Tractable Meta-Reasoning in Propositional Logics of Belief

Gerhard Lakemeyer
Department of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 1A4

Abstract

Finding adequate semantic models of deductive reasoning is a difficult problem, if deductions are to be performed efficiently and in a semantically appropriate way. The model of reasoning provided by possible-worlds semantics has been found deficient both for computational and intuitive reasons. Existing semantic approaches that were proposed as alternatives to possible-worlds semantics either suffer from computational intractability or do not allow agents to have meta-beliefs. This work, based on relevance logic, proposes a model of belief where an agent can hold met a-beliefs and reason about them and other world knowledge efficiently. It is also shown how the model can be extended to include positive introspection without losing efficiency.

I. Introduction

Most knowledge based systems can perform some form of deductive reasoning on their knowledge bases, which contain explicit representations of some aspect of the world. Those systems or agents are usually thought of as knowing or believing[1] some fact about the world, if they are able to deduce the fact (or, more precisely, its representation) from the knowledge base. For important tasks such as planning actions, it is generally not sufficient that agents have beliefs only about some application domain but also about their own knowledge. For example, only after realizing that one does *not know* a friend's phone number one would go about finding out what it is (see Moore [Moor80] and Konolige [Kono84] for further examples on the importance of meta-knowledge and, for that matter, meta-reasoning).

Independent of what beliefs are about, actual reasoning must obey certain computational constraints. For example, resource limitations are a fact of life, and when interacting with the environment, systems can afford to spend only so much time "thinking" before they are forced to act.

We are interested in models of reasoning that give rise to good computational performance, even in the presence of meta-beliefs. In addition, these models should give a semantic account of reasoning, that is, relate it to the notion of truth in a way that conforms as much as possible with our intuitions.

The framework for our investigations are logics of knowledge and belief and their model theories, which have been convenient tools in the study of reasoning for two reasons. For one, they allow us to address meta-beliefs in a direct and elegant way. Secondly, models that predict what beliefs follow from a given set of beliefs can be viewed as a knowledge level specification of a reasoner that has somehow represented those beliefs internally.

One of the best understood models of knowledge and belief is provided by *possible-worlds semantics.* It dates back to Hintikka [Hint62], with more recent investigations in [Moor80] and [Leve82], for example. The major problem with this approach is that it prescribes reasoning that is closed under logical implication, which is strongly believed to be computationally intractable (co-NP hard) even for propositional logics, and is known to be undecidable for first-order logic. Even on intuitive grounds, closure under logical implication has been found much too demanding for real agents, a problem often referred to as *logical omniscience.* For example, logical omniscience has an agent believe all valid sentences, something we are more than willing to give up in return for better performance.

There are mainly two approaches that try to avoid the shortcomings of possible-worlds semantics. One essentially assumes that beliefs are sentences in some syntactically specified *belief set.* Examples of such models are [Eber74] and [MoHe79]. A more sophisticated approach can be found in [Kono84]. As pointed out by Levesque [Leve84), a major drawback of this syntactic approach is the fact that the kinds of sentences believed can be quite arbitrary because they depend on the form of sentences.

The appeal of possible-worlds semantics, despite its problems, is that it avoids relying on syntactic form by defining belief with respect to the classical notion of truth. Researchers following the so-called semantic approach have therefore tried to retain these properties while at the same time avoiding logical omniscience by adopting a modified notion of truth. Levesque [Leve84] was among the first to follow this route. His model of belief employs non-standard worlds, resulting in a kind of tractable inference closely related to relevance logic [AnBe75] (see section II. for more details). Fagin and Halpern develop a logic in [FaHa85] which adds a concept of awareness of primitive proposi-

[1] For the purposes of this paper, the distinction between knowledge and belief is not important, and we use both terms interchangeably.

tions to the standard notion of truth[2].

In a model by Vardi ([Vard86]), the belief set of the syntactic approach is replaced by a set of propositions, where propositions are modelled as sets of certain states of affairs. However, this still forces an agent who believes *a* to believe all sentences that are equivalent to a. (For a comprehensive overview of models of belief in the literature, see [McAr87].)

The semantic approaches mentioned above still leave one important question unanswered, namely whether there are models that allow agents to reason with their meta-belief8 in a tractable way. Even though Vardi and Fagin and Halpern overcome the problem of logical omniscience and allow for meta-beliefs, reasoning still appears to be intractable (see section D. for more details). On the other hand, Levesque, who does provide an efficient algorithm to determine whether certain beliefs imply others, does not allow the agent to have beliefs about itself, precluding any form of meta-reasoning, which is a serious limitation.

The main contribution of this work is that it offers a plausible semantics for the beliefs of an agent who can hold meta-beliefs and is able to draw inferences from them efficiently. The paper is organized as follows: in section 2 we give a general outline of the model discussing its origins and motivations. Section 3 introduces the language *C* and the logic BLK. In addition to a formal semantics, proof theory, and a discussion of its properties, the main tractability result is presented. Section 4 outlines changes to the semantics that allow an agent to do some introspection on its beliefs without losing tractability (resulting in the logic BL4). We conclude with a short summary and an outlook on open questions and future work.

II. The Approach

The model of belief proposed in this paper is an extension of Levesque's model in [Leve84] which is derived from possible-worlds semantics (also referred to as Kripke structures).

A standard Kripke structure consists of a set of *worlds* and a binary *accessibility relation (R)* between worlds. The main characteristic of a world is that it determines the truth of a global set of propositions. An agent at a world *w* is then said to believe a proposition *p* if *p* is true in all worlds accessible from *w.* This is the basic model for what is usually referred to as the logic K (see [HaMo85] or [HuCr68] for an introduction to modal logics). Restrictions on *R* result in models for certain introspective abilities of an agent. If *R* is transitive, for example, we get *positive introspection,* that is, if an agent believes *p* then it believes that it believes *p* (the logic *weak* S4). If, in addition, *R* is Euclidean[3], the agent can perform *negative introspection,* that is, if it does not believe p, it believes that it does not believe *p* (the logic *weak* S5).

In Levesque's work the major deviation from standard Kripke structures is the use of *situations* rather than worlds. In contrast to a world, where everything is either true or false, a situation can support the truth of a proposition, the falsity, both, or neither. Intuitively, only those propositions that are relevant to a situation are supported. A situation in which proposition *p* has both true and false support, can be interpreted as providing evidence both for the truth of *p* and for its falsity. An agent believes a proposition *p* in a situation s, if *p* has true support in all situations accessible from *s.*[4] A more detailed picture of the properties of Levesque's model will follow from the discussion of the logics presented in this paper, which subsume his model. Besides its attractiveness from a computational point of view, it is worth mentioning that, although implication retains the usual properties, believing *a* implies believing B holds if and only if *a tautologically entails B*. Tautological entailment was proposed by relevance logicians as a more intuitive account of implication [AnBe75]. In a sense, the agent in Levesque's model believes all the *relevant* implications of its beliefs.

Our goal is to extend Levesque's semantics in a natural way to allow for meta-beliefs. In particular, beliefs about beliefs should have properties analogous to those that are just about the world. This means that the agent should not be logically omniscient with respect to its own beliefs, and in addition, its reasoning abilities should not be any more powerful when reasoning about itself.[5]

A key feature underlying Levesque's logic is the fact that an agent on the one hand may have no opinion at all about some aspect of the world, and on the other hand, its opinion about something may be unrelated to an opinion about its negation. The extension we propose to this property is to let that "something" also apply to the agent's beliefs about itself.

This is the idea: whenever the agent in a situation *s* wants to confirm a belief a, it does so by making sure that *a* is supported in all elements of some set of situations. Whenever it disconfirms a belief /?, it is because B is not supported by some member of a set of accessible situations. The crucial point is that the sets in both cases need not be the same. It is as if the agent is in (potentially) different modes of thinking when it comes to positive versus negative beliefs of its own. Caveat: We certainly do not want to suggest that humans actually behave like this. Rather, this is an attempt to provide a semantically motivated account for an arguably weak artificial agent. Another aim is

[3] Their logic of *general* awareness where an agent can be aware of arbitrary sentences, however, has a strong syntactic flavour.

[3] *R* is Euclidean iff for all w_1, w_2, and W_3, if w_1 w_2 and $w_1 R w_3$, then $w_2 R w_3$.

[4] Actually, since Levesque does not consider meta-beliefs, he dispenses with the accessibility relation altogether replacing it with a set of possible situations that are visible from any situation.

[5] Both criteria are violated in the obvious extension to Levesque's logic, which simply allows nested beliefs without changing the semantics. In fact, this would lead to a reasoner with the power of *weak* S5 with respect to meta-beliefs.

to stay as close to traditional semantic models as possible. The only change to Kripke structures in addition to allowing situations is the introduction of a second accessibility relation *R*, which is used to determine negative beliefs.

The two logics we are about to formally introduce not only capture this extended notion of explicit belief but are also able to express what is implicit in an agent's belief. Intuitively, by implicit we mean anything that one could possibly deduce given what the agent actually believes. The logic *weak* S5 seems to be an appropriate choice as a model of implicitness (see also [FaHa85] for a similar notion).

In their logic of awareness, Fagin and Halpern allow an agent to have beliefs about what is implicit in its beliefs. At this point, our model is not concerned with those issues because we view implicit beliefs as a purely external characterization of an agent's beliefs and what follows from them. In other words, we assume that the agent does not know about the concept of implicitness (see section V. for further comments).

III. The Logic BLK

A. The Language $\mathcal{L}$

The language $\mathcal{L}$ we are using throughout the paper is a standard propositional language with a countably infinite set $\mathcal{P}$ of propositional letters (or atoms), the logical connectives $\neg$, $\vee$, and $\wedge$, and two modal operators B and L.[6] Sentences of $\mathcal{L}$ are either atoms or are constructed from other sentences, logical connectives or modal operators with the usual formation rules. The only constraint we impose is that no L may appear within the scope of a B. $\mathcal{L}_B$ denotes the sublanguage of $\mathcal{L}$ that does not contain Ls. In a sense, $\mathcal{L}_B$ is the agent's language whereas $\mathcal{L}$ as a whole is used to talk about the agent's beliefs, what is implicit in its beliefs, and the world. Sentences that contain neither B nor L are called *objective*. Sentences where every atom is in the scope of a B or a L are called *subjective*.

B. A Formal Semantics for BLK

Definition 1

A **BLK***-model is a tuple* $M = < S, T, F, R, \bar{R} >$, *where*

1. S *is any set, the set of situations*
2. T *and* F *map* $\mathcal{P}$ *into subsets of* S
3. R, $\bar{R}$ *are binary relations on* S
4. $w \in S$ *is a* world *iff*
   - $w \in T(p) \iff w \notin F(p)$ *for all atoms* p
   - *for all* $s \in S$ $\; wRs \iff w\bar{R}s$
5. *for all* worlds w_1, $w_2 \in S$, *and situations* $s \in S$
   - *if* w_1Rw_2 *and* w_2Rs *then* w_1Rs — *transitivity*
   - *if* w_1Rw_2 *and* w_1Rs *then* w_2Rs — *Euclidean*

[6] We will freely use formulas containing additional connectives like $\supset$ and $\equiv$, where $\alpha \supset \beta$ and $\alpha \equiv \beta$ should be understood in the standard way, viz. as shorthand for $(\neg\alpha \vee \beta)$ and $(\alpha \supset \beta) \wedge (\beta \supset \alpha)$, respectively.

The understanding of T and F is that if $s \in T(p)$ and $t \in F(p)$, then s supports the truth of p and t its falsity. Since T and F map into arbitrary subsets of S, it is easy to see that this captures exactly the notion that a situation can support either the truth, the falsity, both, or neither of a proposition. As noted earlier, the accessibility relations R and $\bar{R}$ capture the intuition that an agent in a given situation s may use different sets of possible situations to confirm or disconfirm a belief. Worlds are, of course, just those situations that look like Kripke worlds. In particular, R and $\bar{R}$ coincide at a world, and as far as worlds are concerned, the model reduces to a transitive and Euclidean Kripke structure. Note that, in addition, a world can see exactly those situations (including non-worlds) that are visible to any world that can reach this world.

Given a model M, we can now define what it means for a situation s in M to support the truth ($\models_T$) or falsity ($\models_F$) of any sentence in $\mathcal{L}$.

1. $M, s \models_T p \iff s \in T(p)$
   $M, s \models_F p \iff s \in F(p)$
2. $M, s \models_T \neg\alpha \iff M, s \models_F \alpha$
   $M, s \models_F \neg\alpha \iff M, s \models_T \alpha$
3. $M, s \models_T \alpha \wedge \beta \iff M, s \models_T \alpha$ and $M, s \models_T \beta$
   $M, s \models_F \alpha \wedge \beta \iff M, s \models_F \alpha$ or $M, s \models_F \beta$
4. $M, s \models_T \alpha \vee \beta \iff M, s \models_T \alpha$ or $M, s \models_T \beta$
   $M, s \models_F \alpha \vee \beta \iff M, s \models_F \alpha$ and $M, s \models_F \beta$
5. $M, s \models_T B\alpha \iff$ for all t, if sRt then $M, t \models_T \alpha$
   $M, s \models_F B\alpha \iff$ for some t, $s\bar{R}t$ and $M, t \not\models_T \alpha$
6. $M, s \models_T L\alpha \iff$ for all worlds w, if sRw then $M, w \models_T \alpha$
   $M, s \models_F L\alpha \iff M, s \not\models_T L\alpha$

For a world w in M and a sentence $\alpha \in \mathcal{L}$ we say that α is *true* in M at w if $M, w \models_T \alpha$; otherwise α is *false*. A sentence α is said to be **BLK**-valid (written $\models\alpha$) iff α is true in all **BLK**-models M and all worlds w in M.

C. Properties of BLK

Our intuition that everything deducible given what is believed is implicit is captured by the following properties of **BLK**:

1. $\models B\alpha \supset L\alpha$
2. $\models B\alpha \supset LB\alpha$
3. $\models \neg B\alpha \supset L\neg B\alpha$
4. $\models L\alpha \supset LL\alpha$
5. $\models \neg L\alpha \supset L\neg L\alpha$

1-3 say in essence that not only *what* is believed is implicit, but also the fact *that* it is or is not believed is implicit. Incidentally, 2 and 3 are not valid in Fagin and Halpern's logic of awareness [FaHa85]. Finally, 4 and 5 indicate that L is essentially a *weak* S5 operator.

Concerning the behaviour of the B operator, we first look at how logical omniscience is avoided. Here we get all the properties of Levesque's logic with the following sentences being satisfiable:

$B\alpha \wedge B(\alpha \supset \beta) \wedge \neg B\beta$	Beliefs are not closed under implication.
$\neg B(\alpha \vee \neg\alpha)$	A valid sentence need not be believed.
$B\alpha \wedge \neg B(\alpha \wedge (\beta \vee \neg\beta))$	A logical equivalent to a belief need not be believed.
$B\alpha \wedge B(\neg\alpha) \wedge \neg B\beta$	Beliefs can be inconsistent without every sentence being believed.

It is important to note that these sentences are satisfiable even when α and β are subjective sentences.

In order to get a better understanding of what kinds of explicit beliefs follow from a given set of beliefs, we first present a proof theory that is both sound and complete with respect to the above semantics:

1. Axioms for standard propositional logic.
2. $\vdash L(\alpha \supset \beta) \supset (L\alpha \supset L\beta)$
3. $\vdash B\alpha \supset L\alpha$
4. $\vdash \sigma \supset L\sigma$, where σ is a subjective sentence[7]
5. $\vdash B\alpha \equiv B\alpha_{CNF}$, where α_{CNF} is α converted into conjunctive normal form (CNF)[8]
6. $\vdash (B\alpha \wedge B\beta) \equiv B(\alpha \wedge \beta)$
7. $\vdash (B\alpha \vee B\beta) \supset B(\alpha \vee \beta)$

   Rules of Inference:
8. if $\vdash \alpha$ and $\vdash (\alpha \supset \beta)$, then $\vdash \beta$
9. if $\vdash \alpha$, then $\vdash L\alpha$
10. if $\vdash (B\alpha \vee B\beta) \supset B\gamma$, then $\vdash B(\alpha \vee \beta) \supset B\gamma$
11. if $\vdash (B\alpha \wedge B\beta) \supset B\gamma$, then
    $\vdash B(B\alpha \wedge B\beta) \supset BB\gamma$ and
    $\vdash \neg B\neg(B\alpha \wedge B\beta) \supset \neg B\neg B\gamma$

Assuming the standard notion of theoremhood we have

Theorem 2 *For any sentence α in $\mathcal{L}$, α is a theorem of* **BLK** *if and only if α is* **BLK***-valid.*[9]

From the proof theory, it is apparent that **BLK** restricted to sentences without nested modal operators reduces to Levesque's logic. All of Levesque's axioms appear (sometimes in a modified form) in **BLK**. The only new axiom and inference rule are 4 and 11, both addressing nested beliefs. Axiom 4 makes sure that L has the right weak S5 properties, whereas rule 11 determines what kinds of explicit beliefs follow from nested beliefs. As a consequence, if we treat sentences of the form $B\alpha$ in the scope of a B as atomic, we get exactly the same relevant implications $\models B\alpha \supset B\beta$ as in Levesque's logic, for example:

$\models B(\alpha \vee \beta) \supset B(\beta \vee \alpha)$

$\models B\alpha \supset B(\alpha \vee \beta)$

$\models B(\alpha \wedge \beta) \supset B\alpha$

In addition, the inference rule 11 allows the agent to perform relevant implications at any level provided the levels are the same on either side of the implication. Examples with beliefs at level two are:

$\models BB(\alpha \vee \beta) \supset BB(\beta \vee \alpha)$

$\models BB\alpha \supset BB(\alpha \vee \beta)$

$\models B\neg B(\alpha \vee \beta) \supset B\neg B\alpha$

$\models BB(\alpha \wedge \beta) \supset BB\alpha$

$\models B(B\alpha \vee B\beta) \supset BB(\alpha \vee \beta)$

D. Computational Properties of BLK

As we have already seen earlier, there are many things a reasoner using this model of belief cannot deduce. A major limitation, for example, is the inability to perform *modus ponens*. The obvious question then is whether we get any mileage out of these limitations in terms of more efficient reasoning. In particular, if the agent's knowledge is represented as a conjunction of sentences KB of $\mathcal{L}_B$, how hard is it then for the agent to determine whether it believes some other sentence α, or more formally, is there a tractable (polynomial time) algorithm to decide whether $\models B\text{KB} \supset B\alpha$?

Before giving the answer, let us first go back to the logic of awareness in [FaHa85] which, as was mentioned in the introduction, does not lend itself to tractability, even when KB and α are in CNF. We can now be a little more precise about this. The problem of deciding $B\text{KB} \supset B\alpha$ is co-NP hard in Fagin and Halperns logic of awareness. Since the proof is straight forward, we will not introduce the logic formally, but rather give an informal argument: let KB and α be objective sentences in CNF and, in particular, let $\alpha = (p \wedge \neg p)$. Since inconsistent sentences are never believed in this logic, $B\text{KB} \supset B\alpha$ is true iff KB is unsatisfiable (in the classical sense), since every satisfiable (objective) sentence can be believed. Deciding satisfiability of CNF formulas is NP-complete and thus the original problem is at least as hard as the complement of an NP-complete problem, for which no known tractable solution exists. (For the other logics in [FaHa85] and [Vard86] a similar argument holds.)

For the logic **BLK**, the problem is in general also too hard. As was shown by Patel-Schneider in [Pate85], the problem is co-NP complete if KB and α are arbitrary objective formulas. However, if α and KB are in CNF, Levesque [Leve84] could show that the problem is $O(|\text{KB}| \times |\alpha|)$, where $|\gamma|$ denotes the length of a sentence γ. (Levesque also notes that deciding whether α is implicit in what the

[7] This axiom schema subsumes the sentences used in properties (2) - (5) of **BLK** above

[8] Equivalently, we could have used axiom schemata that allow conversion into CNF, for example,
$B(\neg(\alpha \wedge \beta)) \equiv B(\neg\alpha \vee \neg\beta)$ or $B(\alpha \vee (\beta \wedge \gamma)) \equiv B((\alpha \vee \beta) \wedge (\alpha \wedge \gamma))$.
Note that subsentences of the form $B\gamma$ are treated as atomic here.

[9] Proofs are generally omitted for reasons of space. They appear in [Lake88].

agent believes is still co-NP complete.) Certainly, converting a formula 7 into CNF can be exponential in |r|. However, in the intended application, if the knowledge of an agent remains relatively stable, it is only *a* that needs to be converted into CNF, where *a* can be assumed to be much smaller than the KB. Also, adding a belief to the KB only involves adding the CNF-version of the new belief without touching the old KB.

In the rest of this section we will demonstrate that a result similar to Levesque's holds for BLK with the important addition that the agent can reason efficiently also about its own beliefs. Since we are allowing meta-beliefs, we first have to introduce a slightly modified version of CNF, which we call *extended conjunctive normal form* (ECNF for short).

Definition 3 *A sentence a in L_B is called an* extended clause *(e-clause) iff*

$$\alpha = \alpha_1 \vee \alpha_2 \vee \ldots \vee \alpha_m ,$$

where each α_i is either a literal (i.e., an atom or its negation) or of the form $B\beta$ or $\neg B\beta$ where β itself is an e-clause.

Definition 4 *A sentence α in $\mathcal{L}_B$ is in* ECNF *iff*

$$\alpha = \alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n$$

and every α_i is an e-clause.

An example of a formula in ECNF is

$$(p \vee B(p \vee r) \vee \neg Bq) \wedge B(q \vee \neg B(\neg s \vee Bp)) \wedge (\neg Bp \vee B(\neg Bp))$$

Lemma 5 *Every α in $\mathcal{L}_B$ can be converted into a formula α_{ECNF} in* ECNF *such that* $\models B\alpha \equiv B\alpha_{\text{ECNF}}$.

The following theorem and corollary essentially give us the specification for an efficient algorithm to determine $\models B\text{KB} \supset B\alpha$ if both α and KB are in ECNF:

Theorem 6 *Let α and β be e-clauses of the form*

$$\alpha = \bigvee_{i=1}^{k-1} \alpha_i \vee \bigvee_{i=k}^{l-1} B\alpha_i \vee \bigvee_{i=l}^{m-1} \neg B\alpha_i \quad \text{and}$$

$$\beta = \bigvee_{j=1}^{n-1} \beta_j \vee \bigvee_{j=n}^{o-1} B\beta_j \vee \bigvee_{j=o}^{p-1} \neg B\beta_j$$

with $\alpha_i (1 \leq i < k)$ and $\beta_j (1 \leq j < n)$ being literals. Then $\models B\alpha \supset B\beta$ if and only if

a) for all i, $1 \leq i < k$, there is a j, $1 \leq j < n$ s.t. $\alpha_i = \beta_j$, and

b) for all i, $k \leq i < l$, there is a j, $n \leq j < o$ s.t. $\models B\alpha_i \supset B\beta_j$, and

c) for all i, $l \leq i < m$, there is a j, $o \leq j < p$ s.t. $\models B\beta_j \supset B\alpha_i$

The if-direction can be proved directly by a simple argument about any **BLK**-model M and world w that support $B\alpha$. The only-if direction involves an induction on the number of Bs occurring in α and β. It is not hard to see that the base case, i.e., α and β are objective sentences, reduces to Levesque's procedure in [Leve84]. Note that we could rephrase a) as $\models B \bigvee_{i=1}^{k-1} \alpha_i \supset B \bigvee_{j=1}^{n-1} \beta_j$. The proof of the only-if direction for part c) crucially depends on the fact that reasoning about explicit beliefs does not allow chaining, for example, $\not\models B(\neg Bq) \supset B(\neg Bp \vee \neg B(\neg p \vee q))$ in contrast to $\models L(\neg Lq) \supset L(\neg Lp \vee \neg L(\neg p \vee q))$. In fact, from part c) we obtain the following corollary:

Corollary 7 *Let α and β be in ECNF, where $\alpha = \bigwedge \alpha_i$ and $\beta = \bigwedge \beta_j$.*
Then $\models B\alpha \supset B\beta$ if and only if for each β_j there is an α_i such that $\models B\alpha_i \supset B\beta_j$.

Proof : The result follows immediately from theorem 6 and the fact that
$\models B(\bigwedge \alpha_i) \supset B(\bigwedge \beta_j) \iff$
$\models B(\bigvee \neg B\beta_j) \supset B(\bigvee \neg B\alpha_i)$. ∎

Given the previous theorem and corollary, it is easy to see that determining $\models B\text{KB} \supset B\alpha$ with KB and α in ECNF and arbitrary nestings of beliefs is no harder than the case where only object level sentences are considered.

Theorem 8 *If* KB *and α are in ECNF, then* $\models B\text{KB} \supset B\alpha$ *can be determined in time* $O(|\text{KB}| \times |\alpha|)$.

Proof : The proof follows a simple induction argument on the number of Bs in both KB and α. Note that the base case corresponds directly to Levesque's result about object level beliefs. ∎

IV. The Logic BL4

So far, we have modelled agents that are able to reason efficiently about their own beliefs, but are only able to draw conclusions from meta-beliefs if they are explicitly represented in their knowledge base. That is, they lack any introspective capability. (Implicit beliefs are, of course, introspective, but the agent does not know about them.) We will now add positive introspection to the agent's abilities, which, as it turns out, does not substantially affect the computational properties of reasoning as discussed for **BLK**.

At first glance it seems that all one has to do is add transitivity to the R relation, which will certainly make $B\alpha \supset BB\alpha$ valid (as it does in standard Kripke structures). However, on closer inspection one notices that this alone does not preserve a nice symmetry between reasoning from positive and negative beliefs which we have in **BLK**, where (1): $BB\alpha \supset BB\beta$ is valid iff (2): $B\neg B\beta \supset B\neg B\alpha$ is. In particular, if we replace β by $B\alpha$, then (1) is valid, but we can easily find models that make (2) false if the only change to the semantics is letting R be transitive. We can regain the above symmetry if we add another transitivity rule which relates both R and $\bar{R}$. These modifications give us the logic **BL4**:

Definition 9 *A model* $M =< S, T, F, R, \bar{R} >$ *is a* **BL4***-model iff*
M is a **BLK***-model and*

- R *is transitive.*
- *for all situations* s, t, *and* u, *if* $s\bar{R}t$ *and* tRu, *then* $s\bar{R}u$.

A sentence α *is* **BL4***-valid iff* α *is true in all* **BL4***-models* M *and worlds* w *in* M.

Note that for relations between worlds nothing changes, since there both R and $\bar{R}$ coincide and are already transitive and Euclidean.

The algorithm to decide whether a belief in ECNF implies another in **BL4** is only a slight extension of the one for **BLK**:

Theorem 10 *Let* α *and* β *be as in theorem 6. Then*
$\models B\alpha \supset B\beta$ *if and only if either*
$\models B\alpha \supset B\beta_j$ *for some* j, $n \leq j < o$, *or*
Conditions a), b), and c) of theorem 6 (with $\models$ *read as* **BL4***-validity).*

The following example illustrates why the new condition is needed:

$$B(p \vee \neg BBq) \supset B(p \vee B(p \vee \neg Bq))$$

We leave it to the reader to verify that this sentence is valid in **BL4** but not in **BLK**.

The analogue of corollary 7 holds in this logic as well. With that it is not hard to show that deciding whether $\models B\mathrm{KB} \supset B\alpha$ for KB and α in ECNF is still polynomial time.

Theorem 11
If KB *and* α *are in ECNF, then* $\models B\mathrm{KB} \supset B\alpha$ *in* **BL4** *can be determined in time* $O(|\mathrm{KB}| \times |\alpha|)$.

Of course, the question arises what happens if we add also negative introspection. In that case it seems that we run into trouble, both from a conceptual and a computational point of view. For one, from $\models B\alpha \supset BB\alpha$ and $\models \neg B\alpha \supset B\neg B\alpha$ one can immediately infer $\models B(\neg B\alpha \vee B\alpha)$, which means that we have brought in a form of logical omniscience through the back door (at least as far as subjective sentences are concerned). Note, however, that this still allows an agent to hold inconsistent beliefs about itself without believing everything. In order to determine $\models B\mathrm{KB} \supset B\alpha$ one now has to test for the validity of $B\alpha$, which in case of objective sentences in CNF is trivial. For formulas in ECNF we have not yet found a complete algorithm to test for validity. We conjecture, however, that it is still tractable, since formulas like $B[B(Bp \wedge (Bp \supset Bq)) \supset BBq]$ need not be valid, that is, there is no obvious way of simulating propositional reasoning inside an ECNF.

V. Summary and Future Work

We have presented a model of explicit belief that avoids the problem of logical omniscience. The major contribution of this paper is that reasoning about beliefs in the sense of deciding $\models_{\mathrm{BKB}} \supset Ba$ is tractable (assuming a certain normal form). This is true even in the case of meta-beliefs on either side of the implication. Furthermore, adding positive introspection does not change the complexity.

There are still many issues left open. One, which is probably more of theoretical interest, is how to allow the agent to know about implicitness. For example, one might want to model the property that the agent thinks it knows exactly what is implicit, i.e., one might want $BBa \equiv BLa$ and $B\neg B\alpha \equiv B\neg L\alpha$ to be valid. (See also [FaHa85], where $B\alpha \equiv BL\alpha$ is valid in the logic of awareness.) Straight forward extensions of BLK and BL4 do not have these properties.

Other issues concern the fact that the logics BLK and BL4 are arguably very weak. For one, they have no notion of consistency, not even with respect to meta-beliefs. For example, the sentence $B(p \wedge \neg Bp)$ is satisfiable, which is a version of G. E. Moore's famous problem (see [Hint62]). In [Lake88] we show that consistency requirements of this form can be accommodated without sacrificing tractable reasoning. Furthermore, the language C itself is not expressive enough for many knowledge representation purposes. Therefore, we are now looking at first-order versions of explicit beliefs. In particular, we are investigating how the results in this paper can be combined with those in [Lake86].

Nevertheless, the logics BLK and BL4 demonstrate already that reasoning about object and meta-beliefs can be done efficiently within a reasonable semantic framework.

Acknowledgements

I am indebted to Hector Levesque for his patience and enthusiasm during our weekly meetings and his comments on earlier drafts of this paper. I would also like to thank Jim des Rivieres and Diane Horton for their suggestions concerning style and contents of the paper. This work was financially supported by a Government of Canada Award and the Department of Computer Science at the University of Toronto.

References

[AnBe75] Anderson, A. R. and Belnap, N. D., *Entailment, The Logic of Relevance and Necessity,* Princeton University Press, 1975.

[Eber74] Eberle, R. A., A Logic of Believing, Knowing and Inferring, *Synthese* 26, 1974, pp. 356-382.

[FaHa85] Fagin, R. and Halpern, J. Y., Belief, Awareness, and Limited Reasoning: Preliminary Report, *Proc. Int. Joint Conf. on AI,* August 1985, pp. 491-501.

[HaMo85] Halpern, J. Y. and Moses, Y. O., A Guide to the Modal Logics of Knowledge and Belief, *Proc. Int. Joint Conf. on Artificial Intelligence,* Los Angeles, CA, August 1985, pp. 480-490.

[Hint62] Hintikka, J., Knowledge and Belief: An Introduction to the Logic of the Two Notions, Cornell University Press, 1962.

[HuCr68] Hughes, G. E. and Cresswell, M. J., An Introduction to Modal Logic, Methuen and Company Ltd., London, England, 1968.

[Kono84] Konolige, K., Belief and Incompleteness, SRI Artificial Intelligence Note 319, SRI International, Menlo Park, 1984.

[Lake86] Lakemeyer, G., Steps Towards a First-Order Logic of Explicit and Implicit Belief, Proc. of the Conf. on Theoretical Aspects of Reasoning about Knowledge, Asilomar, California, 1986, pp. 325-340.

[Lake88] Lakemeyer, G., Ph.D. Thesis, Department of Computer Science, University of Toronto, forthcoming.

[Leve82] Levesque, H. J., A Formal Treatment of Incomplete Knowledge Bases, Tech. Report No. 3, Fairchild Lab. for AI Research, Palo Alto, 1982.

[Leve84] Levesque, H. J., A Logic of Implicit and Explicit Belief, Tech. Rep. No. 32, Fairchild Lab. for AI Research, Palo Alto, 1984.

[McAr87] McArthur, G., Reasoning About Knowledge and Belief: A Review, to appear in: Computational Intelligence, 1987.

[MoHe79] Moore, R. C. and Hendrix, G., Computational Models of Beliefs and the Semantics of Belief-Sentences, Technical Note 187, SRI International, Menlo Park, 1979.

[Moor80] Moore, R. C, Reasoning about Knowledge and Action, Technical Note 181, SRI International, Menlo Park, 1980.

[Pate85] Patel-Schneider, P. F., A Decidable First-Order Logic for Knowledge Representation, Proc. Int. Joint Conf on AI, August 1985, pp. 455-458, (also available as: AI Tech. Report No. 45, Schlumberger Palo Alto Research).

[Vard86] Vardi, M. Y., On Epistemic Logic and Logical Omniscience, Proc. of the Conf on Theoretical Aspects of Reasoning about Knowledge, Asilomar, California, 1986, pp. 293-305.