, w \models \boldsymbol{O}_i\boldsymbol{C}p$. By definition, $w' \in f(i,1)$ and $f' \in f_i^{w'}$ iff $f', w' \models \boldsymbol{C}p$. Since the truth of $\boldsymbol{C}p$ does not depend on the actual world, $f(i,1) = \mathcal{W}$. This means that $f, w \models \neg\boldsymbol{K}_i p$, which is a contradiction. ∎

In fact, the following also holds (as observed for the single agent case in [Levesque and Lakemeyer, 2001]):

- $\models \neg\boldsymbol{O}_i\boldsymbol{K}_i p$

Intuitively, saying that all that is known about the world is $\boldsymbol{K}_i p$ says nothing about the world leading to this property.

4 Relation to Existing Logics

The Logic $\mathcal{OL}_n$

This work not only extends our previous account in the sense of capturing common knowledge, but it is also compatible with it for the language $\mathcal{OL}_n$. (That is, $\mathcal{OL}_n = \mathcal{COL}_n - \boldsymbol{C}$.) We recall some essentials: the prior account was limited to two agents, say $\{a, b\}$; a model of a formula α of *depth* k (definition not reproduced here) is a tuple of the form (e_a, e_b, w), where $e_i \in \mathcal{E}^k$ and $w \in \mathcal{W}$ as defined here. So a formula α of depth k is said to be *satisfiable* iff there is such a tuple. If α is true wrt every (e'_a, e'_b, w), where e'_i is a k'-structure with $k' \geqslant k$, it is said to be *valid.*

As noted, our prior work was shown to be reasonable in terms of salient features of Levesque's logic [Halpern and Lakemeyer, 2001]. Thus, by establishing compatibility, we inherit the reasonableness of only knowing in the many agent case. The proof is not hard but tedious. Here we only go over the main ideas. Suppose $\models'$ denotes the satisfaction relation in [Belle and Lakemeyer, 2010]. Then:

Lemma 8: *Suppose $I = \{a, b\}$ and $\alpha \in \mathcal{OL}_n$ is of depth k. Given any f and w, suppose $f(i,k) = e_i$. Then:*

$$f, w \models \alpha \text{ iff } e_a, e_b, w \models' \alpha.$$

By means of an induction on α, one can argue that formulas of depth k are true wrt (f, w) iff they hold at (e_a, e_b, w). Intuitively, satisfaction wrt $\models$ implies satisfaction wrt $\models'$.

Next, we need a simple property that proper epistemic states can be constructed for any $e_i \in \mathcal{E}^k$:

Proposition 9: *For every pair (e_a, e_b), where $e_i \in \mathcal{E}^k$ for any k, there is a f such that $f(i,k) = e_i$.*

Owing to the definition of proper epistemic states (that is, consistent beliefs across all levels) and the above proposition, satisfaction wrt $\models'$ can be seen to imply satisfaction wrt $\models$, leading to the following result:

Theorem 10: *Suppose $\alpha \in \mathcal{OL}_n$. Then: $\models \alpha$ iff $\models' \alpha$.*

The Logic $KC45_n$

For a propositional language,[3] $\mathcal{OL}_n$ was also related to $K45_n$:

Lemma 11: *[Belle and Lakemeyer, 2010, Lemma 16] Suppose $\mathcal{OL}_n$ is propositional. For any $\alpha \in \mathcal{OL}_n$ not mentioning $\boldsymbol{O}_i$, if α is consistent wrt $K45_n$ then α is satisfiable wrt $\models'$.*

The proof rests on the property that every $K45_n$-consistent formula is satisfiable wrt the *canonical model* for $K45_n$ axioms [Hughes and Cresswell, 1972; Fagin *et al.*, 1995]. A tuple (e_a, e_b, w), where $e_i \in \mathcal{E}^k$, can then be constructed to correspond precisely to the canonical model for formulas of depth k. Of course, by means of Theorem 10, we obtain:

Corollary 12: *Suppose $\mathcal{OL}_n$ is propositional. For any $\alpha \in \mathcal{OL}_n$ not mentioning $\boldsymbol{O}_i$, if α is consistent wrt $K45_n$ then α is satisfiable wrt $\models$.*

From this, letting $KC45_n$ denote $K45_n$ with the common knowledge operator [Fagin *et al.*, 1995], we finally get:

Theorem 13: *Suppose $\mathcal{COL}_n$ is propositional. For any $\alpha \in \mathcal{COL}_n$ not mentioning $\boldsymbol{O}_i$, if α is consistent wrt $KC45_n$ then α is satisfiable wrt $\models$.*

Proof (sketch): The proof uses Lemma 11 and Corollary 12 to argue that if $\boldsymbol{E}^k\alpha$ for any $k > 0$ is satisfiable in $KC45_n$, then it is satisfiable wrt $\models$. Since $\boldsymbol{C}\alpha$ is true iff $\boldsymbol{E}^k\alpha$ is true for every $k > 0$, the argument follows. ∎

5 Extensions

Truthful Knowledge

As noted, knowledge need not be true, but we can require it, using a straightforward definition:

Definition 14: Given any proper f and any w, the pair (f, w) is called a *knowledge model* iff for all i, we have $w \in f(i, 1)$.

(By the definition of proper, w will be considered possible by all agents at all levels.) Logical consequence, then, can be extended in an obvious way; $\Sigma \models \alpha$ means that in every knowledge model where Σ is true, so is α. And indeed, as required of knowledge that is true [Fagin *et al.*, 1995], the following properties are shown to hold:

- $\models \boldsymbol{K}_i\alpha \supset \alpha$
- $\models \boldsymbol{C}\alpha \supset \alpha$

As is needed, knowledge models will be assumed for the developments in subsequent sections.

[3]For a language with quantification, $\mathcal{OL}_n$ differs from the usual treatments of first-order epistemic logic [Fagin *et al.*, 1995, Chapter 3] in assuming a fixed domain of discourse called standard names as in [Levesque and Lakemeyer, 2001].

Public Announcements

In this paper, we are also interested in showing how the logical system can be used for a puzzle in a *dynamical* context that involves *announcements*. Perhaps the simplest way to capture this is to consider the addition of an *announcement modality* [Baltag *et al.*, 1998] to $\mathcal{COL}_n$. So let $\mathcal{COL}_n^+$ be the addition of the modality $[\phi]$ to $\mathcal{COL}_n$, where $\phi \in \mathcal{COL}_n$. To give a semantics for this modality, we first define the intersection operator for epistemic states:

Definition 15: Given any f and f', we define $f^* = f \cap f'$ inductively as follows. For every i:

- $f^*(i, 1) = \{w \mid w \in f(i, 1) \text{ and } w \in f'(i, 1)\}$
- for $k > 1$, $f^*(i,k) = \{(w, e_1, \ldots, e_n) \mid (w, e_1, \ldots, e_n) \in f(i,k) \text{ and } (w, e_1, \ldots, e_n) \in f'(i,k)\}$

For this definition, it is easy to show the following property:

Proposition 16: *Suppose $\alpha, \beta \in \mathcal{COL}_n$ are Boolean combinations of objective formulas and formulas of the form $\boldsymbol{K}_i\phi$ and $\boldsymbol{O}_i\phi$, where $\phi \in \mathcal{COL}_n$. Suppose f and f' are epistemic states, and w is any world. For $j \neq i$, if $f, w \models \boldsymbol{K}_j\alpha$ and $f', w \models \boldsymbol{K}_j\beta$ then $f \cap f', w \models \boldsymbol{K}_j(\alpha \wedge \beta)$; if $f, w \models (\neg\boldsymbol{K}_j\beta \wedge \neg\boldsymbol{K}_j\neg\beta)$ and f' is as before, then $f \cap f', w \models \boldsymbol{K}_j\beta$.*

We define the meaning of $\alpha \in \mathcal{COL}_n^+$ inductively as before, with the following new rule:

- $f, w \models [\phi]\alpha$ iff $f, w \models \phi$ implies $f|_\phi, w \models \alpha$.

Here $f|_\phi = f \cap f'$, where f' is any proper epistemic state such that $f', w \models \bigwedge \boldsymbol{O}_i(\phi \wedge \boldsymbol{C}\phi)$. So, the announcement of ϕ is defined simply in terms of an epistemic state where ϕ and common knowledge about ϕ is all that is believed. (Preconditions for announcements are omitted for simplicity [Baltag *et al.*, 1998].) Intuitively, the only knowing modality ensures that the epistemic state considered is the "maximal" one where ϕ holds at all levels. On intersecting such an epistemic state with the current one f, all structures where ϕ is not believed are discarded.

The operator can be shown to have the following reasonable properties [Lutz, 2006]:

- $\models [\phi]p \equiv (\phi \supset p)$
- $\models [\phi]\neg\alpha \equiv (\phi \supset \neg[\phi]\alpha)$
- $\models [\phi]\alpha \wedge \beta \equiv ([\phi]\alpha \wedge [\phi]\beta)$
- $\models [\phi]\boldsymbol{K}_i\alpha \equiv (\phi \supset \boldsymbol{K}_i(\phi \supset [\phi]\alpha))$
- $\models [\phi]\forall x\alpha \equiv \forall x([\phi]\alpha)$

Proof: We prove the first item only, and the others can be proved by induction using the first item as the base case. By definition, $f, w \models [\phi]p$ iff $f, w \models \phi$ implies $f|_\phi, w \models p$. We claim $f, w \models \phi$ implies $f|_\phi, w \models p$ iff $f, w \models \phi \supset p$. Suppose $f, w \not\models \phi$, then the LHS is vacuously true but then so is the RHS. Conversely, suppose $f, w \models \phi$. Then $f|_\phi, w \models p$ iff (by definition) $p \in w$ iff $f, w \models p$ because the epistemic state is irrelevant iff (by assumption) $f, w \models \phi \supset p$. ∎

6 The Muddy Children Puzzle

The muddy children puzzle [Barwise, 1981; Moses *et al.*, 1986], which is a variant of the cheating wives puzzle [Gamow and Stern, 1958], brings out subtle changes to the knowledge states of a group of logical agents. The situation is as follows. Imagine n children playing together, leading some of them to get mud on their foreheads. Each child can see the foreheads of all other children, but not its own. Along comes the father, who says, "At least one of you has mud on your forehead," thus expressing a fact known to each of them before he spoke. The father then asks the question, "Do any of you know whether your forehead is dirty?," over and over. If two children have muddy foreheads, for example, every child announces "no" the first time the question is asked; but the next time, the dirty ones declares that their forehead is muddy.

Here, a purely deductive account of the puzzle is developed as logical consequences of what is only known initially, thereby complementing the usual approach with model constructions [Fagin *et al.*, 1995].

Formally, assume m_i says that child i has a muddy forehead. Initially, let us suppose that every child has a muddy forehead, that is, the real world w is any world where *AllMuddy* is true, where:

$$AllMuddy \doteq m_1 \wedge \ldots \wedge m_n$$

Next, we take note of three properties of the puzzle to characterize the knowledge bases of the children: (a) first, i sees that every other child has a muddy forehead; (b) second, i knows that others see its forehead; and (c) third, the father announces initially that at least one child has a muddy forehead, which is the only thing assumed to be common knowledge because every child has heard the father's announcement. Putting this together, the initial theory Σ is the conjunction of *AllMuddy* and the knowledge bases for every i:

$$\boldsymbol{O}_i(OthersMuddy \wedge (\boldsymbol{X}m_i \vee \boldsymbol{X}\neg m_i) \wedge \boldsymbol{C}(AtLeastOne))$$

where,

- $OthersMuddy \doteq \bigwedge_{j\neq i} m_j$ captures (a);
- $\boldsymbol{X}\alpha \doteq \bigwedge_{j\neq i} \boldsymbol{K}_j\alpha$ abbreviates "they know" for (b);
- $AtLeastOne \doteq \bigvee m_i$ is the father's message in (c).

Finally, every time the father asks the question of whether the children know their foreheads are muddy, their announcements are lumped together as:

$$No \doteq \neg\boldsymbol{K}_1 m_1 \wedge \ldots \neg\boldsymbol{K}_n m_n.$$

The puzzle, then, is:

Theorem 17: $\models \Sigma \supset \overbrace{[No]\ldots[No]}^{n-1 \text{ times}}(\boldsymbol{K}_1 m_1 \wedge \ldots \wedge \boldsymbol{K}_n m_n)$.

After $(n-1)$ occurrences of *No*,[4] the children know that they have muddy foreheads. To better see the knowledge bases in action, we give the argument for $n = 2$; that is, suppose $I = \{a, b\}$, and we are to prove:

$$\models \Sigma \supset [No](\boldsymbol{K}_a m_a \wedge \boldsymbol{K}_b m_b).$$

Proof: Let (f, w) be a knowledge model for Σ and let f' be the epistemic state where $\boldsymbol{O}_a(No \wedge \boldsymbol{C}(No)) \wedge \boldsymbol{O}_b(No \wedge \boldsymbol{C}(No))$ holds. We prove the case for a concluding that its own forehead is muddy, the other being analogous.

Owing to the closure of knowledge under modus ponens, $(\boldsymbol{K}_b(m_a \vee m_b) \wedge \boldsymbol{K}_b\neg m_a) \supset \boldsymbol{K}_b m_b$ is valid. (In contrast, $(\boldsymbol{K}_b(m_a \vee m_b) \wedge \boldsymbol{K}_b m_a) \supset \boldsymbol{K}_b m_b$ is falsifiable.) Let α denote $\boldsymbol{K}_b(m_a \vee m_b)$, β denote $\boldsymbol{K}_b\neg m_a$, γ denote $\boldsymbol{K}_b m_b$ and δ denote $\boldsymbol{K}_b m_a$. So $(\alpha \wedge \beta) \supset \gamma$ is valid. Owing to the knowledge of valid sentences, the validity of $(\alpha \wedge \neg\gamma) \supset \neg\beta$ means $\boldsymbol{K}_a((\alpha \wedge \neg\gamma) \supset \neg\beta)$ is valid, and so is true at f and f'.

Since *AtLeastOne* is common knowledge in Σ, $f, w \models \boldsymbol{K}_a\boldsymbol{K}_b AtLeastOne$, that is, $f, w \models \boldsymbol{K}_a\alpha$. By construction, $f, w \models \neg\boldsymbol{K}_a\gamma \wedge \neg\boldsymbol{K}_a\neg\gamma$ because of only knowing. (That is, in the absence of only knowing, there are epistemic states where, say, $\boldsymbol{K}_a\gamma$ is believed.) By construction, $f', w \models \boldsymbol{K}_a\neg\gamma$ because of only knowing. Using Proposition 16, $f \cap f', w \models \boldsymbol{K}_a\neg\beta$.

On expanding $\boldsymbol{X}$ in Σ, we get $f, w \models \boldsymbol{K}_a(\boldsymbol{K}_b m_a \vee \boldsymbol{K}_b\neg m_a)$, that is, $f, w \models \boldsymbol{K}_a(\beta \vee \delta)$. By construction, $f', w \models (\neg\boldsymbol{K}_a(\beta \vee \delta) \wedge \neg\boldsymbol{K}_a\neg(\beta \vee \delta))$ because of only knowing. By Proposition 16, $f \cap f', w \models \boldsymbol{K}_a(\beta \vee \delta)$.

Putting it together, we get $f \cap f', w \models \boldsymbol{K}_a\delta$, that is, $f \cap f', w \models \boldsymbol{K}_a\boldsymbol{K}_b m_a$. When knowledge is true the sentence $\boldsymbol{K}_b\psi \supset \psi$ is valid, and so $f \cap f', w \models \boldsymbol{K}_a m_a$. ∎

7 Related Work

As noted, there have been some prominent proposals for multiagent only knowing such as [Halpern, 1993; Lakemeyer, 1993; Halpern and Lakemeyer, 2001; Waaler, 2004; Waaler and Solhaug, 2005]. Besides being propositional, they have problematic features, as discussed at length in [Belle and Lakemeyer, 2010]. See that work on how these problems are avoided using k-structures.

The underlying notion of only knowing is due to Levesque [1990]. Incidentally, there are some related notions, such as *total knowledge* [Pratt-Hartmann, 2000] and *minimal knowledge* [Halpern and Moses, 1984], the latter of which has also received recent multiagent treatments [van Der Hoek and Thijsse, 2002]. However, these notions differ significantly from Levesque's only knowing [Levesque and Lakemeyer, 2001]. Although not the focus of this paper, we note that when the knowledge base itself refers to the agent's beliefs, only knowing also exhibits a form of nonmonotonic reasoning; see [Levesque and Lakemeyer, 2001] and [Belle and Lakemeyer, 2015] for discussions and references in the single and multiagent cases respectively.

Let us also remark that multiagent logics of knowledge are an active focus in artificial intelligence [van der Hoek and Wooldridge, 2012], with a number of extensions for reasoning about time, actions, desires, and intentions, among others. (For an account using sets of possibilities, see [Gerbrandy and Groeneveld, 1997].) Investigations in the area on common knowledge and muddy children variants are

[4] More generally, if $k \leqslant n$ children have mud on their foreheads, $k-1$ announcements of *No* need to be made [Fagin *et al.*, 1995]. For example, if $k = 1$, owing to *AtLeastOne* being common knowledge, the child with the muddy forehead notices that others have clean ones, and concludes immediately that he must be the muddy one.

also fairly prevalent; for example, see [Fagin *et al.*, 1995; van Ditmarsch *et al.*, 2007]. The analysis of the puzzle here is mostly by means of model constructions, with approaches such as [Kraus and Lehmann, 1988; Elgot-Drapkin, 1991] being notable exceptions. In particular, one would observe in this latter work that the muddy children puzzle requires an explicit enumeration of the non-beliefs of the agents. In contrast, we can get off the ground simply in terms of what is believed initially (with help from only knowing), which (to the best of our knowledge) has not been obtained before.

Finally, the inspiration for the infinite sequence of structures is the work in [Fagin *et al.*, 1991]. They introduce what they call *knowledge structures*, also for different levels. While at level 1 our concepts match (that is, just worlds), a level 2 structure of theirs is already different from a 2-structure. Most significantly, they are propositional, do not treat only knowing and require knowledge to always be true. The way that the infinite sequences work in the semantics of the two proposals differs as well: in our case, the C modality causes the epistemic states to repeatedly progress, whereas in theirs there are structures for every nesting of this modality which requires structures corresponding to many ordinal numbers. Nonetheless, they do establish a connection to the logic $KC45_n$ [Fagin *et al.*, 1991, Corollary 5.9], as do we, and so we would agree on formulas of this language.

8 Conclusions

A first-order logic of only knowing and common knowledge was introduced that allows us to investigate the interaction between these notions, and provide a solution to the muddy children puzzle in terms of what is only known initially. Among other things, this logic is shown to be fully compatible with (and extend) previous accounts, and thus is a very general first-order proposal of only knowing in the many agent case.

Beginning with $O\mathcal{L}_n$ [Belle and Lakemeyer, 2010], exploring issues related to axiomatization and nonmonotonic reasoning in the presence of common knowledge would make for interesting future work.

References

[Baltag *et al.*, 1998] A. Baltag, L. S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Proc. TARK*, pages 43–56, 1998.

[Barwise, 1981] J. Barwise. Scenes and other situations. *Journal of Philosophy*, 78(7):369–397, 1981.

[Belle and Lakemeyer, 2010] V. Belle and G. Lakemeyer. Multi-agent only-knowing revisited. In *Proc. KR*, pages 49–60, 2010.

[Belle and Lakemeyer, 2015] V. Belle and G. Lakemeyer. Semantical considerations on multiagent only knowing. *Artif. Intell.*, 223:1–26, 2015.

[Elgot-Drapkin, 1991] J. J. Elgot-Drapkin. Step-logic and the three-wise-men problem. In *Proc. AAAI*, pages 412–417, 1991.

[Fagin *et al.*, 1991] R. Fagin, J. Y. Halpern, and M. Y. Vardi. A model-theoretic analysis of knowledge. *J. ACM*, 38(2):382–428, 1991.

[Fagin *et al.*, 1995] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

[Gamow and Stern, 1958] G. Gamow and M. Stern. *Puzzle Math*. New York: Viking Press, 1958.

[Gerbrandy and Groeneveld, 1997] J. Gerbrandy and W. Groeneveld. Reasoning about information change. *J. of Logic, Lang. and Inf.*, 6(2):147–169, April 1997.

[Halpern and Lakemeyer, 1995] J. Y. Halpern and G. Lakemeyer. Levesque's axiomatization of only knowing is incomplete. *Artif. Intell.*, 74(2):381–387, 1995.

[Halpern and Lakemeyer, 2001] J.Y. Halpern and G. Lakemeyer. Multi-agent only knowing. *Journal of Logic and Computation*, 11(1):251–265, 2001.

[Halpern and Moses, 1984] J. Y. Halpern and Y. Moses. Towards a theory of knowledge and ignorance: Preliminary report. In *Proc. NMR*, pages 125–143, 1984.

[Halpern, 1993] J. Y. Halpern. Reasoning about only knowing with many agents. In *Proc. AAAI*, pages 655–661, 1993.

[Hintikka, 1962] J. Hintikka. *Knowledge and belief: an introduction to the logic of the two notions*. Cornell University Press, 1962.

[Hughes and Cresswell, 1972] G. E. Hughes and M. J. Cresswell. *An introduction to modal logic*. Methuen London, 1972.

[Kraus and Lehmann, 1988] S. Kraus and D. J. Lehmann. Knowledge, belief and time. *Theor. Comput. Sci.*, 58:155–174, 1988.

[Kripke, 1963] S. Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

[Lakemeyer, 1993] Gerhard Lakemeyer. All they know: A study in multi-agent autoepistemic reasoning. In *Proc. IJCAI*, pages 376–381, 1993.

[Levesque and Lakemeyer, 2001] H.J. Levesque and G. Lakemeyer. *The logic of knowledge bases*. The MIT Press, 2001.

[Levesque, 1990] H. J. Levesque. All I know: a study in autoepistemic logic. *Artif. Intell.*, 42(2-3):263–309, 1990.

[Lutz, 2006] C. Lutz. Complexity and succinctness of public announcement logic. In *Proc. AAMAS*, 2006.

[Moses *et al.*, 1986] Y. Moses, D. Dolev, and J. Y. Halpern. Cheating husbands and other stories: A case study of knowledge, action, and communication. *Distributed Computing*, 1(3), 1986.

[Pratt-Hartmann, 2000] I. Pratt-Hartmann. Total knowledge. In *Proc. AAAI*, pages 423–428, 2000.

[van Der Hoek and Thijsse, 2002] W. van Der Hoek and E. Thijsse. A general approach to multi-agent minimal knowledge: With tools and samples. *Studia Logica*, 72(1):61–84, 2002.

[van der Hoek and Wooldridge, 2012] W. van der Hoek and M. Wooldridge. Logics for multiagent systems. *AI Magazine*, 33(3):92–105, 2012.

[van Ditmarsch *et al.*, 2007] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2007.

[Waaler and Solhaug, 2005] A. Waaler and B. Solhaug. Semantics for multi-agent only knowing: extended abstract. In *Proc. TARK*, pages 109–125, 2005.

[Waaler, 2004] A. Waaler. Consistency proofs for systems of multi-agent only knowing. In *Advances in Modal Logic*, pages 347–366. College Publications, 2004.