Mathematical Foundations for Joining Only Knowing and Common Knowledge

Marcos Cramer[1] , **Samuele Pollaci**[2] , **Bart Bogaerts**[2]
[1]Institute of Artificial Intelligence, TU Dresden, Germany
[2]Department of Computer Science, Vrije Universiteit Brussel, Belgium
marcos.cramer@tu-dresden.de, {samuele.pollaci, bart.bogaerts}@vub.be

Abstract

Common knowledge and only knowing capture two intuitive and natural notions that have proven to be useful in a variety of settings, for example to reason about coordination or agreement between agents, or to analyse the knowledge of knowledge-based agents. While these two epistemic operators have been extensively studied in isolation, the approaches made to encode their complex interplay failed to capture some essential properties of only knowing. We propose a novel solution by defining a notion of μ-biworld for countable ordinals μ, which approximates not only the worlds that an agent deems possible, but also those deemed impossible. This approach allows us to define a multi-agent epistemic logic with common knowledge and only knowing operators, and a three-valued model semantics for it. Moreover, we show that we only really need biworlds of depth at most $\omega^2 + 1$. Based on this observation, we define a Kripke semantics on a canonical Kripke structure and show that this semantics coincides with the model semantics. Finally, we discuss issues arising when combining negative introspection or truthfulness with only knowing and show how positive introspection can be integrated into our logic.

1 Introduction

When developing intelligent agents, it is important that they can reason not just about the world they are placed in, but also about the knowledge or beliefs of other agents in their environment. Consider for instance a traffic situation where two cars meet at a crossing. When the driver of the first car observes their light is green, the typical action to be taken would be to keep on driving. The reason for this is twofold: on the one hand, the driver knows the traffic regulations and the fact that the other car should stop. But on the other hand, the driver also knows (or believes) that the other driver also knows the regulations (and will likely respect them). It is this knowledge about the other driver's knowledge that allows the first driver to conclude that it is safe to continue.

The formal study of knowledge and how to reason correctly about it has a long history in knowledge representation, dating back at least to the 1960s (Hintikka 1962). This study becomes particularly interesting when, as in our example, multiple agents are involved. Next to the standard knowledge operator K, we are concerned with two epistemic operators, namely **common knowledge** and **only knowing**, and their intricate interplay. We say that a statement φ is *common knowledge* among a group G of agents (and denote this $C_G\varphi$) if each agent in G knows φ, and also knows that every agent in G knows φ, and knows that everyone in G knows that everyone in G knows φ, and so on. This operator is useful, for instance, for reasoning about coordination or agreement between agents. We say that an agent A *only knows* a statement φ (and denote this $O_A\varphi$) if the agent knows φ (denoted $K_A\varphi$) and moreover *everything* they know follows from φ (so whenever $K_A\psi$ holds, it must be the case that φ entails ψ). This operator is useful for instance when we consider that knowledge-based agents do not know anything except for what follows from their knowledge base and we might want to reason about their knowledge as well.

There have been many papers studying these operators in isolation (Fagin et al. 1995; Meyer and van der Hoek 1995; Halpern and Lakemeyer 2001; Waaler and Solhaug 2005; Belle and Lakemeyer 2010; Belle and Lakemeyer 2015b) and some authors have even studied the combination of the two (Aucher and Belle 2015; Belle and Lakemeyer 2015a; Van Hertum 2016). However, we will argue that there are some essential properties of only knowing that none of these approaches captures. We will start with the good news: it is easy to develop a Kripke semantics for a logic with these two operators: given a Kripke structure $\mathcal{K}$ with set of worlds W and an accessibility relations R_A for every agent A, the semantics for C_G and O_A is given by

- $\mathcal{K}, w \models C_G\varphi$ if $\mathcal{K}, w' \models \varphi$ for all w' reachable from w with edges in $\bigcup_{A\in G} R_A$, and
- $\mathcal{K}, w \models O_A\varphi$ if for every world $w' \in W$, $\mathcal{K}, w' \models \varphi$ if and only if $(w, w') \in R_A$.

Intuitively, the if-part in the definition of the semantics of O_A states that A knows φ (in all worlds A deems possible, φ holds), and the only-if-part ensures that the agent doesn't know anything else (all worlds in which φ holds are indeed deemed possible by the agent in question).

If this is so easy, then where's the catch, one might wonder. Well, the problem lies in the choice of the set W of worlds. The question we tackle is: can we construct a set of worlds W that is rich enough such that, for instance, $O_A\top$ really means that that A only knows tautologies in the language (i.e., that agent A "knows nothing")? In other words, can we construct a canonical Kripke structure for this logic? A naive first attempt at doing so would be as follows.

Definition 1 (World — incorrect definition)**.** *Given a propositional vocabulary Σ, a world w consists of*

- *a (classical propositional) interpretation w^{obj} over Σ (the* objective interpretation *of w), and*
- *for each agent $A \in \mathcal{A}$, a set of worlds A^w.*

However, the attentive reader might have noticed that this circular definition breaks the basic rules of set theory (a world is defined to consist of, among others, a set of worlds). We are not the first to observe this issue. The most common approach to alleviate it is to approximate the knowledge of agents up to a certain level and defining k+1-worlds as consisting, among others, of a set A^w of k-worlds for each agent A. There are two challenges with this approach. The first is that as soon as we add common knowledge to our language, there is a strong need for these worlds to be *infinitely deep*. One way to achieve this is to consider an infinite *precision-increasing* sequence of k-worlds, as is done for instance by Fagin, Halpern, and Vardi (1991), or by Belle and Lakemeyer (2015a). However, that in itself does not suffice: we show that to evaluate certain formulas there is a need to have worlds that are even deeper than this. The second is that in such approximations, no matter how deep one goes, there is never enough information to conclude that this is "all we know": it might always be that by making the approximation more precise, more knowledge at some later level comes in.

This brings us to the main contribution of this paper: a solution to the above problem. First we will define a notion of μ-biworld, where μ can be any countable ordinal (thereby resolving the first challenge), and where the "bi" in biworld stands for the fact that we do not just approximate the set of worlds an agent deems possible, but also the set of worlds an agent deems *impossible*. Intuitively, with each μ+1-biworld w, we will associate a set A^w of μ-biworlds representing the set of μ-biworlds of which the agent deems some extension possible, and a set $\bar{A}^w$ of μ-biworlds of which the agent deems some extension impossible. This immediately allows us to see when all of agent A's knowledge is captured by such a biworld: this is precisely when $A^w \cap \bar{A}^w$ is empty. For limit ordinals, the situation is more complex, and this makes the construction of biworlds highly technical and mathematical. Due to space limitations, proofs are not included in this paper, but all propositions and lemmas are carefully proven in a technical report accompanying this paper (Cramer, Pollaci, and Bogaerts 2023). This transfinite construction of our biworlds is given in Section 2; Section 3 then shows several properties of them, essentially showing that they are well-behaved, in a precise sense.

In Section 4 we define our logic of common knowledge and only knowing as a simple multi-agent epistemic logic extended with the operators O_A and C_G as described above. We show that the biworlds possess enough information so that the formulas in our logic can be evaluated in them. More specifically, we define a three-valued model semantics for our logic: given a formula φ and a biworld w of any depth, we define what φ^w is. This can be true, false, or unknown, where the last case represents the fact that the biworld is not sufficiently "deep" to evaluate the formula. Moreover, we show that we do not need arbitrarily deep biworlds: if a biworld has depth at least $\omega^2 + 1$, then every formula will evaluate in it either to true or to false. Inspired by this observation, we are able to define our canonical Kripke structure: the *worlds*, are precisely the ω^2+1-biworlds that are *completed*, which is a technical term to denote the fact that it identifies for every other biworld of any other level whether or not the agent believes it is possible, i.e., that it characterizes complete knowledge. The accessibility relations can directly be obtained from the definition of the worlds. We then proceed to show that the semantics obtained from this canonical Kripke-structure actually coincides with the valuation we started from. Finally, this allows us to prove several desirable properties the resulting logic satisfies (see Theorem 5 for an extensive list), including the following two:

- For any φ and ψ, if $\varphi \not\models \psi$, then $O_A\varphi \models \neg K_A\psi$.
- For any φ, $O_A\varphi \not\models \bot$.

The first property states precisely that whenever $O_A\varphi$ holds, all formulas not entailed by φ are not known (and in fact we also have that all properties entailed by φ are known by A). The second property states that for *any* formula φ, there is a world in which $O_A\varphi$ holds. In fact, our results are even stronger than this: we show that there is a unique state-of-mind of agent A in which they know precisely φ. These two properties of the O_A operator, while quite simple, are — to the best of our knowledge — not satisfied by any other paper combining common knowledge and only knowing.

While developing our worlds, and our logic, we do not enforce any properties that are often associated to *knowledge*. For instance, our worlds do not guarantee that our agents are *truthful* or *introspective*. The main focus of the paper is on how to create a semantic structure that allows defining a rich enough set of worlds. However, once this semantic structure is in place, it is possible to use it to define a logic that satisfies such properties as well. To illustrate this, we show in Section 5 how we can hard-code into the logic the fact that agents are positively introspective, which in fact means that it is common knowledge that all agents satisfy this property. We also discuss the possibility of adding truthfulness and negative introspection, highlighting a problem that arises from combining them with only knowing. We use this problem to uncover a major mistake in previous work on combining only knowing and common knowledge.

Finally, in Section 6 we discuss related work and explain why previous approaches could not achieve the two properties set out above. We conclude in Section 7.

2 Construction of Biworlds

In this section, we define the concept of biworlds. Intuitively, a biworld w consists of *(1)* an objective interpretation, and *(2)* for each agent A two sets of biworlds A^w and $\bar{A}^w$, where the former set contains all the biworlds the agent deems possible, and the latter the biworlds the agent deems impossible. As explained before, using this definition brings us into set-theoretic problems. To resolve this issue, we define the notion of μ-biworlds, for all countable ordinals μ. Intuitively, a μ-biworld does not describe the complete epistemic state of the different agents, but only their belief about the world up to a certain depth. For some formulas, having a certain depth will suffice to evaluate the formula. For instance, a 1-biworld will suffice to determine whether or not

an agent knows p, but might not suffice to determine whether an agent knows that some other agent knows p. Moreover, we will later show that for our logic, the ordinal ω^2+1 suffices, in the sense that every set of formulas is either true or false in every ω^2+1-biworld. While we only need the theory of biworlds up to ω^2+1, we will still develop the theory in more generality for all *countable* ordinals. From now on, when we say *ordinal*, we mean *countable ordinal*.

If μ is a successor ordinal $\alpha+1$, then a μ-biworld w associates with each agent two sets of α-biworlds, where A^w represents the set of α-biworlds that can be extended (in depth; for some sufficiently large ordinal) to a biworld the agent deems possible, and $\bar{A}^w$ represents the set of α-biworlds that can be extended to a biworld the agent deems impossible. Clearly, a natural condition will then be that the union of these two sets is the set of all α-biworlds.

If there is any biworld in the intersection of A^w and $\bar{A}^w$, this biworld must be extendible both to a biworld the agent deems possible, and to a biworld the agent deems impossible. If for all agents A, the intersection of A^w and $\bar{A}^w$ is empty, this means that the information in w fully specifies which biworlds the agents deem possible and which ones impossible. Once this is fully specified, there is only one way to extend w to a biworld of a greater depth. When there are multiple ways in which a μ-biworld w can be extended to a $\mu+1$-biworld, then we call w an *incompleted* biworld.

Since a biworld in the intersection of A^w and $\bar{A}^w$ must be extendible (for some sufficiently large depth) both to a biworld the agent deems possible and to a biworld the agent deems impossible, it must be incompleted.

We start with the auxiliary notion of a μ-prebiworld, which approximates the more complex notion of a μ-biworld while leaving out some of the structural conditions that we impose on μ-biworlds. We define μ-prebiworlds, restrictions of prebiworlds and a precision order on the prebiworlds through **simultaneous recursion**, but for readability, we split it into Definitions 2, 3 and 4.

Definition 2 (μ-Prebiworld)**.** *Let μ be an ordinal. We define the set of* μ-prebiworlds *over a propositional vocabulary Σ, and a set of agents $\mathcal{A}$ by transfinite induction:*

- *A 0-prebiworld w is an interpretation $\mathcal{I}$ over Σ. We define $w^{obj}=\mathcal{I}$ and $d(w)=0$.*
- *A $\mu+1$-prebiworld w is a triple $(\mathcal{I},(A^w)_{A\in\mathcal{A}},(\bar{A}^w)_{A\in\mathcal{A}})$, where $\mathcal{I}$ is an interpretation over Σ, and for each $A\in\mathcal{A}$, A^w and $\bar{A}^w$ are sets of μ-prebiworlds. We define $d(w)=\mu+1$ and $w^{obj}=\mathcal{I}$.*
- *A λ-prebiworld w for a limit ordinal λ is a precision-increasing transfinite sequence $(w_\alpha)_{\alpha<\lambda}$ of prebiworlds, i.e. for each $\alpha<\lambda$, w_α is an α-prebiworld and for each $\beta\leq\alpha$, $w_\beta\leq_p w_\alpha$. We define $d(w)=\lambda$ and $w^{obj}=w_0$.*

For each ordinal μ, we denote by $\mathcal{W}_p^\mu$ the set of μ-prebiworlds, and we call the integer $d(w)$ the *depth of* w. If $w\in\mathcal{W}_p^\lambda$ with λ a limit ordinal, then for each $\alpha<\lambda$, we denote by $(w)_\alpha$ the α-prebiworld which is the element in position α in the transfinite sequence represented by w.

The following definition captures what it means to restrict a prebiworld of depth μ to a prebiworld of smaller depth α.

Definition 3 (Restriction of a prebiworld)**.** *Assume w is a μ-prebiworld and $\alpha\leq\mu$. The* restriction *of w to α is the α-prebiworld $w|_\alpha$ defined as follows:*

- *If $\alpha=0$, then $w|_\alpha=w^{obj}$.*
- *If α is a limit ordinal, then $w|_\alpha=(w|_\beta)_{\beta<\alpha}$*
- *If $\alpha=\alpha'+1$, we distinguish two cases:*
  - *If μ is a limit ordinal, then $w|_\alpha=(w)_\alpha$.*
  - *If μ is a successor ordinal $\mu=\mu'+1$, then $w|_\alpha^{obj}=w^{obj}$, $A^{w|_\alpha}=\{w'|_{\alpha'}\mid w'\in A^w\}$, and $\bar{A}^{w|_\alpha}=\{w'|_{\alpha'}\mid w'\in\bar{A}^w\}$.*

The precision order on the prebiworlds is based on the notion of restriction:

Definition 4 (Precision order)**.** *If w is a μ-prebiworld and v an α-prebiworld with $\alpha\leq\mu$, we say that v is less precise than w (and denote this $v\leq_p w$) if $w|_\alpha=v$.*

If $w\leq_p v$ for some prebiworlds w and v, we call w a *restriction of* v and v an *extention of* w.

Now we define the set of incompleted prebiworlds and the set of biworlds by simultaneous transfinite recursion. For better readability, we separate this simultaneous definition into Definitions 5 and 6 and explain afterwards why it is a successful definition.

Definition 5 (Incompleted prebiworld)**.** *A μ-prebiworld w is* incompleted *if there exists two distinct $\mu+1$-biworlds v_1 and v_2 such that $v_1,v_2\geq_p w$.*

A prebiworld is called *completed* if it is not incompleted.

Definition 6 (μ-Biworld)**.** *A μ-*biworld *is a μ-prebiworld w such that one of the following conditions holds:*

(a) $\mu=0$;

(b) $\mu=\mu'+1$ is a successor ordinal, and for each agent $A\in\mathcal{A}$ the following hold:
  (1) the union $A^w\cup\bar{A}^w$ is the set of all μ'-biworlds;
  (2) for each $v\in A^w\cap\bar{A}^w$, v is incompleted;

(c) μ is a limit, and w_α is an α-biworld for each $\alpha<\mu$.

For each ordinal μ, the set of μ-biworlds is denoted as $\mathcal{W}^\mu$. We often call a 0-biworld an *objective world*.

Since the definition of *incompleted prebiworlds* of depth μ depends on the existence of biworlds of depth $\mu+1$, an attentive reader may be worried whether the simultaneous definition of *incompleted prebiworlds* and *biworlds* is really a successful definition. We therefore now explain how this definition is to be understood.

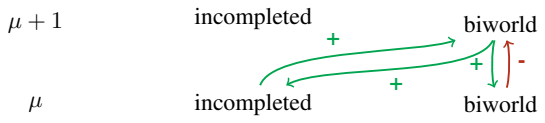

Figure 1: Dependencies between the notions of *incompleted prebiworlds* and *biworlds* at levels μ and $\mu+1$

For this Figure 1 depicts the dependencies between the notions of *incompleted prebiworlds* and *biworlds* at levels μ and $\mu+1$. Here, a green arrow from one notion to another indicates that the second notion positively depends on the first, i.e., will include more objects when the first notion includes more objects. A red arrow, on the other hand, indicates a negative dependency, so that the second notion will

include fewer objects when the first notion includes more objects. For example, the green arrow from *biworld* at level $\mu+1$ to *incompleted* at level μ indicates that determining additional biworlds at level $\mu+1$ can lead to determining a prebiworld w at level μ to be incompleted, as one of the newly determined biworlds at level $\mu+1$ may imply there is more than one way of extending w to a $+1$-biworld. The arrow from *biworld* at level μ to *biworld* at level $\mu+1$ is red, because given a $\mu{+}1$-prebiworld w, determining more biworlds at level μ can show that the union $A^w \cup \bar{A}^w$ is not the set of all μ-biworlds, i.e., that w is not a biworld.

Now the simultaneous definition of *incompleted prebiworlds* and *biworlds* is a transfinite recursion over μ, where at each level μ of the induction, the notions of an *incompleted μ-prebiworld* and a *$\mu{+}1$-biworld* are defined simultaneously by choosing the minimal set of incompleted μ-prebiworlds and $\mu{+}1$-biworlds that satisfy Definitions 5 and 6(b). Such a minimal set exists, because the definitions of *incompleted μ-prebiworld* and *$\mu{+}1$-biworld* only depend positively on each other. Here the negative dependency of *$\mu{+}1$-biworld* on *μ-biworld* is not a problem, because at this stage in the transfinite recursion over μ, the set of μ-biworlds has already been determined: If μ is a successor ordinal $\mu'+1$, it has been determined in the previous step μ' of the transfinite recursion. If μ is a limit ordinal, it has been determined by Definition 6(c) and the fact that for every $\alpha<\mu$, the set of α-biworlds has already been determined. If $\mu=0$, the set of μ-biworlds just coincides with the set of μ-prebiworlds by Definition 6(a).

Example 1. *We consider one of the simplest settings, namely we suppose to have just one agent A, and one propositional variable p in the vocabulary Σ. The set $\mathcal{W}^0$ of 0-biworlds coincides with the set of all 0-prebiworlds by Definition 6, and it is equal to $\mathcal{W}^0 = \{\{p\},\{\emptyset\}\}$. The set $\mathcal{W}^1$ already counts considerably more elements, eighteen to be precise: nine with $\{p\}$ as objective world, and other nine with $\{\emptyset\}$. This follows from the limitations imposed by Item b1 in Definition 6 of biworlds*[1]*. For the sake of conciseness, we omit the explicit description of all the 1-biworlds but one:*

$$v_1 := (\{p\},\{\{p\}\},\{\{p\},\{\emptyset\}\}).$$

Since $\{p\}$ is the objective world of v_1, p is true in v_1. Moreover, the sets A^{v_1} and $\bar{A}^{v_1}$ provide some information on the beliefs of A in v_1. Recall that A^{v_1} (resp. $\bar{A}^{v_1}$) is the set of biworlds that have an extension A deems possible (resp. impossible). Since $\{p\}$ is the only biworld in A^{v_1}, p is true in any biworld A deems possible. As we will see later, this means that A knows p. On the contrary, since $\bar{A}^{v_1}$ contains both $\{p\}$ and $\{\emptyset\}$, p is true in some of the biworlds A deems impossible, and false in others.

Starting with the objective world $v_0 := \{p\}$ and continuing with the biworld v_1, we can inductively build an ω-prebiworld $v := (v_\alpha)_{\alpha<\omega}$ as follows

$$v_\alpha := \begin{cases} \{p\} & \text{if } \alpha = 0 \\ (v_0, \{v_{\alpha'}\}, \mathcal{W}^{\alpha'}) & \text{if } \alpha = \alpha'+1. \end{cases}$$

Since for all successor ordinals $\alpha = \alpha'+1 < \omega$ we have $A^{v_\alpha} \cap \bar{A}^{v_\alpha} = \{v_{\alpha'}\}$, to prove that v is an ω-biworld, it suffices to show that for all $\alpha<\omega$, v_α is incompleted, which can be done by an inductive proof. Moreover, v can be shown to be completed.

[1] In the single-agent setting, $|\mathcal{W}^1| = n3^n$, where $n = |\mathcal{W}^0|$.

We now dive deeper in certain features of (pre)biworlds with limit ordinal depth. Note that for successor ordinals μ, and for any μ-prebiworld, each agent is equipped with two sets of prebiworlds. For limit ordinals μ, however, this is not the case. The following definition aims at retrieving a similar concept for a prebiworld with limit ordinal depth.

Definition 7. *Given a limit ordinal λ, a λ-prebiworld w and an agent $A \in \mathcal{A}$, we define the following sets*

$$A{\uparrow}w := \{v \in \mathcal{W}_p^\lambda \mid \forall \mu<\lambda : (v)_\mu \in A^{(w)_{\mu+1}}\}$$

$$\bar{A}{\uparrow}w := \{v \in \mathcal{W}_p^\lambda \mid \forall \mu<\lambda : (v)_\mu \in \bar{A}^{(w)_{\mu+1}}\}$$

It is clear that if w in the definition above is a biworld, then $A{\uparrow}w$ and $\bar{A}{\uparrow}w$ are sets of biworlds. It is important to notice that the sets defined in Definition 7 do not carry the exact same meaning as the successor-ordinal counterpart: if $v \in A{\uparrow}w$ for some λ-(pre)biworld w, then in w, for each approximation v_α ($\alpha<\lambda$) of v, the agent A deems some extension of v_α as possible. Analogously, if $v \in \bar{A}{\uparrow}w$, then in w, for each approximation v_α ($\alpha<\lambda$) of v, the agent A believes some extension of v_α is impossible.

3 Properties of Biworlds

In this section, we show that the formal definitions stated in Section 2 behave well and that they indeed correspond to the intuitive ideas introduced above. First, we present two propositions regarding the notion of restriction, which show that the induced relation $\leq_p$ is a non-strict partial order. Second, we provide some additional insight into the sets defined in Definition 7. Finally, we focus on certain properties concerning biworlds. In particular, we will show the following fundamental facts in Theorem 1:

Restrictability: the restriction of a biworld is a biworld.

Monotonicity of completedness: an extension of a completed biworld is a completed biworld.

Completability: all biworlds have a a completed extension.

Completedness at successor ordinals: a $\mu{+}1$-biworld w is completed if and only if $A^w \cap \bar{A}^w = \emptyset$ for all agents A.

Given Definitions 3 and 5, the first two properties are sensible to have. Completability will be fundamental in Section 4 and it is clearly desirable: a complete biworld characterizes complete knowledge, providing a full description of the epistemic state of the agents. The last property provides a simple characterization of what it means to be completed for biworlds of successor ordinal depth. While the four properties listed above may seem natural and straightforward, several intermediate results are required to prove them to hold.

As anticipated, we start by showing that Definition 4 defines a non-strict partial order on the set of prebiworlds.

Proposition 1. *Let w be a μ-prebiworld and let $\beta \leq \alpha \leq \mu$. Then $w|_\mu = w$ and $w|_\alpha|_\beta = w|_\beta$.*

Proposition 2. *The relation $\leq_p$ is a non-strict partial order (a reflexive, antisymmetric and transitive relation). The induced strict order $<_p$ is a well-founded order.*

Before proceeding to the properties of biworlds, we focus on the sets of Definition 7. Analogously to what happens

in Item b of Definition 6 for biworlds of successor ordinal depth, the union of $A{\uparrow}w$ and $\bar{A}{\uparrow}w$ is the whole set of biworlds of that depth.

Proposition 3. *Let λ be a limit ordinal, and w be a λ-biworld. For all $A \in \mathcal{A}$, we have $A{\uparrow}w \cup \bar{A}{\uparrow}w = \mathcal{W}^{\lambda}$.*

It is interesting to notice that the second part of Item b of Definition 6 does not hold in general for the sets $A{\uparrow}w$ and $\bar{A}{\uparrow}w$, i.e., their intersection might contain completed biworlds (see Example 2). Nevertheless, this intersection tells us something about the completedness of w, as stated in the following proposition.

Proposition 4. *Let λ be a limit ordinal and w a λ-biworld. If $A{\uparrow}w \cap \bar{A}{\uparrow}w = \emptyset$ for all $A \in \mathcal{A}$, then w is completed.*

Example 2. *Consider the completed ω-biworld v defined in Example 1. It is easy to see that $A{\uparrow}v = \{v\}$ and $\bar{A}{\uparrow}v = \mathcal{W}^{\omega}$. Hence, we get $A{\uparrow}v \cup \bar{A}{\uparrow}v = \mathcal{W}^{\omega}$ and $A{\uparrow}v \cap \bar{A}{\uparrow}v = \{v\}$, which agrees with Proposition 3 and shows that the converse of Proposition 4 does not hold. Moreover, notice that $A{\uparrow}v \cap \bar{A}{\uparrow}v$ contains a completed biworld.*

The following theorem formally states the fundamental properties of biworlds presented at the beginning of this section. The proof of it requires several intermediate technical lemmas, which we omit due to space limitations.

Theorem 1. *Let $\alpha \leq \mu' \leq \mu < \beta$ be ordinals, let w be a μ-biworld, and w' be a μ'-biworld. The following hold:*
1. *The restriction $w|_{\alpha}$ is an α-biworld.*
2. *If w' is completed and $w \geq_p w'$, then w is completed.*
3. *There exists a completed β-biworld v such that $v \geq_p w$.*
4. *If μ is a successor ordinal, then w is completed if and only if for each $A \in \mathcal{A}$, $A^w \cap \bar{A}^w = \emptyset$.*

Theorem 1 directly implies an important corollary:

Corollary 1. *A μ-biworld w is completed if and only if there exists exactly one $\mu{+}1$-biworld v such that $v \geq_p w$. In this case, also v is completed.*

4 The logic $\mathcal{COL}$

In this section we define the syntax and semantics of the logic $\mathcal{COL}$ that allows to speak about knowledge, common knowledge and only knowing. More specifically, we first define a three-valued model semantics, where the value of a $\mathcal{COL}$ formula is either true ($\mathbf{t}$), false ($\mathbf{f}$), or unknown ($\mathbf{u}$). We consider two orders on these truth values: the *precision order* given by $\mathbf{u} \leq_p \mathbf{f}$ and $\mathbf{u} \leq_p \mathbf{t}$, and the *truth order* given by $\mathbf{f} \leq_t \mathbf{u} \leq_t \mathbf{t}$. We will write v^{-1} for the *inverse* of the truth value v, defined by $\mathbf{f}^{-1} = \mathbf{t}$, $\mathbf{t}^{-1} = \mathbf{f}$, and $\mathbf{u}^{-1} = \mathbf{u}$. We will show that our three-valued semantics is precision-monotonic, in the sense that more precise biworlds give more precise results. Moreover, we will show that in a biworld of depth at least ω^2, every formula evaluates to either true or false. This fact prompts us to define an alternative two-valued semantics. In more detail, since dealing with biworlds with a limit ordinal depth may be complicated and counter-intuitive, we will move our focus to the smallest (for the sake of simplicity) successor ordinal at which biworlds evaluate all formulas (and all sets of formulas) as true or false, namely $\omega^2 + 1$. In addition, we restrict to completed $\omega^2{+}1$-biworlds: by Item 4 in Theorem 1, they have the intuitive property that A^w and $\bar{A}^w$ are disjoint, which corresponds to the intuition that these two sets represent the biworlds deemed possible and the biworlds deemed impossible by A. This motivates an alternative semantic characterization of our logic through a canonical Kripke structure consisting of completed $\omega^2{+}1$-biworlds[2], which we show to coincide with the three-valued semantics on the relevant biworlds. Furthermore, we show that $\mathcal{COL}$ is semantically well-behaved. In Section 6, we will show that no previously proposed semantic approach leads to a logic that is semantically well-behaved in the way specified in this section.

To say that an agent A only knows φ can be viewed as a conjunction of the statement that A knows φ, denoted $K_A\varphi$, and the statement that A knows at most φ, denoted $M_A\varphi$. Given a set G of agents, we write $E_G\varphi$ for the statement that every agent in G knows φ, and we write $C_G\varphi$ for the statement that φ is common knowledge within the set G of agents. The language $\mathcal{COL}$ extends propositional logic with these modal operators as follows:

Definition 8. *We define the language $\mathcal{COL}$ by structural induction with the standard recursive rules of propositional logic, augmented with:*

$$K_A(\psi) \in \mathcal{COL} \text{ if } \psi \in \mathcal{COL} \text{ and } A \in \mathcal{A}$$
$$M_A(\psi) \in \mathcal{COL} \text{ if } \psi \in \mathcal{COL} \text{ and } A \in \mathcal{A}$$
$$E_G(\psi) \in \mathcal{COL} \text{ if } \psi \in \mathcal{COL} \text{ and } G \subseteq \mathcal{A}$$
$$C_G(\psi) \in \mathcal{COL} \text{ if } \psi \in \mathcal{COL} \text{ and } G \subseteq \mathcal{A}$$

We use $O_A\varphi$ as syntactic sugar for $K_A\varphi \wedge M_A\varphi$.

In the introduction, we have already discussed what an intuitive Kripke semantics for $O_A\varphi$ would be. Adapting these ideas to the representation of $O_A\varphi$ as $K_A\varphi \wedge M_A\varphi$, we can easily see that the correct way to define the Kripke semantics for $M_A\varphi$ is as follows:

- $\mathcal{K}, w \models M_A\varphi$ if for every world $w' \in W$ such that $\mathcal{K}, w' \models \varphi$, we have $(w, w') \in R_A$.

In order to explain why this is a good semantic characterization of "knowing at most φ", we will sketch a proof that the only way in which $M_A\varphi$ and $K_A\psi$ can both be true is when φ entails ψ: Assume $M_A\varphi$ and $K_A\psi$ are true in a world w. We want to show that φ entails ψ, i.e., that ψ is true in every world w' in which φ is true. Assume w' is a world in which φ is true. But the assumption that $M_A\varphi$ is true in w together with the above definition of the Kripke semantics for $M_A\varphi$ implies that $(w, w') \in R_A$. This together with the assumption that $K_A\psi$ is true in w implies that ψ is true in w', as required.

In preparation for the upcoming discussion of a three-valued semantics for $\mathcal{COL}$, note that if we write $(M_A\varphi)^{\mathcal{K},w} = \mathbf{t}$ and $(M_A\varphi)^{\mathcal{K},w} = \mathbf{f}$ for $\mathcal{K}, w \models M_A\varphi$ and $\mathcal{K}, w \not\models M_A\varphi$ respectively, the above characterization of the semantics of $M_A\varphi$ is equivalent to the following:

[2]For any $\mu \geq \omega^2$, all the results would hold if we considered a canonical Kripke structure consisting of completed μ-biworlds. Taking all completed μ-biworlds for all ordinals μ would also work, but some uniqueness results would be lost, as we would have many worlds representing the same object.

$$(M_A\varphi)^{\mathcal{K},w} = \mathrm{glb}_{\leq_t}\{(\varphi^{\mathcal{K},w'})^{-1} \mid w' \notin R_A^w\}$$

Let us now turn to the three-valued valuation of formulas of $\mathcal{COL}$. All parts of this definition are precisely what one would expect when applying a Kleene-style three-valued semantic approach to logics with a common knowledge operator, taking into account the above rewording of the Kripke semantics of $M_A\varphi$.

Definition 9. *Given a formula $\varphi \in \mathcal{COL}$ and a μ-biworld w, we define the three-valued valuation function φ^w by induction on μ and the structure of φ:*

$$P^w = \mathbf{t} \text{ if } P \in w^{obj} \text{ and } P^w = \mathbf{f} \text{ otherwise}$$
$$(\varphi \wedge \psi)^w = glb_{\leq_t}(\varphi^w, \psi^w)$$
$$(\neg\varphi)^w = (\varphi^w)^{-1}$$
$$(K_A\varphi)^w = \begin{cases} \mathbf{u} & \text{if } \mu = 0 \\ \mathrm{glb}_{\leq_t}\{\varphi^{w'} \mid w' \in A^w\} & \text{if } \mu = \mu'{+}1 \\ \mathrm{lub}_{\leq_p}\{(K_A\varphi)^{(w)_{\mu'}} \mid \mu'{<}\mu\} & \text{if } \mu \text{ is a limit} \end{cases}$$
$$(M_A\varphi)^w = \begin{cases} \mathbf{u} & \text{if } \mu = 0 \\ \mathrm{glb}_{\leq_t}\{(\varphi^{w'})^{-1} \mid w' \in \bar{A}^w\} & \text{if } \mu = \mu'{+}1 \\ \mathrm{lub}_{\leq_p}\{(M_A\varphi)^{(w)_{\mu'}} \mid \mu'{<}\mu\} & \text{if } \mu \text{ is a limit} \end{cases}$$
$$(E_G\varphi)^w = \mathrm{glb}_{\leq_t}\{(K_A\varphi)^w \mid A \in G\}$$
$$(C_G\varphi)^w = \mathrm{glb}_{\leq_t}\{(E_G^k\varphi)^w \mid k \geq 1\}$$

where we define $E_G^k\varphi$ inductively as $E_G^0\varphi = \varphi$, $E_G^{k+1}\varphi = E_G(E_G^k\varphi)$. We say a μ-biworld w satisfies *a formula φ (notation: $w \models \varphi$) if $\varphi^w = \mathbf{t}$. A μ-biworld w satisfies, or is a model of, a theory if it satisfies all formulas in that theory. We say a μ-biworld w* resolves *a formula φ if $\varphi^w \neq \mathbf{u}$.*

The following proposition asserts that the three-valued valuation is $\leq_p$-monotonic:

Proposition 5. *For every pair w, w' of biworlds such that $w \leq_p w'$ and every formula $\varphi \in \mathcal{COL}$, we have $\varphi^w \leq_p \varphi^{w'}$.*

The notion of the *modal depth* of a formula allows us to specify conditions for a formula to be resolved by a biworld:

Definition 10 (Modal depth)**.** *The* modal depth $MD(\varphi)$ *of a formula $\varphi \in \mathcal{COL}$ is defined by inductively as follows:*
- $MD(P) = 0$ *for every propositional atom P*
- $MD(\neg\varphi) = MD(\varphi)$
- $MD(\varphi \wedge \psi) = max(MD(\varphi), MD(\psi))$
- $MD(K_A\varphi) = MD(\varphi) + 1$
- $MD(M_A\varphi) = MD(\varphi) + 1$
- $MD(E_G\varphi) = MD(\varphi) + 1$
- $MD(C_G\varphi) = MD(\varphi) + \omega$, *which is the smallest limit ordinal greater than $MD(\varphi)$.*

Note that the modal depth of any formula in $\mathcal{COL}$ is less than ω^2. Every formula of a given modal depth is resolved at any biworld of at least this depth:

Theorem 2. *If w is a μ-biworld and $\varphi \in \mathcal{COL}$ is a formula such that $MD(\varphi) \leq \mu$, then w resolves φ.*

As explained at the beginning of this section, by Item 4 of Theorem 1, a $\mu{+}1$-biworld w precisely captures the knowledge of every agent A iff w is completed. Hence, Theorem 2 combined with the fact that the modal depth of any formula in $\mathcal{COL}$ is less than ω^2 motivates focusing on completed $\omega^2{+}1$-biworlds, as $\omega^2{+}1$ is the first successor ordinal greater than the modal depth of all formulas.

Definition 11. *A* world *is a completed $\omega^2{+}1$-biworld.*

Definition 12. *The $\omega^2{+}1$-completed* Kripke structure *$\mathcal{K}^* := (U, (R_A)_{A\in\mathcal{A}})$ is the Kripke structure whose underlying world set U is the set of all worlds, and whose accessibility relations R_A are given by*

$$R_A = \{(w, w') \in U^2 \mid w'|_{\omega^2} \in A^w\}.$$

Instead of $(w, w') \in R_A$, we sometimes write wR_Aw'.

Interpreting formulas in this canonical Kripke structure $\mathcal{K}^*$ in the standard way (with the above specified semantics for $M_A\varphi$) amounts to a two-valued valuation of $\mathcal{COL}$.

The following theorem tells us that the two-valued and three-valued valuations fully coincide on worlds:

Theorem 3. *If $\varphi \in \mathcal{COL}$ and $w \in U$, then $\varphi^w = \varphi^{\mathcal{K}^*,w}$.*

The next theorem is of central importance to show that our semantics generally avoids a problem that some previous accounts of only knowing and common knowledge had, namely the problem that $O_A\neg C_Gp$ is not satisfiable in those accounts, even though it should be (see Section 6 for a discussion of this problem in other accounts). The following theorem shows that no such problems can arise in the $\omega^2{+}1$-completed Kripke structure $\mathcal{K}^*$:

Theorem 4. *Let A be an agent. For every formula $\varphi \in \mathcal{COL}$, there is a world w such that $(O_A\varphi)^w = (O_A\varphi)^{\mathcal{K}^*,w} = \mathbf{t}$. Moreover, if w_1 and w_2 are two such worlds, then $A^{w_1} = A^{w_2}$ and $\bar{A}^{w_1} = \bar{A}^{w_2}$.*

We will now show that the two-valued valuation of $\mathcal{COL}$ (and thus by Theorem 3 also the three-valued valuation, when restricted to suitable biworlds) gives rise to a sensible entailment relation between formulas of $\mathcal{COL}$. We define this relation as follows:

Definition 13. *Let $\varphi \in \mathcal{COL}$ be a formula, and $\Gamma \subseteq \mathcal{COL}$ be a set of formulas. We write $\Gamma \models \varphi$ if $\varphi^{\mathcal{K}^*,w} = \mathbf{t}$ for every world w such that $\psi^{\mathcal{K}^*,w} = \mathbf{t}$ for all $\psi \in \Gamma$.*

Note that this is definition of the entailment relation does not coincide with the standard way of defining the entailment with respect to a Kripke semantics, as in our case we fix a canonical Kripke structure rather than quantifying over all Kripke structures. The fact that this entailment relation behaves in a sensible way is captured by the properties listed in the following theorem:

Theorem 5. *Let $\varphi, \psi \in \mathcal{COL}$ two formulas, $\Gamma, \Gamma' \subset \mathcal{COL}$ two sets of formulas, $A \in \mathcal{A}$ an agent, and $G \subseteq \mathcal{A}$ a non-empty set of agents. Then, the following properties hold:*
1. *(Prop) For each propositional tautology φ, we have $\models \varphi$.*
2. *(MP) $\varphi, \varphi \Rightarrow \psi \models \psi$.*
3. *(Mono) If $\Gamma \models \varphi$, then $\Gamma, \psi \models \varphi$.*
4. *(Cut) If $\Gamma \models \varphi$ and $\Gamma', \varphi \models \psi$, then $\Gamma, \Gamma' \models \psi$.*
5. *(K) $\models (K_A(\varphi \Rightarrow \psi) \wedge K_A\varphi) \Rightarrow K_A\psi$.*
6. *(Nec) If $\models \varphi$, then $\models K_A\varphi$.*
7. *(M) If $\varphi \not\models \psi$, then $M_A\varphi \models \neg K_A\psi$.*
8. *(O) $O_A\varphi \not\models \bot$.*
9. *(Fixed point axiom) $\models C_G\varphi \Leftrightarrow E_G(\varphi \wedge C_G\varphi)$.*

10. (Induction rule) If $\varphi \models E_G(\varphi \wedge \psi)$, then $\varphi \models C_G\psi$.

Properties 1, 5, 6, 7, and 8 of Theorem 5 ensure that the semantics properly captures the intended meaning of the only-knowing operator. In more detail, properties 1, 5, 6, and 7 imply that for every φ and ψ, either $O_A\varphi \Rightarrow K_A\psi$ or $O_A\varphi \Rightarrow \neg K_A\psi$. More specifically, the former holds when ψ is entailed by φ, the latter otherwise. This means that $O_A\varphi$ completely determines the agent's knowledge. Property 8 states that for any formula $\varphi \in \mathcal{COL}$, there is a world in which $O_A\varphi$ holds, i.e. it is possible that an agent knows φ and knows nothing beyond φ.

The following example discusses the construction of a world w that shows that $O_A\neg C_Gp$ is satisfiable (in line with item 8 of Theorem 5), something that previous attempts at combining only knowing and common knowledge failed at, even though it should intuitively be the case.

Example 3. *Consider the setting of Example 1. We want to construct a world w that satisfies $O_A\neg C_Gp$, where $G = \{A\}$. By the definition of the only knowing operator, we must have*

$$A^w = \{w' \in \mathcal{W}^{\omega^2} \mid (C_Gp)^{w'} = \mathbf{f}\}$$

$$\bar{A}^w = \{w' \in \mathcal{W}^{\omega^2} \mid (C_Gp)^{w'} = \mathbf{t}\}.$$

We first have to find the above sets.

Notice that, by Theorem 2, we have $A^w \cup \bar{A}^w = \mathcal{W}^{\omega^2}$, as supposed. By Proposition 5, Theorem 2, and the fact that any ω^2-biworld is the extension of some ω-biworld, we can reduce to finding the set of ω-biworlds satisfying C_Gp. It is not hard to see that v from Example 1 satisfies C_Gp. Moreover, since v is completed, by Corollary 1, v has a unique extension v' to depth ω^2, and v'. In an analogous fashion, we can build another ω^2-biworld u' satisfying C_Gp by considering the unique extension of a completed ω-biworld $u := (u_\alpha)_{\alpha<\omega}$ defined as

$$u_\alpha := \begin{cases} \{\emptyset\} & \text{if } \alpha = 0 \\ (u_0, \{v_{\alpha'}\}, \mathcal{W}^{\alpha'}) & \text{if } \alpha = \alpha' + 1 \end{cases}$$

Intuitively, both v' and u' are worlds in which p is common knowledge (for the only agent A), but p is true in the objective world of v' and false in the one of u'. In particular, in u' the agent A is not truthful.

We claim[3] v and u are the only ω-biworlds satisfying C_Gp. Hence, A^w must be $\mathcal{W}^{\omega^2} \setminus \{v', u'\}$ and $\bar{A}^w$ must be $\{v', u'\}$, and we define $w := (\{p\}, \mathcal{W}^{\omega^2} \setminus \{v', u'\}, \{v', u'\})$.

Notice that by Theorem 4, w is unique up to change of objective world. In other words, the world defined as w but with $\{\emptyset\}$ as objective world w_0 satisfies $O_A\neg C_Gp$ too.

5 Truthfulness and Introspection

In the previous three sections we have defined and described the construction of a structure of worlds that is rich enough to allow to formally define the semantics of only knowing and common knowledge in a way that matches basic intuitions about these logical modalities in a precisely specified way. A major challenge in designing this construction was to ensure that we have enough worlds to describe all logically possible epistemic states. For this reason, we decided to keep the construction as general as possible, i.e., not to unnecessarily limit the set of worlds.

However, there are certain properties of the knowledge modality that are often taken for granted in epistemic logic and that require limiting the set of worlds. In particular, the following properties are often assumed to hold:

- Truthfulness: $K_A\varphi \Rightarrow \varphi$ is satisfied in every world for every agent A.
- Positive introspection: $K_A\varphi \Rightarrow K_AK_A\varphi$ is satisfied in every world for every agent A.
- Negative introspection: $\neg K_A\varphi \Rightarrow K_A\neg K_A\varphi$ is satisfied in every world for every agent A.

While all three of these properties are commonly assumed in epistemic logic, there are specific issues about ensuring the first or the third in a logic with a modality $O_A\varphi$ for only knowing or a modality $M_A\varphi$ for knowing at most.

In the case of truthfulness, there is an issue concerning the formula $O_A(K_Ap \vee K_Aq)$. Given the definition of the modality O_A, this entails $K_A(K_Ap \vee K_Aq)$, which by truthfulness entails $K_Ap \vee K_Aq$, so either K_Ap or K_Aq has to be true. But since A only knows $K_Ap \vee K_Aq$ and since $K_Ap \vee K_Aq$ entails neither p nor q, neither p nor q can be known, a contradiction. Thus $O_A(K_Ap \vee K_Aq)$ cannot be satisfiable in a logic with truthfulness, which means that principle (O) from Theorem 5 cannot hold in such a logic (assuming principles (Prop), (Cut) and (M) do hold).

In the case of negative introspection, a more severe problem arises. Suppose M_Aq is true in some world w. Since q does not entail p, by (M) this should entail that $\neg K_Ap$ is true in w, so by negative introspection, $K_A\neg K_Ap$ is true in w. But since q does not entail $\neg K_Ap$, principle (M) also allows us to conclude that $\neg K_A\neg K_Ap$ is true in w, a contradiction. Thus M_Aq is not satisfiable, and similarly no formula of the form $M_A\varphi$ or $O_A\varphi$ is satisfiable for any satisfiable φ. It should be stressed that this problem is not caused by our semantic approach to only knowing, but is a direct consequence of basic properties that only knowing has been assumed to satisfy also in other papers.

Remark 1. *A reader familiar with the literature on only knowing, of which we give an overview in Section 6, may wonder why this problem was not identified in previous papers, e.g. Belle and Lakemeyer (2015b), who define a semantics for only knowing in a logic with negative introspection. What the authors did not realize is that by enforcing negative introspection in their logic, they actually made all statements of the form $O_A\varphi$ unsatisfiable (for any satisfiable formula φ). They wrongly claim on page 5 that $O_i(p \wedge Cp)$ is satisfiable, but the alleged proof is wrong. If one defines $V^1 = \{w \mid p, q \in w, w \in \mathcal{W}\}$, $V^{k+1} = \{(w, V^k, \ldots, V^k) \mid w \in V^1\}$, $f'(i,k) = V^k$ and $w' = \{p\}$, then one can easily see that $f' \notin f_i^{w'}$ and $f', w' \models p \wedge Cp$, contradicting their claim that $f, w \models O_i(p \wedge Cp)$.*

In order to avoid this problem, one would need to make use of autoepistemic logic (Moore 1985), or some multi-agent version thereof (Lakemeyer 1993; Permpoontanalarp

[3]The claim can be proven by induction, and it can already be seen to hold true by writing down all the eighteen 1-biworlds, and reasoning on the conditions that make C_Gp satisfied. We omit this reflection for the sake of conciseness.

and Jiang 1995; Vlaeminck et al. 2012; Van Hertum et al. 2016) in the definition of the semantics of $M_A\varphi$: Intuitively, $M_A\varphi$ should be true in a world in which all worlds that satisfy all formulas entailed by the autoepistemic theory $\{\varphi\}$ are accessible. This would give rise to a logic in which a variant of principle (M) with a negated autoepistemic entailment in the place of the negated entailment is satisfied. We leave it to future work to develop the details of such a theory and investigate whether it behaves as intended.

Given that it is somewhat problematic to incorporate truthfulness and negative introspection in a logic with only knowing, whereas no similar problems arise for positive introspection, we describe how the semantic framework from the previous sections can be used to define a logic of only knowing and common knowledge in which positive introspection is ensured.

Definition 14. *A world w is called* positively introspective (PI) *if for each $A \in \mathcal{A}$ and for any worlds w', w'', $w''R_Aw'R_Aw$, implies $w''R_Aw$. A world w is called* recursively PI *if w is positively introspective and all the worlds that are reachable from w through the union $\bigcup_{A\in\mathcal{A}} R_A$ of all accessibility relations are positively introspective.*

We now define $\mathcal{K}^*_{PI} := (U^{PI}, (R^{PI}_A)_{A\in\mathcal{A}})$ to be the Kripke substructure of $\mathcal{K}^*$, where the underlying world set U^{PI} is the set of all recursively PI worlds, and the accessibility relations are just the ones coming from $\mathcal{K}^*_{PI}$ being a substructure, i.e., for all $A \in \mathcal{A}$, $R^{PI}_A := R_A \cap (U^{PI} \times U^{PI})$.

Now, we can define a modified entailment relation that takes into account positive introspection:

Definition 15. *For $\varphi \in \mathcal{COL}$ and $\Gamma \subseteq \mathcal{COL}$, we write $\Gamma \models_{PI} \varphi$ if $\varphi^{\mathcal{K}^*_{PI},w} = \mathbf{t}$ for every world w such that $\psi^{\mathcal{K}^*_{PI},w} = \mathbf{t}$ for all $\psi \in \Gamma$.*

Note that incorporating positive introspection into the logic in this way modifies the meaning of the modalities M_A and O_A, because knowing at most φ means that any ψ that one knows must be entailed by φ and the entailment relation is different now that truthfulness and introspection are hard-coded into the logic: ψ may be entailed by φ together with truthfulness and introspection even though it was not entailed by φ itself in the original semantics of $\mathcal{COL}$ without truthfulness and introspection. In other words, when $\varphi \models_{PI} \psi$ but $\varphi \not\models \psi$, then $M_A\varphi \not\models_{PI} \neg K_A\varphi$ even though $M_A\varphi \models \neg K_A\varphi$.

The following theorem establishes that positive introspection does indeed hold in this logic and that the properties that we established for $\models$ in Theorem 5 also hold for $\models_{PI}$.

Theorem 6. *Let $\varphi, \psi \in \mathcal{COL}$ be two formulas, $\Gamma, \Gamma' \subset \mathcal{COL}$ two sets of formulas, $A \in \mathcal{A}$ an agent, and $G \subseteq \mathcal{A}$ a non-empty set of agents. Then, the following properties hold:*

1-10. All properties mentioned in Theorem 5 with $\models$ replaced by $\models_{PI}$.

11. (PI) $\models_{PI} K_A\varphi \Rightarrow K_AK_A\varphi$

Given that we could develop a theory with positive introspection, one may wonder what happens if one tries to similarly add negative introspection and/or truthfulness. Due to the problems with negative introspection described above, naively adding negative introspection in this way will yield a logic in which $M_A\varphi$ is not satisfiable for any φ. Adding truthfulness, on the other hand, does not cause such problems. Indeed, truthfulness and positive introspection can meaningfully be added together in a way similar to how positive introspection was added in the above definitions, yielding an entailment relation $\models_{TPI}$. Apart from (O), the properties of Theorem 6 will still hold. We conjecture that the following weakening of (O') holds in this context:

(O') If φ is a formula not involving a modality with subscript A, then $O_A\varphi \not\models_{TPI} \bot$.

The proof of this conjecture is left to future work.

6 Related Work

In this paper, we have studied the interplay of common knowledge and only knowing. The former concept is quite well-known and has been extensively studied (Fagin et al. 1995; Meyer and van der Hoek 1995) since its first mentions in the philosophical (Lewis 1969), and the mathematical literature (Aumann 1976). The latter concept is younger and has been studied intensively more recently.

Levesque (1990) was among the first to introduce the notion of only knowing[4] by presenting a single-agent logic of belief extended with a novel operator O expressing that the agent's beliefs are exactly the ones implied by the knowledge base and nothing more. He intended his logic of only knowing to capture certain types of non-monotonic reasoning patterns, like autoepistemic logic (AEL) (Moore 1985). In the 1990s and early 2000s, single-agent only knowing was successfully studied and implemented (Halpern and Lakemeyer 1995; Rosati 2000; Levesque and Lakemeyer 2000), and Lakemeyer and Levesque (2005) further revealed its potential by showing that it is also possible to capture default logic (DL) (Reiter 1980) and a variant of AEL proposed by Konolige (1988).

Halpern (1993) and Lakemeyer (1993) were among the first to extend Levesque's only knowing to a multi-agent setting. In a joint publication, they (2001) improved upon their independent works with an axiom system satisfying all the desired properties for only knowing. Nevertheless, the proposed axiomatization forces to include directly in the language a validity operator and the resulting semantics is not "as natural as we might like", according to the authors. In the early 2000s, first Waaler (2004) alone and then together with Solhaug (Waaler and Solhaug 2005) tried another route to generalize Levesque's axioms, without encoding the notion of validity into the language itself. Yet, once again these models and the logic itself feel complex and unnatural.

An in-depth analysis of the issues of all the previously mentioned works on multi-agent only knowing is provided by Belle and Lakemeyer (2010; 2015b), together with a natural way to avoid such problems. Belle and Lakemeyer proposed to use models limited to a finite depth k, i.e., models at which beliefs can be nested at most k times, by introducing the concept of k-structure to represent an agent's epistemic

[4]There are several closely-related notions, like ignorance (Halpern and Moses 1985; Konolige 1982), minimal knowledge (van der Hoek and Thijsse 2002), and total knowledge (Pratt-Hartmann 2000).

state. In this way, they successfully extend Levesque's logic, by keeping the original idea of Levesque's worlds and generalizing its features at the same time. However, they do not take into account common knowledge.

Aucher and Belle (2015) proposed a novel formulation comprising both common knowledge and only knowing at level k. To achieve this result, they make use of the concept of so-called $k{+}1$-canonical formulas (Moss 2007), written as conjunctions of a 0-canonical formula and k-canonical formulas nested in knowledge and common knowledge operators. Even though the proposed pointed epistemic models can be fully characterized up to modal depth k by a k-canonical formula, such representation at depth k might feel slightly unnatural, as it characterizes the knowledge of an agent only up to level k, but it also determines the common knowledge of a group of agents, which is infinitary in nature. When disregarding the conjunction of the common knowledge operators in such formulas, one can see a correspondence between our k-biworlds and the proposed k-canonical formulas. However, there is an important limitation to this work. Namely, while our O_A operator fully captures the knowledge and ignorance of an agent A at any depth, the O_A^n operator proposed by Aucher and Belle (2015) expresses only knowing for an agent A just up to level n in the sense that an expression of the form $O_A^1\varphi$ should be read as "If I disregard all my knowledge deeper than level one, I only know φ". Formally, as Aucher and Belle pointed out, for any two modal depths $n > m$, if agent A only knows φ at depth n, then A only knows φ at depth m, i.e., $O_A^n \to O_A^m$, but the reverse implication might not hold.

The limitation of the only knowing operator to a finite depth is overcome in the same year by Belle and Lakemeyer (2015a). They devised an alternative approach to bring common knowledge and only knowing together, the semantic structures of which are close to our ω-biworlds (except that there is no representation of the set $\bar{A}{\uparrow}\omega$). However, as explained in Remark 1, due to hard-coding negative introspection in their logic, all non-trivial formulas of the form $O_A\varphi$ are unsatisifiable, which goes against what we expressed as Property 8 in Theorem 5. Hence, such a semantics does not properly capture the intended meaning of the only-knowing operator.

Van Hertum (2016) observed that in (Belle and Lakemeyer 2015a), the formula $O_A\neg Cp$ cannot be satisfied, and this holds even if negative introspection is dropped (to avoid the problem mentioned in Remark 1). Van Hertum (2016) attempted to overcome this problem by proposing four different semantics, but he does not entirely fulfill his purpose. In more detail, two of the proposed semantics (Section 6.3 of (Van Hertum 2016)) do not allow for arbitrary nesting of the only knowing operator and thus they do not cover the whole $\mathcal{COL}$ language. A third semantics (Section 6.2.2 of (Van Hertum 2016)) seems to solve the problem regarding the satisfiability of formulas like $O_A\neg Cp$, but it is not precision-monotonic. This is rather problematic: in a given μ-world in such semantics, $O_A\varphi$ might be true, but if we give more precise information (extending it to depth $\mu + 1$), the formula might become false. Hence, one could only define what an agent only knows "up to a certain depth", as in (Aucher and Belle 2015). Finally, the semantics presented in Section 6.2.1 makes use of λ-canonical Kripke structures, for any fixed limit ordinal λ, to define a two-valued valuation for $\mathcal{COL}$. In particular, if $\lambda = \omega$, then the proposed semantics corresponds to the semantics of Belle and Lakemeyer (2015a) with some minor adjustments. The hope of the author was that by choosing a large enough limit ordinal λ, the satisfiability issue would be solved. Unfortunately, there exist formulas that are not satisfiable in any λ-canonical Kripke structure, for example $O_A\neg K_A\bot$, which formalizes the statement "all A knows is that their knowledge is consistent". Even though this formula may seem like a corner case, all such corner cases need to be avoided in order to have a well-behaved entailment relation that satisfies the desirable properties listed in Theorem 5. The unsatisfiability of $O_A\neg K_A\bot$ implies that property (O) is not satisfied, which means that the semantics does not properly capture the intended meaning of the only-knowing operator.

7 Conclusion

We defined a multi-agent epistemic logic with common knowledge and only knowing operators, which successfully encodes these notions under the same framework.

First, we introduced the novel concept of μ-biworld for countable ordinals μ, which approximates not only the worlds that an agent deems possible, but also those deemed impossible. This duality proved to be fundamental to successfully deal with the only knowing operator in a multi-agent setting. Moreover, we have shown that the proposed new definitions are indeed sensible, as they satisfy the properties one would expect (Theorem 1).

Second, we defined the language $\mathcal{COL}$, extending propositional logic with the modal operators K_A, M_A, E_G, and C_G, and a three-valued model semantics for it. Furthermore, we defined a canonical Kripke structure over completed $\omega^2{+}1$-biworlds, and the two-valued semantics obtained from it is shown to coincide with the model semantics. This allowed us to prove several desirable properties (Theorem 5) the resulting logic satisfies. In particular, we showed that for any formula φ, there is a unique state-of-mind of a given agent in which they only know φ.

Finally, we have considered how our framework can be extended to satisfy properties like truthfulness, positive introspection and negative introspection. For positive introspection we have shown some positive results, whereas for truthfulness and negative introspection, we have identified certain problems that arise when combining them with only knowing. We have motivated the need for further research related to these problems.

Another line of future work that we envisage is to generalize the construction of biworlds so that it becomes applicable to other areas of research. More concretely, this amounts to the construction of a set-theoretic universe in which there exists a universal set, similarly as in topological set theory. If one takes this alternative set theory as one's metatheory, the incorrect definition from the introduction of this paper could very easily be turned into a correct definition.

Acknowledgements

We are grateful to Marc Denecker and Pieter Van Hertum for the fruitful discussions and feedback on earlier versions of this work. This work was partially supported by Fonds Wetenschappelijk Onderzoek – Vlaanderen (project G0B2221N) and by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen"

References

Aucher, G., and Belle, V. 2015. Multi-agent only knowing on planet Kripke. In Yang, Q., and Wooldridge, M. J., eds., *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2713–2719. AAAI Press.

Aumann, R. J. 1976. Agreeing to disagree. *The Annals of Statistics* 4(6):1236–1239.

Belle, V., and Lakemeyer, G. 2010. Multi-agent only-knowing revisited. *CoRR* abs/1009.2041.

Belle, V., and Lakemeyer, G. 2015a. Only knowing meets common knowledge. In Yang, Q., and Wooldridge, M., eds., *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2755–2761. AAAI Press.

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

Cramer, M.; Pollaci, S.; and Bogaerts, B. 2023. Mathematical foundations for joining only knowing and common knowledge (extended version). *CoRR* abs/2306.03267.

Fagin, R.; Halpern, J. Y.; Moses, Y.; and Vardi, M. Y. 1995. *Reasoning About Knowledge*. MIT Press.

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

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

Halpern, J. Y., and Lakemeyer, G. 2001. Multi-agent only knowing. *J. Log. Comput.* 11(1):41–70.

Halpern, J. Y., and Moses, Y. 1985. Towards a theory of knowledge and ignorance: Preliminary report. In Apt, K. R., ed., *Logics and Models of Concurrent Systems*, volume 13 of *NATO ASI Series*, 459–476. Springer Berlin Heidelberg.

Halpern, J. Y. 1993. Reasoning about only knowing with many agents. In Fikes, R., and Lehnert, W. G., eds., *Proceedings of the 11th National Conference on Artificial Intelligence. Washington, DC, USA, July 11-15, 1993*, 655–661. AAAI Press / The MIT Press.

Hintikka, J. 1962. *Knowledge and Belief*. Ithaca: Cornell University Press.

Konolige, K. 1982. Circumscriptive ignorance. In Waltz, D. L., ed., *Proceedings of the National Conference on Artificial Intelligence, Pittsburgh, PA, USA, August 18-20, 1982*, 202–204. AAAI Press.

Konolige, K. 1988. On the relation between default and autoepistemic logic. *Artif. Intell.* 35(3):343–382.

Lakemeyer, G., and Levesque, H. J. 2005. Only-knowing: Taking it beyond autoepistemic reasoning. In Veloso, M. M., and Kambhampati, S., eds., *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, 633–638. AAAI Press / The MIT Press.

Lakemeyer, G. 1993. All they know: A study in multi-agent autoepistemic reasoning. In Bajcsy, R., ed., *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 - September 3, 1993*, 376–381. Morgan Kaufmann.

Levesque, H. J., and Lakemeyer, G. 2000. *The logic of knowledge bases*. MIT Press.

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

Lewis, D. K. 1969. *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell.

Meyer, J. C., and van der Hoek, W. 1995. *Epistemic logic for AI and computer science*, volume 41 of *Cambridge tracts in theoretical computer science*. Cambridge University Press.

Moore, R. C. 1985. Semantical considerations on nonmonotonic logic. *Artif. Intell.* 25(1):75–94.

Moss, L. S. 2007. Finite models constructed from canonical formulas. *J. Philos. Log.* 36(6):605–640.

Permpoontanalarp, Y., and Jiang, J. Y. 1995. On multi-agent autoepistemic reasoning. In *WOCFAI*, 307–318.

Pratt-Hartmann, I. 2000. Total knowledge. In Kautz, H. A., and Porter, B. W., eds., *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, July 30 - August 3, 2000, Austin, Texas, USA*, 423–428. AAAI Press / The MIT Press.

Reiter, R. 1980. A logic for default reasoning. *Artif. Intell.* 13(1-2):81–132.

Rosati, R. 2000. On the decidability and complexity of reasoning about only knowing. *Artif. Intell.* 116(1-2):193–215.

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

Van Hertum, P.; Cramer, M.; Bogaerts, B.; and Denecker, M. 2016. Distributed autoepistemic logic and its application to access control. In Kambhampati, S., ed., *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 1286–1292. IJCAI/AAAI Press.

Van Hertum, P. 2016. *New Language Constructs and Inferences for the Knowledge Base Paradigm: A Business and Multi-agent Perspective ; Taaluitbreidingen en nieuwe inferenties voor het kennisbank paradigma vanuit een Business en Multi-Agent perspectief*. Ph.D. Dissertation, Katholieke Universiteit Leuven, Belgium.

Vlaeminck, H.; Vennekens, J.; Bruynooghe, M.; and Denecker, M. 2012. Ordered Epistemic Logic: Semantics,

complexity and applications. In Brewka, G.; Eiter, T.; and McIlraith, S. A., eds., *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Knowledge Representation and Reasoning, Rome, 10-14 July 2012*, 369–379. AAAI Press.

Waaler, A., and Solhaug, B. 2005. Semantics for multiagent only knowing: extended abstract. In van der Meyden, R., ed., *Proceedings of the 10th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2005), Singapore, June 10-12, 2005*, 109–125. National University of Singapore.

Waaler, A. 2004. Consistency proofs for systems of multiagent only knowing. In Schmidt, R. A.; Pratt-Hartmann, I.; Reynolds, M.; and Wansing, H., eds., *Advances in Modal Logic 5, papers from the fifth conference on "Advances in Modal logic," held in Manchester, UK, 9-11 September 2004*, 347–366. King's College Publications.