Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning
Main Track

# A Logic of Limited Belief with Introspection Based on Possible Worlds

**Gerhard Lakemeyer**[1] , **Hector J. Levesque**[2]
[1]Faculty of Computer Science, RWTH Aachen University, Germany
[2]Department of Computer Science, University of Toronto, Canada
gerhard@cs.rwth-aachen.de, hector@cs.toronto.edu

**Abstract**

The starting point of this paper is earlier work, where we proposed an epistemic logic which characterizes the beliefs of a knowledge-based agent in terms of increasing levels of complexity. At the lowest level, the agent is only able to draw simple conclusions from its knowledge base. Higher levels lead to more and more inferences and computing the beliefs at any particular level turns out to be tractable. What makes this logic arguably appealing is the fact that the underlying semantics is based on possible worlds. However, the work is still limited in that only beliefs about what is true in the world are considered, that is, an agent's beliefs about its own beliefs are ignored. In this paper we will close this gap and generalize the earlier work by proposing a model of limited belief where an agent is able to fully introspect on its own beliefs without sacrificing tractability.

## 1 Introduction

A fundamental problem in knowledge representation and reasoning (KR&R) has been that expressive languages for symbolic representations of knowledge and belief tend to lead to intractable reasoning procedures. As a result much of the work in KR&R has focused on limiting the representation language in some form or another, with description logics perhaps being the most prominent example (Baader *et al.* 2007). Starting with the work by Levesque (1984), a much smaller community has been concerned with devising models of belief, which capture forms of tractable reasoning without sacrificing the expressiveness of the representation language. In (Lakemeyer and Levesque 2020), we proposed an epistemic logic where the beliefs of a knowledge-based agent are characterized in terms of levels of increasing complexity. At the lowest level, the agent is only able to draw simple conclusions from its knowledge base. Higher levels lead to more and more inferences and are harder to compute but tractable. What makes this logic arguably appealing is the fact that the underlying semantics is based on possible worlds. While classical (2-valued) possible-world semantics (Kripke 1959; Hintikka 1962) is known to suffer from logical omniscience (Hintikka 1975), which makes reasoning intractable, we used a 3-valued variant, which does not suffer from this problem. Moreover, we show a close connection between the semantics of belief levels and a limited use of Resolution. However, the work is still limited in that only beliefs about what is true in the world are considered, that is, an agent's beliefs about its own beliefs are ignored. In this paper we will close this gap in the propositional case and propose a model of limited belief where an agent is able to fully introspect on its own beliefs without sacrificing tractability.

The rest of the paper is organized as follows. The next section discusses the approach in informal terms, followed by a formal introduction of the propositional fragment of our earlier logic in Section 3. Section 4 generalizes the logic to account for nested beliefs. Section 5 discusses properties of the logic. In Section 6, we briefly study a variant of the logic, where beliefs are required to be consistent. Section 7 develops the main tractability result. Section 8 discusses related work. Then we conclude.

## 2 The Approach

We begin by introducing the main ideas behind our earlier model of limited belief. There we consider a very expressive first-order language with equality. Handling quantifiers in this logic involves introducing fairly heavy technical machinery, including the use of both Skolemization and its dual in the semantics. We avoid these complications here by focusing on the propositional fragment of the logic. When stating results of our earlier logic, they are meant to apply to the propositional fragment of the logic, even though many of them apply to the first-order version as well. Belief is modeled using modal operators $\boldsymbol{B}_0, \boldsymbol{B}_1, \boldsymbol{B}_2, \ldots$ and $\boldsymbol{O}$. $\boldsymbol{B}_k\phi$ may be read as "$\phi$ is believed at level $k$." In the semantics, an epistemic state $e$ is a set of three-valued possible worlds. The beliefs at level 0 are then simply the formulas which are true at all worlds in $e$. Beliefs at higher levels are interpreted with respect to appropriate subsets of $e$. (See the next section for details about how these subsets are formed.) The logic also includes the modal operator $\boldsymbol{O}$ where $\boldsymbol{O}\phi$ may be read as "$\phi$ is only-believed." It is used mainly to characterize the beliefs of a knowledge base (KB), which is a finite collection of formulas understood conjunctively. With only-believing, the beliefs of a KB at any level can be expressed in terms of formulas of the form $(\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$, which may be read as "if all the agent believes are the sentences in KB, then it believes $\phi$ at level $k$." In the following theorem, we write $\models \alpha$ to mean that the formula $\alpha$ is valid (see the next section for formal definitions).

**Theorem 1** (Lakemeyer and Levesque 2020)**.**

- *expressiveness: If* KB $= \bigwedge \phi_i$ *then* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_0\phi_i)$ *for all* $i$*;*
- *cumulativity:* $\models (\boldsymbol{B}_k\phi \supset \boldsymbol{B}_{k+1}\phi)$ *for all* $k$*;*
- *soundness: if* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$ *then* $\models (\text{KB} \supset \phi)$*;*
- *eventual completeness: if* $\models (\text{KB} \supset \phi)$ *then there is a* $k$ *such that* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$*;*
- *tractability: if* KB $= \bigwedge \phi_i$ *and every* $\phi_i$ *is of constant size compared to the size of* KB*, then computing whether* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$ *holds is polynomial in data complexity.*

Expressiveness means that every sentence in the KB is already believed at level 0. Cumulativity means that whatever is believed at level $k$ will also be believed at any higher level. Besides being sound, belief is also eventually complete, that is, whatever follows classically from the knowledge base is believed at some level. Finally, computing the beliefs $\phi$ of a KB at any level $k$ is tractable in data complexity, that is, in the size of the KB only, considering $k$ and the size of $\phi$ to be constant. The assumption that every $\phi_i$ in KB and $\phi$ are small is needed, as the computation requires that these formulas are converted into conjunctive normal form (see Section 3.2 for details.)

Let us now turn to the case where we allow beliefs to be nested as in $\boldsymbol{B}_1(p \wedge \neg\boldsymbol{B}_0 p)$, that is, the agent believes at level 1 that $p$ is true and that it does not believe $p$ at level 0. What kinds of properties do we want when we allow such nested beliefs? To start with, as an agent's beliefs are fully determined by its knowledge base, it seems reasonable to assume that the agent is fully introspective, just as in classical accounts of belief like (Hintikka 1962; Halpern and Moses 1992), that is, both $(\boldsymbol{B}_k\alpha \supset \boldsymbol{B}_k\boldsymbol{B}_k\alpha)$ (positive introspection) and $(\neg\boldsymbol{B}_k\alpha \supset \boldsymbol{B}_k\neg\boldsymbol{B}_k\alpha)$ (negative introspection) should come out valid. As we will see, introspection can easily be obtained semantically by keeping the idea of interpreting belief wrt. a set of possible worlds.

When dealing with levels of belief, perhaps the most interesting aspect of nested beliefs is the interaction of different levels of belief. For starters, we believe that introspection should carry over to the vantage point of higher belief levels. For example, both $(\boldsymbol{B}_0 p \supset \boldsymbol{B}_5\boldsymbol{B}_0 p)$ and $(\neg\boldsymbol{B}_0 p \supset \boldsymbol{B}_5\neg\boldsymbol{B}_0 p)$ should be valid. This means that the beliefs at level 5 should include whatever is believed and not believed at lower levels. In other words, if the agent is able to figure out at level 0 what it believes and does not believe at that level, then surely it should continue to hold these lower level beliefs at higher levels as no extra effort is involved.

Assuming that cumulativity ($\models (\boldsymbol{B}_k\alpha \supset \boldsymbol{B}_{k+1}\alpha)$) continues to hold, it seems desirable to also have $\models (\boldsymbol{B}_k\alpha \supset \boldsymbol{B}_k\boldsymbol{B}_{k+1}\alpha)$, that is, if an agent is able to figure out that $\alpha$ holds at level $k$, then it should believe at level $k$ that it will continue to believe $\alpha$ at level $k+1$, or any other higher level for that matter.

Now suppose that an agent does not believe $\alpha$ at level $k$. What should it believe at level $k$ about believing $\alpha$ at level $k+1$ or higher? Our answer is: nothing. This is because if the agent does not believe $\alpha$ at level $k$, then it may be the case that it believes $\alpha$ at a higher level, but it may also be the case that $\alpha$ is never believed no matter the level. The intuition is that when reasoning is limited to level $k$, then the agent cannot figure out whether or not $\alpha$ holds at a higher level without doing the extra work, and this cannot be done at level $k$. Hence, for example, we would like that both $(\neg\boldsymbol{B}_k\alpha \supset \neg\boldsymbol{B}_k\boldsymbol{B}_{k+1}\alpha)$ and $(\neg\boldsymbol{B}_k\alpha \supset \neg\boldsymbol{B}_k\neg\boldsymbol{B}_{k+1}\alpha)$ come out valid. As we will see later, this aspect is the most demanding when it comes to devising an appropriate model of belief.

When considering the beliefs of a KB, we will continue to assume that the KB itself is objective, that is, does not refer to its own beliefs.[1] With that we can characterize the beliefs of a knowledge base at level $k$ as the valid sentences of the form $(\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\alpha)$ for arbitrary $\alpha$. In particular, we will show that computing whether $(\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\alpha)$ is valid is again tractable (under reasonable assumptions).

## 3 The Logic $\mathcal{LB}$

In this section we formally introduce the propositional fragment of our earlier logic of limited belief, which we call $\mathcal{LB}$. Given an infinite set of atomic propositions, also called atoms for short, formulas, which we sometimes also call sentences, are formed in the usual way using the Boolean operators $\neg$ and $\vee$ as well as modal operators $\boldsymbol{B}_k$ for all natural numbers $k \geq 0$ and $\boldsymbol{O}$ with the restriction that modal operators may not be nested.

Other Boolean connectives $\wedge, \supset, \equiv$ are used as abbreviations. A literal is either an atom or its negation. A formula is called *objective* if it does not mention any $\boldsymbol{B}_k$ and $\boldsymbol{O}$. It is called *subjective* if every atom occurs within the scope of a $\boldsymbol{B}_k$ for some $k$ or $\boldsymbol{O}$. A formula of the form $\boldsymbol{B}_k\phi$ or $\boldsymbol{O}\phi$ is sometimes called a *belief formula*. As a convention, we will denote objective formulas by $\phi$ and $\psi$ and arbitrary formulas by $\alpha$ and $\beta$. We also write $\boldsymbol{O}$KB to say that the (finite) conjunction of objective formulas in the knowledge base are all that is believed.

### 3.1 Semantics

The semantics is based on possible worlds, which come in two flavours, *standard* worlds and *extended* worlds. Standard worlds are mappings from the set of atoms to $\{0, 1\}$, where 1 stands for *true* and 0 for *false*. Extended worlds are mappings from the set of atoms to $\{0, 1, *\}$. When an atom is assigned the value $*$, it means that there is support for both the truth and the falsity of $p$. Note that standard worlds are just a special case of extended worlds. When the context is clear, we sometimes call extended worlds simply worlds.

We begin by defining how extended worlds assign meaning to objective formulas. We write $w \models_{\text{T}} \phi$ when $w$ supports the truth of $\phi$ and $w \models_{\text{F}} \phi$ when it supports its falsity.

**Definition 1** (True- and false-support for extended worlds)**.** *Let $w$ be an extended world.*

1. $w \models_{\text{T}} p$ *iff* $w[p] \neq 0$*;*
   $w \models_{\text{F}} p$ *iff* $w[p] \neq 1$.

[1] As shown in (Levesque and Lakemeyer 2001), allowing a KB to refer to its own beliefs can be useful in order to model defaults, but that is beyond the scope of this paper.

2. $w \models_{\mathrm{T}} \neg\phi$ *iff* $w \models_{\mathrm{F}} \phi$;
   $w \models_{\mathrm{F}} \neg\phi$ *iff* $w \models_{\mathrm{T}} \phi$.
3. $w \models_{\mathrm{T}} (\phi \vee \psi)$ *iff* $w \models_{\mathrm{T}} \phi$ *or* $w \models_{\mathrm{T}} \psi$;
   $w \models_{\mathrm{F}} (\phi \vee \psi)$ *iff* $w \models_{\mathrm{F}} \phi$ *and* $w \models_{\mathrm{F}} \psi$.

An *epistemic state*, often denoted as $e$, is a set of extended worlds.

For ease of presentation, we define the semantics of arbitrary formulas in stages. We begin by only considering formulas with modal operators $\boldsymbol{B}_0$ and $\boldsymbol{O}$.

Outside of belief formulas, the logic is classical and we define the satisfaction of formulas in terms of a standard world $w$ and an epistemic state $e$, written $e, w \models \alpha$, as follows:

**Definition 2** (Satisfaction)**.**
*Let $w$ be a standard world and $e$ an epistemic state.*

1. $e, w \models p$ *iff* $w[p] = 1$ *for atom $p$;*
2. $e, w \models \neg\alpha$ *iff* $e, w \not\models \alpha$;
3. $e, w \models (\alpha \vee \beta)$ *iff* $e, w \models \alpha$ *or* $e, w \models \beta$;
4. $e, w \models \boldsymbol{B}_0\phi$ *iff for all* $w \in e$, $w \models_{\mathrm{T}} \phi$;
5. $e, w \models \boldsymbol{O}\phi$ *iff for all* $w$, $w \in e$ *iff* $w \models_{\mathrm{T}} \phi$.

When a formula $\alpha$ is subjective, we often write $e \models \alpha$ instead of $e, w \models \alpha$. Similarly, we write $w \models \phi$ for objective $\phi$. According to the semantics, a sentence is believed in an epistemic state $e$ at level 0 just in case it has true-support in all worlds in $e$. As the following example shows, beliefs at level 0 are indeed very limited in general.

**Example 1.** *Let $w$ be a world and $p$ and $q$ atoms with $w[p] = *$ and $w[q] = 0$. Let $\phi = (p \vee q) \wedge (\neg p \vee q)$ and $e = \{w\}$. Then $e \models \boldsymbol{B}_0\phi$ because both $p$ and $\neg p$ have true-support at $w$, yet $e \not\models \boldsymbol{B}_0 q$ even though $\phi$ entails $q$ in classical logic.*

Also note the difference between believing at level 0 and only-believing. It is easy to see that $\boldsymbol{O}\phi$ is uniquely satisfied by $e = \{w \mid w \models \phi\}$, which we also denote as $\Re[\![\phi]\!]$. Given distinct atoms $p$ and $q$, this means that $\boldsymbol{O}p$ entails $\neg\boldsymbol{B}_0 q$, yet $\boldsymbol{B}_0 p$ does not entail $\neg\boldsymbol{B}_0 q$ since $e^* \models \boldsymbol{B}_0 p \wedge \boldsymbol{B}_0 q$ with $e^* = \{w \mid w \models (p \wedge q)\}$. In other words, $\boldsymbol{B}_0\phi$ should really be read as "the agent believes at least $\phi$ at level 0" (and perhaps more) while $\boldsymbol{O}\phi$ means that "the agent believes exactly $\phi$" (and no more). Being able to draw conclusions about what an agent does not believe is perhaps the main reason why only-believing is important. This feature will become even more relevant later when we consider beliefs about beliefs, in particular, introspection.

Let us now turn to the meaning of belief formulas $\boldsymbol{B}_k\phi$ for arbitrary $k$. As a preliminary step, we need the following definitions:

**Definition 3** (Unsupported literals)**.**
$U(w) = \{p \mid w[p] = 0\} \cup \{\neg p \mid w[p] = 1\}$.

The unsupported literals of $w$ are those that $w$ says cannot be true.

**Definition 4** (Eliminated world)**.**
*$e$ eliminates world $w$ iff there is an atom $p$ such that for every world $w' \in e$, if $U(w) \subseteq U(w')$, then $w'[p] = *$.*

Intuitively, $e$ eliminates $w$ if there is some $p$ such that the claims made by $w$ (in terms of what cannot be true) depend on $p$ having value $*$. In other words, if we only kept worlds in $e$ where $p$ had value 0 or 1, no worlds would support the claims made by $w$. See Example 2 below for an illustration.

**Definition 5** (Successor epistemic state)**.**
$S(e) = e \setminus \{w \mid e$ *eliminates* $w\}$.

To apply the function $S$ more than once, let $S^0(e) = e$ and $S^k(e) = S(S^{k-1}(e))$ for $k > 0$. In the following we will sometimes also refer to the sets $S^k(e)$ as *successor epistemic states.* Note that if $w$ is standard, it is never eliminated since for no $p$ do we have $w[p] = *$. In other words, if $w \in e$ and $w$ is standard, then for every $k$, $w \in S^k(e)$.

With the definition of $S^k(e)$ in hand, we can now give meaning to belief formulas $\boldsymbol{B}_k\phi$ for arbitrary $k$. All that needs to be done is to replace rule (4) in Definition 2 by this:

4. $e, w \models \boldsymbol{B}_k\phi$ iff for all $w' \in S^k(e)$, $w' \models_{\mathrm{T}} \phi$.

To conclude the semantics, a formula $\alpha$ is valid ($\models \alpha$) if for all standard worlds $w$ and epistemic states $e$, $e, w \models \alpha$.[2]

**Example 2.** *Let $e = \{w_{01}, w_{11}, w_{*0}, w_{**}\}$, where $w_{01}[p] = 0$ and $w_{01}[q] = 1$, and similarly for the others. We have $U(w_{01}) = \{p, \neg q\}$, $U(w_{11}) = \{\neg p, \neg q\}$, $U(w_{*0}) = \{q\}$, and $U(w_{**}) = \{\}$. Then $e$ eliminates $w_{*0}$ because $w_{*0}[p] = *$ and there is no other world $w' \in e$ such that $U(w_{*0}) \subseteq U(w')$. Indeed, $w_{*0}$ is the only world eliminated by $e$, that is, $S(e) = \{w_{01}, w_{11}, w_{**}\}$. Then we have the following:*

$e \models \boldsymbol{B}_0((p \vee q) \wedge (\neg p \vee q))$
$e \not\models \boldsymbol{B}_0 q$ *but*
$e \models \boldsymbol{B}_1 q$ *since for all* $w \in S(e)$, $w \models_{\mathrm{T}} q$.

## 3.2 Some Properties of $\mathcal{LB}$

Here we only discuss some of the main properties of $\mathcal{LB}$. Similar to classical epistemic logic we have

- $\models (\boldsymbol{B}_k\phi \vee \boldsymbol{B}_k\psi \supset \boldsymbol{B}_k(\phi \vee \psi))$;
- $\models (\boldsymbol{B}_k\phi \wedge \boldsymbol{B}_k\psi \equiv \boldsymbol{B}_k(\phi \wedge \psi))$.

Both properties are easily obtained because belief at level $k$ for a given epistemic state $e$ is defined as true-support in all worlds in $S^k(e)$. (We remark that in the first-order case, $\models \boldsymbol{B}_k\phi \wedge \boldsymbol{B}_k\psi \supset \boldsymbol{B}_k(\phi \wedge \psi)$ only holds for formulas without quantifiers.) Furthermore, for the same reasons as above transformation rules like De Morgan's laws or distribution of $\vee$ over $\wedge$ are applicable at any level. In particular, we get

- $\models \boldsymbol{B}_k\phi \equiv \boldsymbol{B}_k\phi_{\mathrm{CNF}}$, where $\phi_{\mathrm{CNF}}$ is $\phi$ converted to conjunctive normal form (CNF).

The main properties of the logic were already stated in Theorem 1 of Section 2. It is instructive to take a closer look at how the beliefs of a KB at level $k$ are computed. As usual, a *clause* is meant to be a set of literals. For an objective formula $\phi$ in CNF, let the clausal form of $\phi$, denoted

[2] We remark that in (Lakemeyer and Levesque 2020) we defined validity only for representable epistemic states, where $e = \Re[\![\phi]\!]$ for some $\phi$. This was needed because of complications arising from Skolemization, which is of no concern in the propositional case.

as $CF(\phi)$, consist of all clauses obtained from the conjunctions in $\phi$. For example, if $\phi = (p \vee \neg q \vee r) \wedge (s \vee \neg p \vee q)$ then $CF(\phi) = \{\{p, \neg q, r\}, \{s, \neg p, q\}\}$.

**Definition 6.** *For any set of clauses $C$, $RP(C)$ is the set of clauses defined by*

$$RP(C) = C \cup \{(b \cup d) \mid \text{for some } p, (\{p\} \cup b) \in C, (\{\neg p\} \cup d) \in C\}.$$

$RP(C)$ simply applies one step of Resolution to $C$. We can iterate this by letting $RP^0(C) = C$ and $RP^k(C) = RP(RP^{k-1}(C))$ for $k > 0$. We then obtained the following tight connection between $S^k(e)$ and Resolution:

**Theorem 2** (Lakemeyer and Levesque 2020)**.** *Let $C$ be a set of clauses. Let $e = \{w \mid w \models_{\text{T}} C\}$.*
*Then $S^k(e) = \{w \mid w \models_{\text{T}} RP^k(C)\}$.*

Here $w \models_{\text{T}} C$ means that $w \models_{\text{T}} c$ for all $c \in C$, where $c$ is understood as a disjunction of literals. The theorem essentially tells us that the elimination of worlds as defined by $S^k(e)$ is a semantic version of applying $k$ steps of Resolution to the clauses that $e$ believes at level 0. With that, it is not hard to show how to compute the beliefs of a KB at level $k$.

**Theorem 3** (Lakemeyer and Levesque 2020)**.** *Suppose both* KB *and $\phi$ are in CNF. Then $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$ iff for every clause $c \in CF(\phi)$, either $c$ contains complementary literals or there is a clause $c'$ in $RP^k(CF(\text{KB}))$ such that $c' \subseteq c$.*

Tractability then obtains by assuming that each sentence in a KB and $\phi$ are small and $k$ is considered a constant.

**Corollary 1** (Lakemeyer and Levesque 2020)**.** *Let* KB *and $\phi$ be as above, and assume every conjunct in* KB *and $\phi$ to be of constant size. Then for every $k$, computing whether $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$ is polynomial in the size of* KB.

The tractability result follows directly from the above theorem given that $RP^k(CF(\text{KB}))$ is of polynomial size when $k$ is considered to be a constant.

While a KB may in general be arbitrarily large, it seems reasonable to assume that each sentence in the KB as well as the query are very small compared to the size of the whole KB and can be reasonably considered to be of constant size. This way the transformation of the sentences in the KB and the query into CNF, which can lead to an exponential blowup, is also not problematic as the conversion is applied to sentences of constant size only.

# 4 The Logic $\mathcal{LBI}$: $\mathcal{LB}$ plus Introspection

Let us now turn to our logic with nested beliefs, which we call $\mathcal{LBI}$. The language is essentially like the one for $\mathcal{LB}$ except that for any formula $\boldsymbol{B}_k\alpha$, $\alpha$ may now contain subformulas $\boldsymbol{B}_j\beta$ for any $j$. For example, $\boldsymbol{B}_1(p \wedge \neg \boldsymbol{B}_0 p \wedge \boldsymbol{B}_4 \boldsymbol{B}_1 p)$ is now a well-formed formula. We continue to only consider $\boldsymbol{O}\phi$ where $\phi$ is objective and we do not allow $\boldsymbol{O}$ to occur within the scope of a $\boldsymbol{B}_k$. Formulas which do not mention $\boldsymbol{O}$ are called *basic*. Finally, it is convenient to add to the language two special atoms $\top$ and $\bot$ for truth and falsity, respectively. Atoms other than $\top$ and $\bot$ are also called *normal* atoms.

## 4.1 The Semantics

Standard and extended worlds are defined exactly the same way as in $\mathcal{LB}$ with the addition that every world $w$ maps $\top$ to 1 and $\bot$ to 0. For a world $w$ and objective formula $\phi$, $w \models_{\text{T}} \phi$ and $w \models_{\text{F}} \phi$ are defined exactly as in Definition 1.

As before, an epistemic state $e$ is any set of extended worlds, including the empty set.

Similar to Definition 2 the logic is classical (two-valued) outside of belief and satisfaction is defined for standard worlds only and epistemic states:

**Definition 7** (Satisfaction)**.**
*Let $w$ be a standard world and $e$ an epistemic state. Then the satisfaction relation $e, w \models \alpha$ is inductively defined as follows:*

1. $e, w \models p$ *iff* $w[p] = 1$ *for atom $p$;*
2. $e, w \models \neg\alpha$ *iff* $e, w \not\models \alpha$*;*
3. $e, w \models (\alpha \vee \beta)$ *iff* $e, w \models \alpha$ *or* $e, w \models \beta$*;*
4. $e, w \models \boldsymbol{B}_k\alpha$ *iff* $e, k, w \models_{\text{T}} \boldsymbol{B}_k\alpha$ *(as defined below).*
5. $e, w \models \boldsymbol{O}\phi$ *iff for all $w$, $w \in e$ iff $w \models_{\text{T}} \phi$.*

Note that the semantics of $\boldsymbol{O}\phi$ is exactly as in $\mathcal{LB}$ as $\phi$ is assumed to be objective. Let us now turn to how (possibly nested) beliefs are interpreted in terms of $e, k, w \models_{\text{T}} \boldsymbol{B}_k\alpha$ and $e, k, w \models_{\text{F}} \boldsymbol{B}_k\alpha$. Notice that the support relations now explicitly mention $k$ on the LHS. As we will see, this is needed as the interpretation of beliefs inside another belief depends on what the outermost belief level is.

In the following the definitions of unsupported literals (Definition 3), the elimination of worlds (Definition 4) and successor epistemic states $S(e)$ (Definition 5) carry over as is.

**Definition 8** (Truth and falsity support within belief)**.**
*Let $w$ be an extended world, $e$ an epistemic state, and $k \geq 0$. Then*

1. $e, k, w \models_{\text{T}} p$ *iff* $w[p] \neq 0$*;*
   $e, k, w \models_{\text{F}} p$ *iff* $w[p] \neq 1$.
2. $e, k, w \models_{\text{T}} \neg\alpha$ *iff* $e, k, w \models_{\text{F}} \alpha$*;*
   $e, k, w \models_{\text{F}} \neg\alpha$ *iff* $e, k, w \models_{\text{T}} \alpha$.
3. $e, k, w \models_{\text{T}} (\alpha \vee \beta)$ *iff* $e, k, w \models_{\text{T}} \alpha$ *or* $e, k, w \models_{\text{T}} \beta$*;*
   $e, k, w \models_{\text{F}} (\alpha \vee \beta)$ *iff* $e, k, w \models_{\text{F}} \alpha$ *and* $e, k, w \models_{\text{F}} \beta$.
4. $e, k, w \models_{\text{T}} \boldsymbol{B}_j\alpha$ *iff*
   *if $j \leq k$ then for all $w' \in S^j(e)$, $e, k, w' \models_{\text{T}} \alpha$,*
   *o.w. for all $e' \subseteq S^k(e)$ for all $w' \in e'$, $e, k, w' \models_{\text{T}} \alpha$;*
5. $e, k, w \models_{\text{F}} \boldsymbol{B}_j\alpha$ *iff*
   *if $j \leq k$ then for some $w' \in S^j(e)$, $e, k, w' \not\models_{\text{T}} \alpha$,*
   *o.w. for all $e' \subseteq S^k(e)$ for some $w' \in e'$, $e, k, w' \not\models_{\text{T}} \alpha$;*

Rules (1)–(3) correspond exactly to those of Definition 1, where we defined the true- and false-support for extended worlds. The more interesting parts are Rules (4) and (5) for belief. Notice that in both cases the interpretation is different depending on whether $j \leq k$ or $j > k$. If $j \leq k$ then the semantics mimics exactly what we saw in the logic $\mathcal{LB}$, that is, $\alpha$ is believed at level $j$ just in case $\alpha$ has true-support in all worlds in $S^j(e)$. The intuition is that belief levels that are no bigger than the outermost belief level can be treated in the "usual" way. For example, $e \models \boldsymbol{B}_2\boldsymbol{B}_1 p$ holds just in

case $e \models \boldsymbol{B}_1 p$ holds, and $e \models \boldsymbol{B}_2 \neg \boldsymbol{B}_1 p$ holds just in case $e \models \neg \boldsymbol{B}_1 p$ holds. We will look at examples in more detail below. The case when $j > k$ is quite different. The beliefs at level $j$ are determined by the worlds in $S^j(e)$. However, this would require reasoning effort beyond what can be done at level $k$. The intuition behind the semantics is that the agent has no idea what $S^j(e)$ looks except that it must be a subset of $S^k(e)$. This is why we consider all possible subsets $e'$ of $e$ when determining whether $\boldsymbol{B}_j \alpha$ has true- or false-support. As later proofs show the agent never actually has to consider all subsets. It suffices to look only at two edge cases, namely $e' = S^k(e)$ and $e' = \{\}$.

The reader may have noticed that Rule (4) can actually be simplified as follows:

4. $e, k, w \models_{\mathrm{T}} \boldsymbol{B}_j \alpha$ iff
   if $j \leq k$ then for all $w' \in S^j(e),\ e, k, w' \models_{\mathrm{T}} \alpha$,
   o.w. $e, k, w \models_{\mathrm{T}} \boldsymbol{B}_k \alpha$.

In other words, if $j > k$, then $\alpha$ is believed at level $j$ just in case it is already believed at level $k$. We still stick to the somewhat redundant original definition to be more aligned with the otherwise-case of Rule (5).

To complete the semantics, a formula $\alpha$ is valid ($\models \alpha$) if for all standard worlds $w$ and epistemic states $e$, $e, w \models \alpha$.

If $\alpha$ is subjective, we sometimes write $e \models \alpha$ instead of $e, w \models \alpha$ and $e, k \models_{\mathrm{T}} \alpha$ ($e, k \models_{\mathrm{F}} \alpha$) instead of $e, k, w \models_{\mathrm{T}} \alpha$ ($e, k, w \models_{\mathrm{F}} \alpha$). For objective $\phi$, we continue to write $w \models \phi$ instead of $e, w \models \phi$ and also write $w \models_{\mathrm{T}} \phi$ ($w \models_{\mathrm{F}} \phi$) instead $e, k, w \models_{\mathrm{T}} \alpha$ ($e, k, w \models_{\mathrm{F}} \alpha$).

## 5 Properties of $\mathcal{LBI}$

To begin with, we remark that $\mathcal{LBI}$ and $\mathcal{LB}$ coincide precisely when restricted to formulas of the language of $\mathcal{LB}$.

**Proposition 1.** *Let $\alpha$ be a formula of $\mathcal{LB}$. Then $\alpha$ is valid in $\mathcal{LB}$ iff $\alpha$ is valid in $\mathcal{LBI}$.*

**Proof:** The proposition follows easily after observing that for objective $\phi$, $e, k, w \models \boldsymbol{B}_k \phi$ iff for all $w \in S^k(e)$, $w \models_{\mathrm{T}} \phi$, which is exactly how $\mathcal{LB}$ interprets objective beliefs at level $k$. ∎

In general, of course, $\mathcal{LBI}$ goes beyond $\mathcal{LB}$ because of nested beliefs. Before considering general properties of the logic, let us look at some examples.

**Example 3.** *Let $\phi = (p \vee q) \wedge (\neg p \vee q)$ and $e = \{w \mid w \models_{\mathrm{T}} \phi\}$. Then we have the following:*

- $e \models \boldsymbol{B}_0(p \vee q) \wedge \boldsymbol{B}_0(\neg p \vee q)$ *since both disjuncts have true-support in all worlds in e;*
- $e \models \neg \boldsymbol{B}_0 q$ *since for some $w \in e$, $w[p] = *$ and $w[q] = 0$;*
- $e \models \boldsymbol{B}_1 q$ *since $q \in RP(CF(\phi))$ and $RP(CF(\phi)) = S(e)$ by Theorem 2;*
- $e \models \boldsymbol{B}_0 \neg \boldsymbol{B}_0 q$ *since $e, 0, w \models_{\mathrm{T}} \neg \boldsymbol{B}_0 q$ for all $w \in e$;*
- $e \models \boldsymbol{B}_1 \boldsymbol{B}_1 q \wedge \boldsymbol{B}_1 \neg \boldsymbol{B}_0 q$ *since for all $w \in S(e)$, $e, 1, w \models_{\mathrm{T}} \boldsymbol{B}_1 q$ and $e, 1, w \models_{\mathrm{T}} \neg \boldsymbol{B}_0 q$;*
- $e \models \neg \boldsymbol{B}_0 \boldsymbol{B}_1 q$ *since there is an $e' \subseteq e$ (choose $e' = e$) such that $w \not\models_{\mathrm{T}} q$ for some $w \in e'$.*
- $e \models \neg \boldsymbol{B}_0 \neg \boldsymbol{B}_1 q$ *since there is an $e' \subseteq e$ (choose $e' = \{w \mid w \models_{\mathrm{T}} q\}$) such that for all $w \in e'$, $w \models_{\mathrm{T}} q$;*
- $e \models \neg \boldsymbol{B}_0(\boldsymbol{B}_1 q \vee \neg \boldsymbol{B}_1 q)$*, which follows from the two previous items.*

The last example is perhaps the most surprising and generalizes in that valid formulas like $(\boldsymbol{B}_j \phi \vee \neg \boldsymbol{B}_j \phi)$ are not believed at level $k$ in case $k < j$. One way to make sense of this is that the agent is completely oblivious when it comes to beliefs at higher levels than the current one and treats them all as distinct.

Let us now turn to general properties of the logic.

It is easy to show that the agent is now able to fully introspect on its own beliefs at every level:

**Theorem 4** (Positive Introspection)**.**
$\models (\boldsymbol{B}_k \alpha \equiv \boldsymbol{B}_k \boldsymbol{B}_k \alpha)$.

**Proof:** $e \models \boldsymbol{B}_k \alpha$ iff for all $w \in S^k(e)$, $e, k, w \models_{\mathrm{T}} \alpha$ iff for all $w' \in S^k(e)$ for all $w \in S^k(e)$, $e, k, w \models_{\mathrm{T}} \alpha$ iff for all $w' \in S^k(e)$, $e, k, w' \models \boldsymbol{B}_k \alpha$ iff $e \models \boldsymbol{B}_k \boldsymbol{B}_k \alpha$. ∎

**Theorem 5** (Negative Introspection)**.**
$\models (\neg \boldsymbol{B}_k \alpha \supset \boldsymbol{B}_k \neg \boldsymbol{B}_k \alpha)$.

**Proof:** Let $e \models \neg \boldsymbol{B}_k \alpha$. Then there exists a $w \in S^k(e)$ s.t. $e, k, w \not\models_{\mathrm{T}} \alpha$. Then for all $w' \in S^k(e)$ there exists a $w \in S^k(e)$ s.t. $e, k, w \not\models_{\mathrm{T}} \alpha$. Then for all $w' \in S^k(e)$, $e, k, w' \models_{\mathrm{F}} \boldsymbol{B}_k \alpha$. Then for all $w' \in S^k(e)$, $e, k, w' \models_{\mathrm{T}} \neg \boldsymbol{B}_k \alpha$, from which $e \models \boldsymbol{B}_k \neg \boldsymbol{B}_k \alpha$ follows. ∎

The converse also holds except when the successor epistemic state becomes empty.

**Theorem 6.** $\models (\boldsymbol{B}_k \neg \boldsymbol{B}_k \alpha \supset (\neg \boldsymbol{B}_k \alpha \vee \boldsymbol{B}_k \bot))$

**Proof:** Let $e \models \boldsymbol{B}_k \neg \boldsymbol{B}_k \alpha$. Then for all $w \in S^k(e)$, $e, k, w \models_{\mathrm{F}} \boldsymbol{B}_k \alpha$ and thus for all $w \in S^k(e)$ there is a $w' \in S^k(e)$ such that $e, k, w' \not\models_{\mathrm{T}} \alpha$.

If $S^k(e) \neq \{\}$ this means that there indeed is a $w' \in S^k(e)$ such that $e, k, w' \not\models_{\mathrm{T}} \alpha$ and hence $e \models \neg \boldsymbol{B}_k \alpha$.

If $S^k(e) = \{\}$ then $e \models \boldsymbol{B}_k \bot$. (Note that $e \models \boldsymbol{B}_k \bot$ iff $S^k(e) = \{\}$ since no world has true-support for $\bot$.) ∎

The next theorem states that when an agent believes $\alpha$ at some level then it believes at that level that it will continue to believe $\alpha$ at any higher level.

**Theorem 7** (Introspective Cumulativity)**.**
$\models (\boldsymbol{B}_k \alpha \supset \boldsymbol{B}_k \boldsymbol{B}_{k+1} \alpha)$.

**Proof:** Let $e \models \boldsymbol{B}_k \alpha$. Then for all $w \in S^k(e)$, $e, k, w \models \alpha$. Then for all $w \in S^k(e)$ for all $e' \subseteq S^k(e)$ for all $w' \in e'$, $e, k, w' \models_{\mathrm{T}} \alpha$ (because $e' \subseteq S^k(e)$). Therefore, for all $w \in S^k(e)$, $e, k, w \models_{\mathrm{T}} \boldsymbol{B}_{k+1} \alpha$, from which $e \models \boldsymbol{B}_k \boldsymbol{B}_{k+1} \alpha$ follows. ∎

The theorem clearly generalizes when replacing $\boldsymbol{B}_{k+1}$ by any $\boldsymbol{B}_j$ with $j > k$.

**Corollary 2.** $\models (\boldsymbol{B}_k \alpha \supset \boldsymbol{B}_k \boldsymbol{B}_j \alpha)$ *for $j > k$.*

**Theorem 8.** *Let $\alpha$ be an arbitrary basic formula and $j > k$. Then:*

1. $\models (\neg \boldsymbol{B}_k \alpha \supset \neg \boldsymbol{B}_k \boldsymbol{B}_j \alpha)$;
2. $\models (\neg \boldsymbol{B}_k \alpha \supset \neg \boldsymbol{B}_k \neg \boldsymbol{B}_j \alpha)$.

**Proof:**

1. Let $e \models \neg \boldsymbol{B}_k \alpha$. Then for some $w \in S^k(e), e, k, w \not\models_{\mathrm{T}} \alpha$. Then there are $w \in S^k(e)$, $e' \subseteq S^k(e)$ (choose $e' = S^k(e)$) and $w' \in e'$ s.t. $e, k, w' \not\models_{\mathrm{T}} \alpha$. Then for some $w \in S^k(e)$, $e, k, w \not\models_{\mathrm{T}} \boldsymbol{B}_j \alpha$, from which $e \models \neg \boldsymbol{B}_k \boldsymbol{B}_j \alpha$ follows.
2. Again, let $e \models \neg \boldsymbol{B}_k \alpha$. Then for some $w \in S^k(e)$, $e, k, w \not\models_{\mathrm{T}} \alpha$. Let $e' \subseteq S^k(e)$ such that for all $w' \in e'$, $e, k, w' \models_{\mathrm{T}} \alpha$. Such $e'$ must exist. In the worst case, $e' = \{\}$ (which happens, for example, when $e \models \boldsymbol{B}_k \neg \alpha$, $\alpha$ is objective and $S^k(e)$ consists of standard worlds only, i.e. $S^k(e)$ contains no worlds which satisfy $\alpha$). Given the existence of an $e' \subseteq S^k(e)$ such that for all $w' \in e'$, $e, k, w' \models_{\mathrm{T}} \alpha$, we have that for some $w \in S^k(e)$, $e, k, w \not\models_{\mathrm{F}} \boldsymbol{B}_j \alpha$ or, equivalently, $e, k, w \not\models_{\mathrm{T}} \neg \boldsymbol{B}_j \alpha$, from which $e \models \neg \boldsymbol{B}_k \neg \boldsymbol{B}_j \alpha$ follows. ∎

It turns out that we can strengthen Theorem 8.2 even more:

**Theorem 9.** *Let $\alpha$ be an arbitrary basic formula. Then $\models \neg \boldsymbol{B}_k \neg \boldsymbol{B}_j \alpha$ for $j > k$.*

**Proof:** Let $e$ be an epistemic state. $e \models \neg \boldsymbol{B}_k \neg \boldsymbol{B}_j \alpha$ iff $e, k \not\models_{\mathrm{T}} \boldsymbol{B}_k \neg \boldsymbol{B}_j \alpha$ iff for some $w \in S^k(e)$, $e, k, w \not\models_{\mathrm{T}} \neg \boldsymbol{B}_j \alpha$ iff $e, k, w \not\models_{\mathrm{F}} \boldsymbol{B}_j \alpha$ iff for some $e' \subseteq S^k(e)$ and for all $w' \in e'$, $e, k, w' \models_{\mathrm{T}} \alpha$, which holds vacuously if we choose $e'$ to be the empty set. ∎

The theorem asserts that an agent never rules out that it may believe $\alpha$ at a belief level higher than the current one. This holds for arbitrary basic formulas including those which are obviously inconsistent such as $p \wedge \neg p$. In other words, the agent does not rule out that it may ultimately come to believe that its beliefs are inconsistent.

One may wonder whether this is always reasonable. For example, suppose the agent is given a knowledge base with the assurance that the KB is consistent. In other words, it is guaranteed that there is no $k$ such that $(\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_k(p \wedge \neg p))$ is valid. In the next section we will explore such a scenario by slightly modifying the semantics so that the beliefs of an agent are always consistent.

## 6 Assuming Consistency

To guarantee consistency of beliefs it is not sufficient to only rule out the empty epistemic state. After all, there are extended worlds which support the truth of $p \wedge \neg p$ by assigning $p$ the value $*$. As we will see, it does suffice though to require that an epistemic state contains at least one standard world.

**Definition 9.** *An epistemic state $e$ is called consistent if it contains at least one standard world.*

Since we will only consider consistent epistemic states in this section, we need to slightly modify the definitions of the true- and false-support of belief formulas by requiring that the subsets $e'$ in the otherwise-cases are consistent:

4. $e, k, w \models_{\mathrm{T}} \boldsymbol{B}_j \alpha$ iff
   if $j \leq k$ then for all $w' \in S^j(e)$, $e, k, w' \models_{\mathrm{T}} \alpha$,
   o.w. for all consistent $e' \subseteq S^k(e)$
   for all $w' \in e'$, $e, k, w' \models_{\mathrm{T}} \alpha$;
5. $e, k, w \models_{\mathrm{F}} \boldsymbol{B}_j \alpha$ iff
   if $j \leq k$ then for some $w' \in S^j(e)$, $e, k, w' \not\models_{\mathrm{T}} \alpha$,
   o.w. for all consistent $e' \subseteq S^k(e)$
   for some $w' \in e'$, $e, k, w' \not\models_{\mathrm{T}} \alpha$;

Note, in particular, that this rules out $e'$ being the empty set. We leave all other definitions unchanged.

Let us re-define validity as follows: A formula $\alpha$ is $c$-valid ($\models_c \alpha$) if for all standard worlds $w$ and consistent epistemic states $e$, $e, w \models \alpha$.

We begin by noting that positive and negative introspection as well as introspective cumulativity continue to hold for consistent epistemic states. Indeed, for negative introspection we can strengthen the result as we do not have to worry about the empty epistemic state.

**Theorem 10** (Positive Introspection)**.**
$\models_c (\boldsymbol{B}_k \alpha \equiv \boldsymbol{B}_k \boldsymbol{B}_k \alpha)$.

The proof is identical to the proof of Theorem 4.

**Theorem 11** (Negative Introspection)**.**
$\models_c (\neg \boldsymbol{B}_k \alpha \equiv \boldsymbol{B}_k \neg \boldsymbol{B}_k \alpha)$.

**Proof:** The proof of the only-if direction is identical to that of Theorem 5. Conversely, let $e \models \boldsymbol{B}_k \neg \boldsymbol{B}_k \alpha$, that is, for all $w' \in S^k(e)$, $e, k, w' \models_{\mathrm{T}} \neg \boldsymbol{B}_k \alpha$. Since $S^k(e)$ contains at least one standard world $w^*$, we have $e, k, w^* \models_{\mathrm{T}} \neg \boldsymbol{B}_k \alpha$ iff $e, k, w^* \models_{\mathrm{F}} \boldsymbol{B}_k \alpha$ iff $e, k, w^* \not\models_{\mathrm{T}} \boldsymbol{B}_k \alpha$, from which $e \models \neg \boldsymbol{B}_k \alpha$ follows. ∎

**Theorem 12** (Introspective Cumulativity)**.**
$\models_c (\boldsymbol{B}_k \alpha \supset \boldsymbol{B}_k \boldsymbol{B}_{k+1} \alpha)$.

The proof is identical to the proof of Theorem 7.

In contrast to the previous section, believing a sentence now rules out believing its negation. This is easiest to see when the sentence is objective.

**Theorem 13.** $\models_c (\boldsymbol{B}_k \phi \supset \neg \boldsymbol{B}_k \neg \phi)$ *for objective $\phi$.*

**Proof:** Let $e \models \boldsymbol{B}_k \phi$. Then for all $w \in S^k(e)$, $e, k, w \models \phi$. Since $e$ is consistent, there is a standard word $w^* \in e$. Then $w^* \in S^k(e)$. Since $w^* \models \phi$ we have $w^* \not\models \neg \phi$. Hence $e \models \neg \boldsymbol{B}_k \neg \phi$. ∎

In order to generalize the theorem to arbitrary $\alpha$ we need the following lemma.

**Lemma 1.** *For any standard world $w$, consistent epistemic state $e$ and basic formula $\alpha$,*

1. *if $e, k, w \models_{\mathrm{T}} \alpha$ then $e, k, w \not\models_{\mathrm{F}} \alpha$;*
2. *if $e, k, w \models_{\mathrm{F}} \alpha$ then $e, k, w \not\models_{\mathrm{T}} \alpha$.*

**Proof:** The proof is by induction on the number of operators in $\alpha$, henceforth denoted as $|\alpha|$. If $\alpha$ is an atomic proposition $p$ then the lemma holds because $w$ is a standard world, that is, $e, k, w \models_{\mathrm{T}} p$ iff $w[p] = 1$ and $e, k, w \models_{\mathrm{F}} \alpha$ iff $w[p] = 0$. The cases for $\neg \alpha$ and $(\alpha \vee \beta)$ follow easily by induction given the definitions of $\models_{\mathrm{T}}$ and $\models_{\mathrm{F}}$. Let us now consider the case $\boldsymbol{B}_j \alpha$:

1. Let $e, k, w \models_{\mathrm{T}} \boldsymbol{B}_j \alpha$. If $j \leq k$ then for all $w' \in S^k(e)$, $e, k, w' \models_{\mathrm{T}} \alpha$, from which $e, k, w \not\models_{\mathrm{F}} \boldsymbol{B}_j \alpha$ follows immediately. If $j > k$ then for all consistent $e' \subseteq S^k(e)$ and

for all $w' \in e'$, $e, k, w' \models_{\mathrm{T}} \alpha$. Then there is some consistent $e' \subseteq S^k(e)$ (choose $e' = e$, for example) such that for all $w' \in e'$, $e, k, w' \models_{\mathrm{T}} \alpha$, from which $e, k, w \not\models_{\mathrm{F}} \boldsymbol{B}_j\alpha$ follows.

2. Let $e, k, w \models_{\mathrm{F}} \boldsymbol{B}_j\alpha$. If $j \leq k$ then for some $w' \in S^k(e)$, $e, k, w' \not\models_{\mathrm{T}} \alpha$, from which $e, k, w \not\models_{\mathrm{T}} \boldsymbol{B}_j\alpha$ follows immediately. If $j > k$ then for all consistent $e' \subseteq S^k(e)$ there is a $w' \in e'$ such that $e, k, w' \not\models_{\mathrm{T}} \alpha$. Then there is some consistent $e' \subseteq S^k(e)$ (again, choose $e' = e$) and a $w' \in e'$ such that $e, k, w' \not\models_{\mathrm{T}} \alpha$. Thus $e, k, w \not\models_{\mathrm{T}} \boldsymbol{B}_j\alpha$. ∎

**Theorem 14.** $\models_c (\boldsymbol{B}_k\alpha \supset \neg\boldsymbol{B}_j\neg\alpha)$ *for all* $j \geq k$.

**Proof:** Let $e \models \boldsymbol{B}_k\alpha$ and suppose $e \models \boldsymbol{B}_j\neg\alpha$. Then for all $w' \in S^k(e)$, $e, k, w' \models_{\mathrm{T}} \alpha$ and for all $w' \in S^j(e)$, $e, k, w' \models_{\mathrm{F}} \alpha$. However, since $e$ is consistent, $S^j(e)$ contains at least one standard world $w^*$, which is also in $S^k(e)$. By Lemma 1, since $e, k, w^* \models_{\mathrm{T}} \alpha$ by assumption, $e, k, w^* \not\models_{\mathrm{F}} \alpha$, contradiction. ∎

The following result shows that, under the assumption of consistency, if the agent believes a sentence at some level, then it believes that it will never believe its negation.

**Theorem 15.** $\models_c (\boldsymbol{B}_k\alpha \supset \boldsymbol{B}_k\neg\boldsymbol{B}_j\neg\alpha)$ *for all* $j \geq k$.

**Proof:** If $j = k$ the theorem follows from Theorem 14 and negative introspection (Theorem 11).

Now let $e \models \boldsymbol{B}_k\alpha$ and $j > k$. Then for all $w \in S^k(e)$, $e, k, w \models_{\mathrm{T}} \alpha$. Suppose $e \models \neg\boldsymbol{B}_k\neg\boldsymbol{B}_j\neg\alpha$. Then for some $w \in S^k(e)$, $e, k, w \not\models_{\mathrm{F}} \boldsymbol{B}_j\neg\alpha$. Thus for some consistent $e' \subseteq S^k(e)$ and for all $w' \in e', e, k, w' \models_{\mathrm{T}} \neg\alpha$. But $e'$ contains at least one standard world $w^*$ from $S^k(e)$ because $e'$ is consistent. By assumption $e, k, w^* \models_{\mathrm{T}} \alpha$ and therefore, by Lemma 1, $e, k, w^* \not\models_{\mathrm{T}} \neg\alpha$, contradiction. ∎

Note that this theorem also shows that Theorem 9 fails if we assume consistency.

While the assumption of consistency arguably has a number of nice properties, it has one serious downside. As the following theorem shows, consistency forces the agent to be able to do full logical reasoning already at level 0!

**Theorem 16.** *Let* $\phi$ *and* $\psi$ *be objective. Then* $\models (\phi \supset \psi)$ *iff* $\models_c (\boldsymbol{O}\phi \supset \boldsymbol{B}_0\neg\boldsymbol{B}_1\neg\psi)$.

**Proof:** For the only-if direction, assume $\models (\phi \supset \psi)$. If $\models \neg\phi$, then $\models_c \neg\boldsymbol{O}\phi$ (because $\Re[\![\phi]\!]$ contains no standard worlds) and the RHS of the theorem holds vacuously. Otherwise, let $e \models \boldsymbol{O}\phi$, that is, $e = \Re[\![\phi]\!]$ is consistent and hence contains a standard world $w^*$. We need to show that for any $w \in e$, $e, 0, w \models_{\mathrm{F}} \boldsymbol{B}_1\neg\psi$, that is, for all consistent $e' \subseteq e$ there is a world $w' \in e'$ such that $w' \not\models_{\mathrm{T}} \neg\psi$. Consider any such $e'$. Since $e'$ is consistent, it contains a standard world $w'$ and $w' \models_{\mathrm{T}} \phi$. By assumption, $w' \models_{\mathrm{T}} \psi$ and hence $w' \not\models_{\mathrm{T}} \neg\psi$.

Conversely, assume $\models_c (\boldsymbol{O}\phi \supset \boldsymbol{B}_0\neg\boldsymbol{B}_1\neg\psi)$. If $\models \neg\phi$ we obviously have $\models (\phi \supset \psi)$. Otherwise, let $w^*$ be a standard world such that $w^* \models \phi$. We need to show that $w^* \models \psi$. $\Re[\![\phi]\!] \models \boldsymbol{O}\phi$. By assumption $\Re[\![\phi]\!] \models \boldsymbol{B}_0\neg\boldsymbol{B}_1\neg\psi$. Thus for all consistent $e' \subseteq e$ there is a $w' \in e'$ such that $w' \not\models_{\mathrm{T}} \neg\psi$. Since $w^* \in \Re[\![\phi]\!]$ and $\{w^*\} \subseteq e$, we have $w^* \not\models_{\mathrm{T}} \neg\psi$ and hence $w^* \models \psi$. ∎

Given the fact that deciding whether $\models (\phi \supset \psi)$ is co-NP-complete, the theorem tells us essentially that requiring epistemic states to be consistent may be asking for too much if we are interested in keeping reasoning tractable. So let us go back to the logic $\mathcal{LBI}$ as originally defined in Section 3. As we will see, reasoning can be shown to be reducible to reasoning in $\mathcal{LB}$ and hence remains tractable.

## 7 Computing the Beliefs of a KB

In this section we are interested in computing whether $(\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_k\alpha)$ is valid for any objective KB and arbitrary basic $\alpha$.

For any basic $\alpha$ let us call a subformula $\boldsymbol{B}_j\beta$ depth-0 if it does not occur within the scope of another belief operator. The following inductive definition replaces the leftmost occurrence of a depth-0 subformula $\boldsymbol{B}_j\beta$ in $\alpha$ by either $\top$, $\bot$ or a new atomic proposition $q$.

**Definition 10.** *Let* KB *be objective and* $\alpha$ *basic. Let* $q$ *be a normal atomic proposition that does not occur in* KB *and* $\alpha$. *For any* $k \geq 0$ *we define* $\alpha{\downarrow}^{\mathrm{KB},k}$ *inductively as follows:*

- $\alpha{\downarrow}^{\mathrm{KB},k} = \alpha$ *if* $\alpha$ *is objective;*
- $(\neg\alpha){\downarrow}^{\mathrm{KB},k} = \neg(\alpha{\downarrow}^{\mathrm{KB},k})$;
- $(\alpha \vee \beta){\downarrow}^{\mathrm{KB},k} = \begin{cases} (\alpha{\downarrow}^{\mathrm{KB},k} \vee \beta) & \text{if } \alpha \text{ is not objective} \\ (\alpha \vee \beta{\downarrow}^{\mathrm{KB},k}) & o.w. \end{cases}$
- $(\boldsymbol{B}_j\alpha){\downarrow}^{\mathrm{KB},k} = \begin{cases} \top & \text{if } j \leq k \text{ and } \Re[\![\mathrm{KB}]\!] \models \boldsymbol{B}_j\alpha \\ \bot & \text{if } j \leq k \text{ and } \Re[\![\mathrm{KB}]\!] \models \neg\boldsymbol{B}_j\alpha \\ \top & \text{if } j > k \text{ and } \Re[\![\mathrm{KB}]\!] \models \boldsymbol{B}_k\alpha \\ q & o.w. \end{cases}$

**Example 4.** *Let* KB $= p \wedge (p \supset q)$.
*Then* $\boldsymbol{B}_0p{\downarrow}^{\mathrm{KB},0} = \top$, $\boldsymbol{B}_0q{\downarrow}^{\mathrm{KB},0} = \bot$,
*and* $((\boldsymbol{B}_1q \vee \neg\boldsymbol{B}_1q){\downarrow}^{\mathrm{KB},0}){\downarrow}^{\mathrm{KB},0} = (r \vee \neg s)$.

*Note that since* $\alpha{\downarrow}^{\mathrm{KB},k}$ *replaces depth-0 belief formulas one at a time, identical occurrences may be replaced by different atoms, as in the last example.*

With the above definition in place, we can show that evaluating a belief $\alpha$ at level $k$ for a given KB reduces to recursively replacing any beliefs occurring in $\alpha$ by an appropriate atom, resulting in an objective formula $\alpha'$. As a final step, $\alpha$ is believed at level $k$ just in case $\alpha'$ is. Before diving into the technical details, let us consider some examples.

**Example 5.** *Let* KB $= p \wedge (p \supset q)$. *We are interested in determining whether* $\models (\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_k\alpha)$ *or, equivalently, whether* $\Re[\![\mathrm{KB}]\!] \models \boldsymbol{B}_k\alpha$.

1. *In order to show that* $\models (\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_0(p \wedge \boldsymbol{B}_1p \wedge \neg\boldsymbol{B}_0q))$ *holds, we first replace* $\boldsymbol{B}_1p$ *by* $\top$ *because* $\models (\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_1p)$ *and* $\boldsymbol{B}_0q$ *by* $\bot$ *because* $\not\models (\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_0q)$. *Then we confirm that* $\models (\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_0(p \wedge \top \wedge \neg\bot))$ *holds and we are done.*
2. $\not\models (\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_0(\boldsymbol{B}_1q \vee \neg\boldsymbol{B}_1q))$. *The entailment fails because at level* 0 *neither* $\boldsymbol{B}_1q$ *nor* $\neg\boldsymbol{B}_1q$ *is believed. This can be determined by replacing* $\boldsymbol{B}_1q$ *and* $\boldsymbol{B}_1q$ *by new atoms* $r$ *and* $s$ *and confirming that* $\not\models (\boldsymbol{O}\mathrm{KB} \supset \boldsymbol{B}_0(r \vee \neg s))$.

The next lemma establishes that, in case the leftmost occurrence of a belief in $\alpha$ is replaced by $\top$ or $\bot$, the true- and false-support for $\alpha$ and $\alpha{\downarrow}^{\text{KB},k}$ remains the same.

**Lemma 2.** *Let $e = \Re[\![\text{KB}]\!]$ for* KB *objective, $\alpha$ basic, $w$ any world, and $k \geq 0$. If $\alpha{\downarrow}^{\text{KB},k}$ does not mention a new normal atom not occurring in* KB *and $\alpha$, then*

$e, k, w \models_{\text{T}} \alpha$ *iff* $e, k, w \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$;
$e, k, w \models_{\text{F}} \alpha$ *iff* $e, k, w \models_{\text{F}} \alpha{\downarrow}^{\text{KB},k}$.

The proof is by a simple induction on $\alpha$ using Corollary 2 for the case $\boldsymbol{B}_j\alpha$ where $j > k$.

We now turn to the case when $\alpha{\downarrow}^{\text{KB},k}$ replaces the leftmost occurrence of a belief formula in $\alpha$ by a new atom. The following lemma establishes that $\alpha$ is believed at level $k$ just in case $\alpha{\downarrow}^{\text{KB},k}$ is believed independent of the truth value of the new atom.

For any world $w$ let $w_{q:0}$ denote the world which is exactly like $w$ except $w_{q:0}[q] = 0$. Similarly, $w_{q:1}$ ($w_{q:*}$) denotes the world which is exactly like $w$ except $w_{q:0}[q] = 1$ ($w_{q:0}[q] = *$).

**Lemma 3.** *Let $e = \Re[\![\text{KB}]\!]$ for* KB *objective, $\alpha$ basic, $w$ any world, and $k \geq 0$. If $\alpha{\downarrow}^{\text{KB},k}$ mentions a new normal atom $q$, that is, the leftmost occurrence of a belief formula is replaced by $q$, then*

$e, k, w \models_{\text{T}} \alpha$ *iff* $e, k, w_{q:0} \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$ *and* $e, k, w_{q:1} \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$;
$e, k, w \models_{\text{F}} \alpha$ *iff* $e, k, w_{q:0} \models_{\text{F}} \alpha{\downarrow}^{\text{KB},k}$ *and* $e, k, w_{q:1} \models_{\text{F}} \alpha{\downarrow}^{\text{KB},k}$.

**Proof:** The proof is by induction on $|\alpha|$.

Since we assume that $\alpha{\downarrow}^{\text{KB},k}$ contains a new atom, $\alpha$ cannot be objective. Hence the base case is $\boldsymbol{B}_j\alpha$ with $j > k$, $(\boldsymbol{B}_j\alpha){\downarrow}^{\text{KB},k} = q$ and $e \not\models \boldsymbol{B}_k\alpha$. Therefore $e, k, w \not\models_{\text{T}} \boldsymbol{B}_j\alpha$ (choose $e' = e$ in Definition 1.4) and $e, k, w \not\models_{\text{F}} \boldsymbol{B}_j\alpha$ (choose $e' = \{\}$ in Definition 1.5) We also have $e, k, w_{q:0} \not\models_{\text{T}} q$ and $e, k, w_{q:1} \not\models_{\text{F}} q$. Hence both the LHS and RHS of the statements of the lemma are false. In other words, the lemma holds trivially in this case.

Suppose the lemma holds for $|\alpha| < n$ and let $|\alpha| = n$. Here we only consider $(\alpha \vee \beta){\downarrow}^{\text{KB},k} = (\alpha{\downarrow}^{\text{KB},k} \vee \beta)$ and the case for $\models_{\text{F}}$.

$e, k, w \models_{\text{F}} (\alpha \vee \beta)$ iff $e, k, w \models_{\text{F}} \alpha$ and $e, k, w \models_{\text{F}} \beta$ iff (by induction) and $e, k, w_{q:0} \models_{\text{F}} \alpha{\downarrow}^{\text{KB},k}$ and $e, k, w_{q:1} \models_{\text{F}} \alpha{\downarrow}^{\text{KB},k}$ and $e, k, w \models_{\text{F}} \beta$ iff (since $q$ does not occur in $\beta$) ($e, k, w_{q:0} \models_{\text{F}} \alpha{\downarrow}^{\text{KB},k}$ and $e, k, w_{q:0} \models_{\text{F}} \beta$) and ($e, k, w_{q:1} \models_{\text{F}} \alpha{\downarrow}^{\text{KB},k}$ and $e, k, w_{q:1} \models_{\text{F}} \beta$) iff $e, k, w_{q:0} \models_{\text{F}} (\alpha \vee \beta){\downarrow}^{\text{KB},k}$ and $e, k, w_{q:1} \models_{\text{F}} (\alpha \vee \beta){\downarrow}^{\text{KB},k}$ ∎

In case $\alpha{\downarrow}^{\text{KB},k}$ does not introduce a new normal atom, Lemma 2 is already all we need in order to prove the main theorems (Theorem 17 and 19 below). The case where $\alpha{\downarrow}^{\text{KB},k}$ introduces a new atom is more tricky. We need to show that for all $w \in S^k(e)$, $e, k, w \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$ iff for all $w \in S^k(e)$, $e, k, w_{q:0} \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$ and $e, k, w_{q:1} \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$. For that we need to take a closer look at $\Re[\![\text{KB}]\!]$ where KB does not mention $q$. Clearly, for any world $w$, if $w \in \Re[\![\text{KB}]\!]$, then so are $w_{q:0}, w_{q:1}$, and $w_{q:*}$. What needs to be shown is that for any $e$ with such a property, if $e$ eliminates one of the three worlds, then $e$ eliminates all three. The following lemmas establish what we need.

In the following, for any world $w$ and normal atom $q$, let $A_{w,q} = \{w_{q:0}, w_{q:1}, w_{q:*}\}$.

**Lemma 4.** *Let $\phi$ be objective, $w$ a world, and $q$ a normal atom not occurring in $\phi$. Then*

*if* $w \models_{\text{T}} \phi$ *then* $w' \models_{\text{T}} \phi$ *for all* $w' \in A_{w,q}$;
*if* $w \models_{\text{F}} \phi$ *then* $w' \models_{\text{F}} \phi$ *for all* $w' \in A_{w,q}$.

The proof is by a simple induction on $|\phi|$.

**Lemma 5.** *Let $e$ be an epistemic state, $q$ a normal atom such that for all worlds $w$, if $w \in e$ then for all $w' \in A_{w,q}$, $w' \in e$. Then if $e$ eliminates $w$ then $e$ eliminates every $w \in A_{w,q}$.*

The proof is by contradiction using Definition 4 and considering all cases of $w$ assigning a value to $q$.

**Lemma 6.** *Let $e = \Re[\![\text{KB}]\!]$ for* KB *objective, $\alpha$ basic, $w$ a world, and $k \geq 0$. Then*

*if* $e, k, w \models_{\text{T}} \alpha$ *then* $e, k, w_{q:*} \models_{\text{T}} \alpha$;
*if* $e, k, w \models_{\text{F}} \alpha$ *then* $e, k, w_{q:*} \models_{\text{F}} \alpha$.

The proof is by a simple induction on $|\alpha|$.

**Lemma 7.** *Let $e = \Re[\![\text{KB}]\!]$. If $q$ is a normal atom not in* KB, *then for all worlds $w$, if $w \in S^k(e)$ then for all $w' \in A_{w,q}$, $w' \in S^k(e)$.*

The proof is by simple induction using Lemma 4 and 5.

**Lemma 8.** *Let* KB, *$e = \Re[\![\text{KB}]\!]$, and $\alpha$ basic. Let $q$ be a normal atom not in* KB. *Then*

*for all* $w \in S^k(e)$, $e, k, w \models_{\text{T}} \alpha$ *iff*
*for all* $w \in S^k(e)$, $e, k, w_{q:0} \models_{\text{T}} \alpha$ *and* $e, k, w_{q:1} \models_{\text{T}} \alpha$.

**Proof:** For the only-if direction, suppose for all $w \in S^k(e)$, $e, k, w \models_{\text{T}} \alpha$. The lemma follows because for any $w \in S^k(e)$, both $w_{q:0}$ and $w_{q:1}$ are also in $S^k(e)$ by Lemma 7.

Conversely, suppose for all $w \in S^k\alpha$, $e, k, w_{q:0} \models_{\text{T}} \alpha$ and $e, k, w_{q:1} \models_{\text{T}} \alpha$. We need to show that $e, k, w \models_{\text{T}} \alpha$ for any $w \in S^k(e)$. Let $w \in S^k(e)$. Then by Lemma 7, for all $w' \in A_{w,q}$, $w' \in S^k(e)$. If $w = w_{q:0}$ or $w = w_{q:1}$, then $e, k, w \models_{\text{T}} \alpha$ follows by assumption. If $w = w_{q:*}$ then $e, k, w \models_{\text{T}} \alpha$ follows from Lemma 6. ∎

We now have all the pieces to prove the main theorem:

**Theorem 17.** *For any objective* KB *and basic $\alpha$,* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\alpha)$ *iff* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k(\alpha{\downarrow}^{\text{KB},k}))$.

**Proof:** Since $e \models \boldsymbol{O}\text{KB}$ iff $e = \Re[\![\text{KB}]\!]$, it suffices to show that $\Re[\![\text{KB}]\!] \models \boldsymbol{B}_k\alpha$ iff $\Re[\![\text{KB}]\!] \models \boldsymbol{B}_k(\alpha{\downarrow}^{\text{KB},k})$. Let $e = \Re[\![\text{KB}]\!]$.

Case 1: $\alpha{\downarrow}^{\text{KB},k}$ does not mention any normal atom not in $\alpha$. Then $e \models \boldsymbol{B}_k\alpha$ iff for all $w \in S^k(e)$, $e, k, w \models_{\text{T}} \alpha$ iff (by Lemma 2) for all $w \in S^k(e)$, $e, k, w \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$ iff $e \models \boldsymbol{B}_k(\alpha{\downarrow}^{\text{KB},k})$.

Case 2: $\alpha{\downarrow}^{\text{KB},k}$ mentions a normal atom not in $\alpha$.
Then $e \models \boldsymbol{B}_k\alpha$ iff for all $w \in S^k(e)$, $e, k, w \models_{\text{T}} \alpha$ iff (by Lemma 3) for all $w \in S^k(e)$, $e, k, w_{q:0} \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$ and $e, k, w_{q:1} \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$ iff (by Lemma 8) for all $w \in S^k(e)$, $e, k, w \models_{\text{T}} \alpha{\downarrow}^{\text{KB},k}$ iff $e \models \boldsymbol{B}_k(\alpha{\downarrow}^{\text{KB},k})$. ∎

With Theorem 17 we are now able to prove by induction that deciding whether $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\alpha)$ holds can be done in polynomial time. The base case involves objective $\phi$ and essentially reduces to the earlier tractability result (Corollary 1 of Section 3). The only slight complication is when $\alpha$ mentions $\top$ or $\bot$, which can be introduced by $\alpha{\downarrow}^{\text{KB},k}$.

**Theorem 18.** *Let* KB *be in CNF not mentioning* $\top$ *or* $\bot$*, and let* $\phi$ *be objective and of constant size. Then for any* $k$*,* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$ *is computable in time polynomial in the size of* KB.

**Proof:** We begin by transforming $\phi$ into CNF and let $\phi_{\text{CNF}} = \bigwedge c_i$ be the resulting formula. While CNF-conversion can lead to an exponential blowup, this is still poly-time as $\phi$ is assumed to be of constant size. We also have $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$ iff $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi_{\text{CNF}})$. Since $\models (\boldsymbol{B}_k(\phi \wedge \psi) \equiv \boldsymbol{B}_k\phi \wedge \boldsymbol{B}_k\psi)$ we have $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi_{\text{CNF}})$ iff $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k c_i)$ for all clauses $c_i$. There are three cases:

1. If $c_i$ does not mention $\top$ and $\bot$ then the theorem holds because $\mathcal{LB}$ and $\mathcal{LBI}$ agree in this case (Proposition 1) and the theorem follows by Corollary 1.
2. If $c_i$ mentions $\top$ then $\models \boldsymbol{B}_0 c_i$ and thus $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k c_i)$ holds trivially.
3. If $c_i$ mentions $\bot$ then we either end up with case (1) after removing $\bot$ from the clause or $c_i = \bot$. In this case $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\bot)$ iff $S^k(\Re[\![\text{KB}]\!]) = \{\}$ or equivalently, by Theorem 2, $RP^k(CF(\text{KB}))$ contains the empty clause, which can be computed in polynomial time because $k$ is constant. ∎

**Theorem 19.** *Let* KB *be in conjunctive normal form not containing* $\top$ *and* $\bot$*. Let* $\alpha$ *be an arbitrary basic formula. Deciding whether* $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\alpha)$ *holds is tractable, that is, can be done in time polynomial in the size of* KB.

**Proof:** The proof is by induction on the number of belief formulas occurring in $\alpha$. If $\alpha$ contains no belief formulas, then $\alpha$ is objective and the claim holds by Theorem 18. Suppose the theorem holds for $\alpha$ containing $< n$ belief subformulas and let $\alpha$ now contain $n$ such subformulas. By Theorem 17, $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\alpha)$ iff $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k(\alpha{\downarrow}^{\text{KB},k}))$. Since $\alpha{\downarrow}^{\text{KB},k}$ contains at most $n-1$ belief subformulas, by induction, computing whether $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k(\alpha{\downarrow}^{\text{KB},k}))$ is tractable. Computing $\alpha{\downarrow}^{\text{KB},k}$ itself requires finding the leftmost occurrence of a belief formula $\boldsymbol{B}_j\beta$ within $\alpha$, which is trivial, and then replacing it by either $\top$, $\bot$, or a new atom $q$. This requires checking whether $\Re[\![\text{KB}]\!] \models \boldsymbol{B}_j\beta$ or $\Re[\![\text{KB}]\!] \models \boldsymbol{B}_k\beta$ in case $j > k$. Note that $\Re[\![\text{KB}]\!] \models \boldsymbol{B}_j\beta$ iff $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_j\beta)$ and $\Re[\![\text{KB}]\!] \models \boldsymbol{B}_k\beta$ iff $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\beta)$. Since $\beta$ contains less than $n$ belief formulas, the induction hypothesis applies again, that is, computing these entailments is again tractable. ∎

To recap, let us walk through the argument again why the earlier tractability result for nonnested beliefs generalizes to nested beliefs. As a preliminary step, Theorem 18, which forms the base case for Theorem 19, establishes that the nonnested case remains tractable even if the query mentions $\top$ or $\bot$, which may be introduced when applying $\alpha{\downarrow}^{\text{KB},k}$. In the induction step of Theorem19 we make use of the fact that determining whether $\alpha$ is believed at level $k$ is the same as determining whether $\alpha{\downarrow}^{\text{KB},k}$ is believed (Theorem 17). The induction applies because $\alpha{\downarrow}^{\text{KB},k}$ mentions fewer belief formulas than $\alpha$. What remains to be shown is that computing $\alpha{\downarrow}^{\text{KB},k}$ itself is tractable. As detailed in Definition 10, $\alpha{\downarrow}^{\text{KB},k}$ replaces the leftmost belief formula $\boldsymbol{B}_j\beta$ in $\alpha$ by either $\top$, $\bot$ or a new atom. This involves computing whether $\beta$ is believed at level $j$ or $k$, depending on whether $j \leq k$ or not. Since $\beta$ is no bigger than $\alpha$, the induction hypothesis applies again, and we are done. We remark that the induction proof indeed forms a valid argument because the decision problem depends only on the size of the KB, keeping the size of $\alpha$ and $k$ constant. Note that this means, in particular, that the number of belief formulas in $\alpha$ is also constant and $\alpha{\downarrow}^{\text{KB},k}$ is no bigger than $\alpha$.

With Theorem 19, a practical algorithm to decide whether $(\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\alpha)$ is valid can easily be derived by first replacing the innermost occurrences of subformulas $\boldsymbol{B}_j\phi$ in $\alpha$, where $\phi$ is objective, by $(\boldsymbol{B}_j\phi){\downarrow}^{\text{KB},k}$, which involves computing whether $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_j\phi)$ if $j \leq k$ or $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\phi)$ otherwise. The procedure is then iterated until $\alpha$ is reduced to an objective formula $\psi$. As a final step one needs to check whether $\models (\boldsymbol{O}\text{KB} \supset \boldsymbol{B}_k\psi)$ holds.

## 8 Related Work

Modeling belief in terms of possible worlds goes back to (Kripke 1959; Hintikka 1962). While the classical approach invariably leads to logical omniscience and intractable reasoning, Hintikka already observed the usefulness of nonstandard worlds to address this problem (Hintikka 1975). Research in KR developed this idea further by considering four-valued worlds as a basis of weaker models of belief or inference, see, for example, (Levesque 1984; Cadoli and Schaerf 1996; Patel-Schneider 1985; Frisch 1987; Delgrande 1995; Cadoli and Schaerf 1996) with close connections to earlier work on *tautological entailment* (Anderson and Belnap 1975; Dunn 1976). Tractability results were obtained for knowledge bases in CNF. One limitation of this line of work was that it essentially only considered beliefs at level 0.

Limited nested beliefs with tractability results were first considered in (Lakemeyer 1987) and later by (Fagin *et al.* 1995) by extending the four-valued approach to accessibility relations. Lakemeyer (Lakemeyer 1996) considered a four-valued first-order logic with full introspection and obtained decidability results. Again, only beliefs at level 0 were considered.

Models of belief with belief levels together with tractability results for every level were first considered by Liu et al. (2004) and further developed in (Liu and Levesque 2005; Lakemeyer and Levesque 2013; Lakemeyer and Levesque 2014; Lakemeyer and Levesque 2016; Schwering 2017). The semantics used epistemic states defined in terms of sets of clauses called setups. Higher belief levels were obtained by splitting cases, for example, by adding a literal to a setup, checking whether the belief obtains using the mod-

ified setup, and then doing the same by adding the complement of the literal to the setup. Klassen (2015) also used the splitting idea, but based the semantics on a three-valued version of neighborhood semantics (Montague 1968; Scott 1970). In a sense, the logic we proposed in (Lakemeyer and Levesque 2020) and which forms the basis of the current paper, can be thought of as a continuation of the logics using setups. In a preliminary version (Lakemeyer and Levesque 2019), we use setups and define belief levels directly in terms of $RP(C)$ (Definition 6) before giving a possible-world account in (Lakemeyer and Levesque 2020).

In this line of work on belief levels, we already considered nested beliefs once before (Lakemeyer and Levesque 2013). Here beliefs are also fully introspective and we obtained decidability results for first-order knowledge bases of a restricted form. Compared to our new approach, there is one important difference in how beliefs about higher levels of belief are handled. For example, in (Lakemeyer and Levesque 2013) both $(\boldsymbol{B}_8\alpha \supset \boldsymbol{B}_0\boldsymbol{B}_8\alpha)$ and $(\neg\boldsymbol{B}_8\alpha \supset \boldsymbol{B}_0\neg\boldsymbol{B}_8\alpha)$ are valid. In other words, the agent is expected to already know at level 0 whether or not it is able to figure out $\alpha$ at level 8. We now believe that this assumption is much too strong and that our new approach is arguably more intuitive.

Besides modeling belief levels explicitly in an epistemic logic, there has also been work on forms of tractable entailment relations with increasing complexity, starting with (Dalal 1996). In more recent work D'Agostino (2015) considers $k$-consequence relations in propositional logic based on a multi-valued nondeterministic semantics, which goes back to (Crawford and Etherington 1998). Higher levels of the consequence relation are obtained by splitting on formulas. Among other things, this work also offers complete proof systems. This approach was later extended beyond classical propositional logic (D'Agostino et al. 2021; D'Agostino and Solares-Rojas 2024). As there is no explicit notion of belief, nested beliefs do not play a role.

## 9 Conclusion

In this paper we generalized our earlier logic of limited belief to account for nested beliefs at any level. The main technical result is that computing whether only-believing a knowledge base consisting of objective formulas entails believing a sentence $\alpha$ at level $k$ remains tractable even when $\alpha$ mentions arbitrary belief formulas. We also saw that requiring beliefs to be consistent leads to intractability.

In future work it will be interesting to explore nested beliefs in the full first-order language considered by LL. This would allow us to express interesting *de re* vs. *de dicto* distinctions such as $\exists x\boldsymbol{B}_1(P(x) \wedge \boldsymbol{B}_0(\exists y P(y) \wedge \neg\boldsymbol{B}_0 P(y)))$, which can be read as "the agent believes at level 1 of some individual that it has property $P$ yet at level 0 it only believes that something has property $P$ without knowing who that individual is." Besides defining an appropriate semantics it remains to be seen whether tractability results can be obtained as well.

## Acknowledgments

Gerhard Lakemeyer acknowledges that much of the work on this paper was conducted while he was also affiliated with the University of Toronto.

## AI Declaration

The authors have not employed any Generative AI tools.

## References

A.R. Anderson and N.D. Belnap. 1975. *Entailment: The Logic of Relevance and Necessity*. Princeton University Press.

F. Baader, D. Calvanese, D.L. McGuiness, D. Nardi, P.F. Patel-Schneider (eds.) *The Description Logic Handbook.* Second Edition, Cambridge University Press, 2007.

M. Cadoli and M. Schaerf. 1996. On the complexity of entailment in propositional multivalued logics. *Annals of Mathematics and Artificial Intelligence*, 18(1):29–50.

J.M. Crawford and D. Etherington. 1998. A non-deterministic semantics for tractable inference. In *Proc. of AAAI-98*, 286–291.

M. Dalal. 1996. Semantics of an anytime family of reasoners. In *Proc. of the 12th European Conference on Artificial Intelligence (ECAI-96)*, 360–364.

M. D'Agostino. 2015. An informational view of classical logic. Theor. Comput. Sci. 606, 79–97.

M. D'Agostino, C. Larese, and S. Modgil. 2021. Towards Depth-bounded Natural Deduction for Classical First-order Logic. *IfCoLog J. of Logics and their Applications* 8(2): 423–452.

M. D'Agostino and A. Solares-Rojas. 2024. Tractable depth-bounded approximations to FDE and its satellites. *J. Log. Comput.* 34(5): 815–855.

J.P. Delgrande. 1995. A framework for logics of explicit belief. *Computational Intelligence*, 11(1):47–88.

J.M. Dunn. 1976. Intuitive semantics for first-degree entailments and coupled trees. *Philosophical Studies*, 29:149–168.

A. Frisch. 1987. Inference without chaining. In *Proc. of IJCAI-87*, 515–519.

J.Y. Halpern and Y. Moses. 1992. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54(3):319–379.

R. Fagin, J.Y. Halpern, and M.Y. Vardi. 1995. A nonstandard approach to the logical omniscience problem. *Artif. Intell.* 79(2): 203–240, 1995.

J. Hintikka. 1962. *Knowledge and Belief.* Cornell University Press.

J. Hintikka. 1975. Impossible Possible Worlds Vindicated. *Journal of Philosophical Logic,* 4, 475–484.

T.Q. Klassen, S.A. McIlraith, and H.J. Levesque. 2015. Towards Tractable Inference for Resource-Bounded Agents. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, 89–95, AAAI Press.

S. Kripke. 1959. A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24:1–14.

G. Lakemeyer. 1987. Tractable Meta-Reasoning in Propositional Logics of Belief. *Proc. of the 10th International Joint Conference on Artificial Intelligence (IJCAI)*, 401–408.

G. Lakemeyer. 1996. *Limited Reasoning in First-Order Knowledge Bases with Full Introspection Artif. Intell.* 84(1-2): 209–255.

G. Lakemeyer and H.J. Levesque. 2013. Decidable Reasoning in a Logic of Limited Belief with Introspection and Unknown Individuals. *23rd International Joint Conference on Artificial Intelligence (IJCAI)* 969–975.

G. Lakemeyer and H.J. Levesque. 2014. Decidable Reasoning in a Fragment of the Epistemic Situation Calculus. *Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR-14),* Vienna.

G. Lakemeyer and H.J. Levesque. 2016. Decidable Reasoning in a Logic of Limited Belief with Function Symbols. In *Proc of KR-16*, 288–297.

G. Lakemeyer and H.J. Levesque. 2019. A Tractable, Expressive, and Eventually Complete First-Order Logic of Limited Belief. *Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, 1764–1771.

G. Lakemeyer and H.J. Levesque. 2020. A Logic of Limited Belief Based on Possible Worlds. *Proc. of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 624–635.

H.J. Levesque. A logic of implicit and explicit belief. In *Proc. of AAAI-84*, 198–202, 1984.

H.J. Levesque and G. Lakemeyer. The Logic of Knowledge Bases. MIT Press, 2001.

Y. Liu, G. Lakemeyer, and H.J. Levesque. 2004. A Logic of Limited Belief for Reasoning with Disjunctive Information. *Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR),* 587–597.

Y. Liu and H.J. Levesque. 2005. Tractable reasoning in first-order knowledge bases with disjunctive information. *In Proc. of the Twentieth National Conference on Artificial Intelligence (AAAI-05).*

R. Montague. 1968. Pragmatics. In Klibansky, R. (Ed.), *Contemporary Philosophy*, Firenze: La Nuova Italia Editrice, 102—122.

P. Patel-Schneider. 1985. A decidable first-order logic for knowledge representation. In *Proc. of IJCAI-85*, 455–458.

D. Scott. 1970. Advice on Modal Logic. In Lambert, K. (Ed.), *Philosophical Problems in Logic*, Dordrecht, Holland: D.Reidel.

C. Schwering. 2017. A Reasoning System for a First-Order Logic of Limited Belief. In *Proc. of IJCAI-17*, 1247–1253.