On Perfect Introspection with Quantifying-in

EXTENDED ABSTRACT

Gerhard Lakemeyer
Institut für Informatik III
Universität Bonn
Römerstr. 164
W-5300 Bonn 1, Germany
e-mail: gerhard@uran.informatik.uni-bonn.de

Abstract

Agents with perfect introspection may have incomplete beliefs about the world, but they possess complete knowledge about their own beliefs. This fact suggests that the beliefs of introspective agents should be completely determined by their objective beliefs, that is, those beliefs that are only about the domain in question and not about other beliefs. Introspection and logical reasoning alone should suffice to reconstruct all other beliefs from the objective ones. While this property has been shown to hold for propositional belief logics, there have so far only been negative results in the case of first-order belief logics with quantifying-in.

In this paper we present a logic of belief with quantifying-in, where the beliefs of a perfectly introspective agent are indeed uniquely determined by the objective beliefs. The result is obtained by weakening the notion of belief of an existing logic that does not have this property.

1 Introduction

In areas such as artificial intelligence or distributed systems, there has been considerable interest in agents with *perfect introspection*, that is, agents, who know everything they know and if they don't know something, then they know that they don't know it. For example, the so-called autoepistemic logics developed in artificial intelligence use introspection to model certain forms of nonmonotonic reasoning(e.g. [Moo85, Kon88, MT89]). In distributed systems, perfect introspection has been found useful in modeling the knowledge ascribed to the various components of a system (e.g. [HM90]).

The most widely used formalism to model the beliefs of introspective agents is *possible-world semantics* [Kri63], whose application to knowledge and belief[1] goes back to Hin-

[1] In this paper, the distinction between knowledge and belief is not important. Hence we will use the two terms interchangeably.

tikka [Hin62, Hin69].[2] One of the properties of perfectly introspective agents is their complete knowledge about themselves. In other words, while such agents may have incomplete beliefs about the world, they always have complete knowledge about their own beliefs by way of their ability to introspect. Thus it seems that the beliefs of a perfectly introspective agent should be completely determined by its *objective beliefs*, that is, those beliefs that are only about the world and not about other beliefs. In other words, given the objective beliefs, it should be possible to reconstruct the others by introspection and logical reasoning alone. Unfortunately, while intuitively compelling, this is not necessarily the case given a sufficiently expressive language of belief. A prominent example where the beliefs of an agent are not determined by the objective ones is a logic by Levesque [Lev84, Lev90], which is among the few attempts to extend autoepistemic logic to the first-order case with *quantifying-in.* This paper demonstrates how Levesque's logic can be modified so that the resulting logic indeed has the property that the beliefs of an agent are determined by the objective ones.

To begin with, for expressively weak logics such as *propositional* logics of belief, our intuitions are indeed captured by the formalisms. For example, in the case of propositional possible-world semantics, Halpern and Moses [HM84], among others,[3] proved that the beliefs of an agent can be reduced to the objective ones. In [LL88], it was shown that this result holds even if we weaken the logic in the sense that beliefs are not necessarily closed under (classical) logical consequence.

In the first-order case, however, we run into problems because of the phenomenon of *quantifying-in.* To illustrate this notion, let us consider a first-order language with a modal operator **B** to refer to beliefs. Quantifying-in simply means that a variable bound outside the scope of a **B** may be used within its scope. Such constructs are important in modeling distinctions such as "knowing who" vs. "knowing that." For example, the sentence $\exists x(\mathtt{Murderer}(x) \wedge \neg \mathbf{B}\mathtt{Murderer}(x))$ as part of the beliefs of a detective tells us that he knows that somebody is a murderer, but that he does not know yet who that person is. There has been considerable interest in quantifying-in among philosophers. An excellent collection of essays on the subject appears in [Lin71]. Different ways of handling quantifying-in in possible-world semantics are described in [Hin69], for example.

In artificial intelligence, one of the best analyzed logics of belief with introspection and quantifying-in is due to Levesque [Lev84, Lev90]. It is particularly appealing because of its simple possible-world semantics, which naturally extends existing formalisms without quantifying-in. As mentioned before, it unfortunately has the property that the beliefs of an agent are generally not uniquely determined by the objective beliefs alone. Quantifying-in is indeed the culprit here because, if we restrict the language to sentences without quantifying-in, the problem disappears.

Besides Levesque's approach, little seems to be known about whether other possible-world approaches to quantifying-in suffer from the same problem. In artificial intelligence, Konolige [Kon92] recently proposed various forms of treating quantifying-in in autoepistemic logic.

[2] An introduction to possible-world models of knowledge and belief can be found in [HM85].

[3] This result was independently obtained by R. Moore, M. Fitting, and J. van Benthem.

In contrast to Levesque's semantics, Konolige's is not based on possible worlds.[4] However, we strongly believe that his suffers from the same problem as Levesque's. In [Lak91], we presented another approach which deals with quantifying-in and which is derived from Levesque's work mainly by weakening the deductive capabilities of an agent. Here the question whether the beliefs are uniquely determined by the objective ones was still open.

In this paper, we present a logic which can be placed in between the logics of [Lev84, Lev90] and [Lak91]. Moreover, we show that this new logic has the desired property that the beliefs of an agent are uniquely determined by the objective beliefs. This logic is weaker than Levesque's because beliefs are no longer closed under first-order logical consequence, but stronger than the one in [Lak91] because beliefs are still closed under *modus ponens*. Introducing such an intermediate logic has two benefits. On the one hand, the result can be shown to extend to the logic of [Lak91]. On the other hand, the formalism is much simpler than in [Lak91], which allows us to focus on the main result without being distracted by unnecessary complications.

In the next section, we introduce Levesque's logic *KL* and demonstrate the problem more rigorously. In Section 3, we present the modified logic and show that it has indeed the property that the beliefs of an agent are uniquely determined by the objective beliefs alone. Except for some proof sketches, proofs are generally omitted from this extended abstract. They can be found in [Lak92a].

2 Levesque's logic *KL*

KL was originally defined in [Lev84] and was extended to an autoepistemic logic in [Lev90]. Here we only deal with the original *KL*, since the problem at hand is independent of the autoepistemic component.

2.1 Syntax and Semantics

The language $\mathcal{L}$ is a modal first-order dialect with equality and function symbols.[5] In addition, there is a countably infinite set N of standard names, which are treated syntactically like constants and are written as $\#1, \#2, \ldots$ How standard names differ from constants semantically will become clear below.

The **formulas** of $\mathcal{L}$ are constructed in the usual way from the atomic formulas, the connectives $\neg$ and $\vee$, the quantifier $\exists$,[6] and the modal operator **B**. We also require that all variables that are bound within the scope of a **B** are distinct. This is done merely as a technical convenience, which simplifies the modified semantics proposed in the next section. Given the usual meaning of free variables, **sentences** are formulas without free variables. Functions and predicates applied to standard names are called **primitive** terms and sentences, respectively. The sets of all closed terms, primitive terms and primitive sentences are denoted as $\mathcal{T}$, $\mathcal{T}_{\mathrm{PRIM}}$,

[4] Instead, he uses a fixed-point approach similar to [Moo85, Kon88] to define the beliefs of an agent.

[5] *Constants* are regarded as 0-ary function symbols.

[6] Other logical connectives like $\wedge$, $\supset$, and $\equiv$ and the quantifier $\forall$ are used freely and are defined in the usual way in terms of $\neg$, $\vee$, and $\exists$.

and $\mathcal{P}_{\text{PRIM}}$, respectively. A formula is called **subjective** if all predicate and function symbols appear within the scope of a **B**, and **objective** if it does not contain any **B**'s. **Literals** and **clauses** have their usual meaning.

Sequences of terms or variables are sometimes written in vector notation. E.g., a sequence of variables $\langle x_1, \ldots, x_k \rangle$ is abbreviated as $\vec{x}$. If a formula α contains the free variables $x_1, \ldots, x_k$, $\alpha[x_1/t_1, \ldots, x_k/t_k]$ (abbreviated as $\alpha[\vec{x}/\vec{t}]$) denotes α with every occurrence of x_i replaced by t_i.

For the semantics, we need to specify when an atomic sentence is *true* and when a sentence is *believed.* The truth of an arbitrary sentence is then defined by the usual recursive rules.

While the truth of atomic sentences is determined by *worlds*, belief is modeled in possible-world fashion. The idea is that an agent at a given world is thought to imagine a set of possible worlds and the agent's beliefs are simply all those sentences that are true in all worlds the agent imagines. Since we are interested in fully introspective agents, a very simple approach will do in this case. In particular, we can assume that the *same* set of worlds is imagined (or accessible) at every world of a given model.

For the specification of worlds, standard names play a special role. Standard names were originally introduced by Kaplan [Kap71] as a means to adequately deal with quantifying-in. Intuitively, a standard name is a name that always denotes the same entity. For example, the numeral '9' comes close to being a standard name, since virtually everyone associates the same number with it. As for quantifying-in, standard names provide an elegant and simple way to give meaning to the notion of "knowing who." For example, a detective can be said to know who the murderer is if he not only knows that some individual is the murderer but also knows the standard name of that person.

The meaning of standard names can be captured in a very intuitive way in possible-world semantics. One merely requires that a standard name denotes the same domain element in *all* possible worlds. Non-standard names, on the other hand, are allowed to vary in their denotation at different worlds. Following terminology introduced by Kripke [Kri80], we also call standard names *rigid designators* and non-standard names *non-rigid designators.*

With standard names denoting the same individual at all worlds, the difference between "knowing who" and "knowing that" has a very simple characterization. For example, if a detective knows who the murderer is, then the same individual with a certain standard name is the murderer in all the worlds the detective imagines. On the other hand, just knowing that somebody is the murderer means that there is a murderer in every accessible world, but it may be a different person in different worlds.

In *KL* it is assumed that all worlds have the *same* domain of discourse, which is, in addition, *isomorphic* to the set of standard names, i.e., every domain element has a unique standard name. One advantage of such an approach is that, while being relatively simple, it still provides a high degree of expressiveness. As argued in [Lev84], the commitment to a countably infinite domain is of little practical concern in areas such as knowledge representation. In particular, an infinite domain does not prevent us from considering predicates with finite extensions.

With the assumption that the domain of discourse is isomorphic to the set of standard names, there is actually no need to distinguish between the two, and the set of standard names is used as *the* domain. This, by the way, also greatly simplifies the technical aspect of the semantics. For example, it allows quantification to be understood substitutionally.[7]

Since standard names represent distinct elements of the universe, a world can now simply be specified as a mapping from the *primitive* sentences into the truth values $\mathbf{t}$ and $\mathbf{f}$ and from the *primitive* terms into the standard names.

Definition 1 (Worlds)
A world w is a function $w : \mathcal{P}_{\text{PRIM}} \cup \mathcal{T}_{\text{PRIM}} \longrightarrow \{\mathbf{t}, \mathbf{f}\} \cup N$ such that

1. $w[\mathtt{P}(n_1, \ldots, n_k)] \in \{\mathbf{t}, \mathbf{f}\}$ *for all primitive sentences* $\mathtt{P}(n_1, \ldots, n_k)$.
2. $w[n = m] = \mathbf{t}$ *iff n and m are identical standard names.*
3. $w[f(n_1, \ldots, n_k)] \in N$ *for all primitive terms* $f(n_1, \ldots, n_k)$.

Note the special treatment of =, which reflects the fact that each standard name represents a unique individual.

Given a world w and its assignment of standard names to the primitive terms, we can define the denotation of arbitrary terms recursively.

Definition 2 (Denotation Functions)
A world w defines a denotation function $d_w : \mathcal{T} \longrightarrow N$ with the following properties:

1. $d_w[n] = n$ *for all standard names n.*
2. $d_w[f(n_1, \ldots, n_k)] = w[f(n_1, \ldots, n_k)]$ *for all primitive terms* $f(n_1, \ldots, n_k)$.
3. $d_w[f(t_1, \ldots, t_k)] = w[f(d_w[t_1], \ldots, d_w[t_k])]$ *otherwise.*

For sequences of terms $\vec{t} = \langle t_1, \ldots, t_k \rangle$ we write $d_w(\vec{t})$ instead of $\langle d_w(t_1), \ldots, d_w(t_k) \rangle$.

Note that standard names denote themselves because they *are* the fixed universe of discourse for all worlds.

With these definitions, we are now able to specify what it means for a world w and set of worlds M to satisfy an arbitrary sentence α ($M, w \models \alpha$). In the following, let $\mathtt{P}(\vec{t})$ be an atomic sentence, and let α and β be sentences except in rule 4, where α may contain the free variable x.

$$\begin{array}{lcl}
M, w \models \mathtt{P}(\vec{t}) & \Longleftrightarrow & w[\mathtt{P}(d_w[\vec{t}])] = \mathbf{t} \\
M, w \models \neg\alpha & \Longleftrightarrow & M, w \not\models \alpha \\
M, w \models \alpha \vee \beta & \Longleftrightarrow & M, w \models \alpha \text{ or } M, w \models \beta \\
M, w \models \exists x \alpha & \Longleftrightarrow & M, w \models \alpha[x/n] \text{ for some } n \in N \\
M, w \models \mathbf{B}\alpha & \Longleftrightarrow & \text{for all } w', \text{ if } w' \in M, \text{ then } M, w' \models \alpha
\end{array}$$

[7] It should be noted that Levesque's standard-name approach is by far not the only way of dealing with quantifying-in. Some interesting alternatives are presented in [Kon92]. Other approaches in the context of possible-world semantics are discussed in [Hin69], for example.

A sentence α is valid ($\models\alpha$) iff $M, w\models\alpha$ for all worlds w and non-empty sets of worlds M.

Notation: We often write $M\models\alpha$ if α is subjective, since the truth of α depends solely on the set of worlds M. Similarly, we write $w\models\alpha$ for objective α.

2.2 A Proof Theory

The following proof theory, which is a variant of the proof theory given in [Lev84], shows that *KL* is essentially *weak S5* with consistency (or *KD45*). This is not surprising since belief is modeled using a globally accessible and non-empty set of worlds M (see, for example, [HM85]).

Axioms

A1 Axioms of first-order logic with a restricted axiom of specialization
A2 $n_i = n_i \wedge n_i \neq n_j$ for all distinct standard names n_i and n_j
A3 $\mathbf{B}\alpha$ for all axioms α under **A1** and **A2**
K $\mathbf{B}(\alpha \supset \beta) \supset (\mathbf{B}\alpha \supset \mathbf{B}\beta)$
D $\mathbf{B}\alpha \supset \neg\mathbf{B}\neg\alpha$
4 $\mathbf{B}\alpha \supset \mathbf{B}\mathbf{B}\alpha$
5 $\neg\mathbf{B}\alpha \supset \mathbf{B}\neg\mathbf{B}\alpha$
BF $\forall x\mathbf{B}\alpha \supset \mathbf{B}\forall x\alpha$

Inference Rules

MP From α and $\alpha \supset \beta$ infer β
UG From $\alpha[x/n_1], \ldots, \alpha[x/n_k]$ infer $\forall x\alpha$,
where the n_i range over all standard names occurring in α and one not in α.

The only non-standard aspect of the logic is due to standard names, which serve as the universe of discourse for all worlds and, at the same time, are part of the language. As a result we need an axiom specifying that all standard names are distinct (Axiom **A2**). Since the logic allows for quantifying-in, the normal first-order axiom of specialization ($\forall x\alpha \supset \alpha[x/t]$) for any closed term t needs to be modified. In particular, we have to require that specialization applies only if t is rigid (a standard name) or if x does not occur within the scope of a $\mathbf{B}$. The fact that we generally cannot substitute arbitrary terms for x is illustrated by the sentence $\forall x\neg\mathbf{B}\mathtt{Murderer}(x)\wedge\mathbf{B}\mathtt{Murderer}(\mathit{driver}(\mathit{red\_car}))$, which is *satisfiable* and which captures the intuition that a detective may know that the driver of the red car is the murderer without knowing who that individual is. Finally, the rule of universal generalization (**UG**) is somewhat more complicated than in classical logic. Instead of just one arbitrary standard name we have to instantiate all names occurring in the formula plus one name not occurring in the formula. Again this has to do with the special meaning of standard names together with the meaning of equality in *KL*.

More details about the properties of *KL* can be found in [Lev84].

2.3 The Problem with Epistemic States in *KL*

We now turn to the problem that the beliefs of an agent as modeled by *KL* are not uniquely determined by the objective beliefs. For that purpose, let us define an *epistemic state* in *KL* as the set of all sentences believed at some set of worlds.

Definition 3 (Epistemic State) *A set of sentences* Γ *is called an epistemic state iff there is a set of worlds* M *such that* $\Gamma = \{\alpha \mid M \models \mathbf{B}\alpha\}$.

If the language did not permit quantifying-in,[8], epistemic states would indeed be uniquely determined by their objective sentences. (For example, a proof in [HM84] for a propositional language could easily be lifted to the quantificational case.) However, quantifying-in destroys this intuitive property, as shown in [Lev81, Lev90]. In particular, Levesque shows that there are epistemic states that agree on all objective sentences yet disagree on the sentence

$$\delta = \exists x(\mathrm{P}(x) \land \neg\mathbf{B}\mathrm{P}(x)).$$

Theorem 1 (Levesque) *Epistemic states are in general not uniquely determined by the objective sentences they contain.*

Proof : (Sketch) The proof is constructive and provides two sets of worlds which agree on all objective beliefs yet disagree on believing δ.[9]

Let P be a unary predicate and let $n_1, n_2, n_3, \ldots$ be an ordering of the standard names. For this ordering, let EVEN $= \{n_2, n_4, n_6, \ldots\}$ and ODD $= \{n_1, n_3, n_5, \ldots\}$.
Let $\overline{\mathcal{T}^{\mathrm{P}}} = \mathcal{T}_{\mathrm{PRIM}} - \{\mathrm{P}(n) \mid n \in N\}$. We now construct two sets of worlds M_1 and M_2, where $\mathrm{P}(n)$ is known for all odd n at both M_1 and M_2. The only difference between the two sets is that every world in M_2 is forced to satisfy at least one $\mathrm{P}(m)$ for an even m. (For simplicity, everything other than P is known to be false and all terms are known to be equal to n_1 at both M_1 and M_2.) More formally, let

$$M_1 = \{w \mid \text{for all } n \in \text{ODD}, w \models \mathrm{P}(n);\ \text{for all } \alpha \in \overline{\mathcal{T}^{\mathrm{P}}}, w \models \neg\alpha;\ \text{for all closed terms } t,\ w \models t = n_1\}$$
$$M_2 = \{w \mid \text{for all } n \in \text{ODD}, w \models \mathrm{P}(n);\ \text{for all } \alpha \in \overline{\mathcal{T}^{\mathrm{P}}}, w \models \neg\alpha;\ \text{for all closed terms } t,\ w \models t = n_1;$$
$$\text{for some } m \in \text{EVEN}, w \models \mathrm{P}(m)\}.$$

From these definitions it is clear that $M_2 \subseteq M_1$. The worlds that are in M_1 but not in M_2 are just those which satisfy all of the odd P's and *none* of the even P's. With that it is easy to show that $M_2 \models \mathbf{B}\delta$ yet $M_1 \models \neg\mathbf{B}\delta$. On the other hand, the two sets can be shown to agree on all objective beliefs. (Levesque's original proof in [Lev81] goes through with only minor modifications to take care of function symbols.) ■

As an immediate consequence of this theorem, Levesque shows that *KL* is *irreducible*, that is, given a sentence α, it is generally *not* the case that there is a sentence α' such that **B** occurs

[8] Not permitting quantifying-in means that all the subformulas $\mathbf{B}\alpha$ of a sentence must be *sentences*, that is, they may not contain free variables.

[9] Levesque's original proof in [Lev81] considered the slightly simpler case of a language without function symbols.

non-nested in α' and $\models\alpha \equiv \alpha'$. In particular, the sentence $\mathbf{B}[\exists x(\mathtt{P}(x)\wedge\neg\mathbf{B}\mathtt{P}(x))]$ is irreducible. (Note that *KL*, when restricted to sentences without quantifying-in, is reducible.)

The sentence δ expresses the fact that someone is a P yet the agent does not know who that individual is, that is, he does not know the individual's standard name. Now let us consider the sentence $\delta' = \mathtt{P}(c)\wedge\forall x(x = c \supset \neg\mathbf{B}\mathtt{P}(x))$ for some constant c, which says that c is a P yet it is not known of whoever refers to c that he or she is a P. The main difference between δ and δ' is that the former merely talks about the existence of an unknown P, while the latter, in addition, has a name for it, if only a non-standard one. At first sight, this additional piece of information does not seem terribly informative. After all, c could be simply a Skolem constant or some internal identifier an agent chooses at random. However, it turns out that, while $\mathbf{B}\delta'$ may or may not be reducible, it at least has the property that any two epistemic states which agree on all objective sentences also agree on δ'.

Lemma 2.1 *Let M and M' be two sets of worlds such that the corresponding epistemic states contain exactly the same objective sentences. Then $M\models\mathbf{B}\delta'$ iff $M'\models\mathbf{B}\delta'$.*

Proof: (Sketch) The key observation in proving this lemma is that any M that believes δ' also believes $c \neq n^*$ for all standard names n^* for which $\mathtt{P}(n^*)$ is believed. Therefore any M' that has the same objective beliefs as M believes $\mathtt{P}(c)$ as well as those $c \neq n^*$, which implies that M' believes δ'. ■

3 The Logic KL^-

With this observation, the idea to modify *KL* is as follows. We give existentially quantified variables within belief a *constructive* interpretation so that (roughly) an agent believes an existential only if he has a *name* for it.[10] More concretely, we restrict the semantics of belief in such a way that a sentence with existentially quantified variables is believed iff a corresponding sentence with all existentially quantified variables replaced by terms, which need not be standard names, is believed. The effect is that a sentence such as δ can only be believed if there is a sentence of the form δ' which is believed as well, thus guaranteeing that any two epistemic states that agree on their objective sentences also agree on all other sentences.

The constructive interpretation of existential quantifiers within belief has an intuitionistic flavor. However, the differences between our logic and intuitionistic logic [vDa86] are considerable, since in our case disjunction and negation retain their classical interpretation.

3.1 Syntax and Semantics

The language of KL^- is the same as for *KL* except that we add a new primitive called **f-term**. An f-term is a term other than a variable or a standard name followed by the

[10] This idea was previously used in [Lak91] and [Pat87]. These logics, however, go much further in that they also employ a four-valued semantics, which is not necessary for our purposes.

symbol $^{\triangle}$. F-terms allow us to import the denotation of a term occurring inside a belief from outside that belief. For example, given a set of worlds M, predicate $\mathtt{P}$ and constant a, we can express the fact that we believe $\mathtt{P}(a)$ at M yet that we don't know *who* a is as *for all* $w^* \in M$, $M, w^* \models \mathtt{P}(a) \wedge \neg\mathbf{B}\mathtt{P}(a^{\triangle})$. Here the denotation of a inside the $\mathbf{B}$ is determined at w^*. This technical device is important to give a substitutional account of existential quantifiers within belief in the context of quantifying-in, where we want to substitute arbitrary terms, not just standard names as in KL.

For those who are suspicious of f-terms, Lemma 3.3 shows that they are indeed dispensable. However, they are useful in presenting the semantics of belief in a more intuitive fashion.

An **extended term** is either a term, an f-term, or of the form $f(t_1, \ldots, t_k)$, where f is a k-ary function symbol and the t_i are extended terms. (In other words, f-terms may not occur nested within an extended term.)[11]

Atomic formulas (or atoms) are now predicate symbols with extended terms as arguments. Formulas and sentences are then defined as usual.

Before turning to the formal semantics of KL^-, we need to introduce some notation and terminology with regards to the substitution of terms for existentially quantified variables.

Definition 4 *A primitive or formula contained in a formula α occurs at the* **objective level** *of α if it does not occur within the scope of a modal operator.*

Definition 5 *Existentially Quantified Variables*
Let α be a formula. Let x be a variable that is bound at the objective level of some formula β such that either $\beta = \alpha$ or $\mathbf{B}\beta$ is a subformula of α.
x is said to be **existentially (universally) quantified** *in α iff x is bound within in the scope of an* **even (odd)** *number of $\neg$-operators in β.*

For example, in $\exists x \neg\mathbf{B}\exists y \mathtt{P}(x, y)$, both x and y are considered existentially quantified.

Definition 6 *Admissible Terms*
Let α be a formula and x existentially quantified in α. A term t is said to be an **admissible** *substitution for x with respect to α iff every variable y in t is universally quantified in α and x occurs within the scope of y.*

If the context is clear, we often say t is admissible for *x or t* is admissible.

Definition 7 *Let α be a sentence and let $\vec{x} = \langle x_1, \ldots, x_k \rangle$ be a sequence of the existentially quantified variables bound at the objective level of α. $\alpha^{\nexists}$ denotes α with all $\exists x_i$ removed.*

Example 3.1 Let $\alpha = \exists w \forall x \exists y \mathtt{P}(w, x, y) \wedge \mathbf{B}\exists z \mathtt{Q}(z)$. Then $\alpha^{\nexists} = \forall x \mathtt{P}(w, x, y) \wedge \mathbf{B}\exists z \mathtt{Q}(z)$.

Note that existential quantifiers within modalities are left untouched.

Definition 8 *A formula α is* **existential-free** *iff α contains no existential (universal) quantifiers at the objective level within the scope of an even (odd) number of $\neg$-operators.*

[11] For example, $f(a)^{\triangle}$ and $g(a, b^{\triangle}, h(a^{\triangle}))$ are f-terms. $f(x^{\triangle})$, where x is a variable, $\#27^{\triangle}$ and $g(a, b^{\triangle}, h(a^{\triangle})^{\triangle})$ are not.

Definition 9 *Let α be a formula with free variables $\vec{x} = \langle x_1, \ldots, x_k \rangle$. ($\alpha$ may contain other free variables as well.) Let $\vec{t} = \langle t_1, \ldots, t_k \rangle$ be a sequence of terms.*
$\alpha\mathbf{[}\vec{x}/\vec{t}\mathbf{]}$ is α with every occurrence of x_i at the objective level replaced by t_i and every occurrence of x_i inside the scope of a modal operator replaced by $t_i{}^{\triangle}$ if t_i is neither a variable nor a standard name and by t_i otherwise.[12]

Example 3.2
Let a and b be constants and let $\alpha = \mathtt{P}(x_1) \wedge \mathbf{B}(\mathtt{Q}(x_1, x_2, x_3))$.
Then $\alpha\mathbf{[}x_1/a, x_2/b, x_3/\#27\mathbf{]} = \mathtt{P}(a) \wedge \mathbf{B}(\mathtt{Q}(a^{\triangle}, b^{\triangle}, \#27))$.

Note the difference to $\alpha[x_1/a, x_2/b, x_3/\#27] = \mathtt{P}(a) \wedge \mathbf{B}(\neg\mathtt{Q}(a, b, \#27))$, that is, $[\ldots]$ indicates regular substitutions, while $\mathbf{[}\ldots\mathbf{]}$ indicates that substitutions within modalities use f-terms.

Definition 10 *Let α be a sentence and w a world with corresponding denotation function d_w. α^w is obtained from α by replacing every occurrence of $t^{\triangle}$ by the standard name $d_w(t)$, if t is closed, and by t otherwise.*

Example 3.3 Let $\alpha = \mathtt{P}(a^{\triangle}) \wedge \forall x \mathbf{B}(\mathtt{Q}(f(x)^{\triangle}) \vee \mathtt{R}(a^{\triangle}))$ and w a world with $w[a] = \#1$. Then $\alpha^w = \mathtt{P}(\#1) \wedge \forall x \mathbf{B}(\mathtt{Q}(f(x)) \vee \mathtt{R}(\#1))$.

We are now ready to present the semantics of KL^-. Note that only the rule for $\mathbf{B}$ has changed compared to the semantics of KL.

$M, w \models \mathtt{P}(\vec{t}) \iff w[\mathtt{P}(d_w[\vec{t}])] = \mathbf{t}$
$M, w \models \neg\alpha \iff M, w \not\models \alpha$
$M, w \models \alpha \vee \beta \iff M, w \models \alpha$ or $M, w \models \beta$
$M, w \models \exists x \alpha \iff M, w \models \alpha[x/n]$ for some $n \in N$
Let $\vec{x} = \langle x_1, \ldots, x_k \rangle$ be a sequence of the existentially quantified variables bound
at the objective level of α.
$M, w \models \mathbf{B}\alpha \iff$ there are admissible $\vec{t}$ such that for all $w' \in M$, $M, w' \models \alpha^{w\sharp}\mathbf{[}\vec{x}/\vec{t}\mathbf{]}$

Note the use of α^w in the last rule, where all closed f-terms within $\mathbf{B}\alpha$ are replaced by their *denotation* relative to the current world w.

As in KL we say that a sentence α is valid in KL^- ($\models \alpha$) iff $M, w \models \alpha$ for all worlds w and non-empty sets of worlds M.

3.2 Some Properties of Belief in KL^-

It is not hard to see that KL and KL^- coincide with respect to sentences without f-terms and without existentially quantified variables within beliefs. In general, however, belief in KL^- is weaker than in KL because of the more restrictive interpretation of existentially quantified variables, which it shares with the logic of [Lak91]. So far we have not obtained a complete

[12] Standard names and variables are exceptions because they can not be made into f-terms.

axiomatization of KL^-. In this subsection, we list the main properties of KL^- comparing them to KL and the logic of [Lak91].

First of all, despite the restricted treatment of existentially quantified variables, belief in KL^- remains closed under modus ponens and beliefs are consistent, in contrast to [Lak91].

$$\models \mathbf{B}(\alpha \supset \beta) \supset (\mathbf{B}\alpha \supset \mathbf{B}\beta)$$
$$\models \mathbf{B}\alpha \supset \neg\mathbf{B}\neg\alpha$$

As in KL, agents have perfect introspection, which is not surprising since we have not changed the structure of the possible-world model, namely a globally accessible set of worlds.

$$\models \mathbf{B}\alpha \supset \mathbf{B}\mathbf{B}\alpha$$
$$\models \neg\mathbf{B}\alpha \supset \mathbf{B}\neg\mathbf{B}\alpha$$

As far as the properties of quantifiers are concerned, KL^- behaves just like the logic of [Lak91]. The main difference compared to KL is that beliefs are not closed under first-order logical consequence because existential generalization from disjunctions is no longer valid. For example,

$$\not\models \mathbf{B}(\mathrm{P}(a) \vee \mathrm{P}(b)) \supset \mathbf{B}\exists \mathrm{P}(x).^{13}$$

While not immediately apparent, one of the consequences of this limitation of belief is that not all normal form transformations of classical logic (and of KL) are valid for belief. For example, beliefs in general do not have an equivalent prenex conjunctive normal form. (Some of the normal form transformations that do hold are discussed in the next subsection.)

The properties of quantifying-in, again shared with [Lak91], are much the same as those of KL. In particular, we obtain the same distinctions between "knowing who" and "knowing that".

$$\models \exists x\mathbf{B}\alpha \supset \mathbf{B}\exists x\alpha$$
$$\not\models \mathbf{B}\exists x\alpha \supset \exists x\mathbf{B}\alpha$$

Finally, in contrast to KL, the Barcan formula $\forall x\mathbf{B}\alpha \supset \mathbf{B}\forall x\alpha$ is valid only for existential-free sentences. However, the converse is valid without restrictions, just as in KL.

3.3 Epistemic States are Objectively Determined

We now turn to the main result of the paper (Theorem 2), namely that epistemic states in KL^- are uniquely determined by their objective sentences.

To be compatible with the language of KL and to simplify things, we restrict epistemic states in KL^- to sentences *without* f-terms, which we call **ordinary** sentences.[14]

Definition 11 (Epistemic States in KL^-) *A set of sentences Γ is called an epistemic state iff there is a set of worlds M such that $\Gamma = \{\alpha \mid \alpha$ is ordinary and $M \models \mathbf{B}\alpha\}$.*

[13] It is easy to come up with an example, where the implication fails. Simply take a set of two worlds w_1 and w_2 such that P(a) holds at w_1, P(b) holds at w_2, and for no term t, P(t) holds at both w_1 and w_2.

[14] None of the following results hinges on this restriction.

Definition 12 *Semi Negation Normal Form*
A formula α is in semi negation normal form (SNNF) iff no quantifier, disjunction, conjunction, or negation at the objective level of α occurs within a negation.

Lemma 3.1 *Let α and α' be ordinary sentences such that α' is α transformed into SNNF. Then $\models \mathbf{B}\alpha \equiv \mathbf{B}\alpha'$.*

Definition 13 *Semi Prenex Conjunctive Normal Form*
A formula α is said to be in **semi prenex conjunctive normal form** *(SPCNF) iff all the quantifiers are to the left of the formula and the matrix, that is, the quantifier free part of the formula, consists of a conjunction of disjunctions, where every disjunct is either a literal or of the form $\mathbf{B}\beta$ or $\neg\mathbf{B}\beta$ and β is an arbitrary formula.*

Lemma 3.2 *Let α and α' be ordinary existential-free sentences such that α' is α transformed into SPCNF. Then $\models \mathbf{B}\alpha \equiv \mathbf{B}\alpha'$.*

Note that the lemma does not hold for sentences which are not existential-free. For example, $\not\models \mathbf{B}[(\forall x \exists y \mathrm{P}(x,y)) \vee (\forall u \exists v \mathrm{Q}(u,v))] \equiv \mathbf{B}\forall u \exists v \forall x \exists y (\mathrm{P}(x,y) \vee \mathrm{Q}(u,v))$.

Lemma 3.3 *Let α be an ordinary sentence, x a variable that occurs free in α, y a variable that occurs nowhere in α, and let t be a closed term.*

1. $\models \mathbf{B}\alpha[x/t^{\triangle}] \equiv \forall y(y = t \supset \mathbf{B}\alpha[x/y]$
2. $\models \neg\mathbf{B}\alpha[x/t^{\triangle}] \equiv \forall y(y = t \supset \neg\mathbf{B}\alpha[x/y]$.

Definition 14 *Flat Formulas*
A formula α is called **flat** *iff no existentially quantified variable that is bound at the objective level of α occurs within a subformula $\mathbf{B}\beta$ of α.*

Using Lemma 3.3, it is possible to convert any belief into an equivalent flat belief. Thus we obtain:

Lemma 3.4 *Epistemic states are uniquely determined by their flat sentences.*

Lemma 3.5 *Epistemic states are uniquely determined by their existential-free sentences.*

Proof : By Lemma 3.4, it suffices to show that two epistemic states with the same existential-free sentences agree on their flat sentences.

Let Γ and Γ' be two epistemic states with corresponding sets of worlds M and M'. Let α be a flat sentence such that $M \models \mathbf{B}\alpha$. Then there are admissible terms $\vec{t}$ for the existentially quantified variables $\vec{x}$ at the objective level of α such that $M \models \mathbf{B}\alpha^{\not\exists}[\vec{x}/\vec{t}]$. (Note that we can simply substitute $\vec{t}$ for $\vec{x}$ because α is flat.) Since $\alpha^{\not\exists}$ is existential-free, we obtain $M' \models \mathbf{B}\alpha^{\not\exists}[\vec{x}/\vec{t}]$ by assumption. Thus $M' \models \mathbf{B}\alpha$.
The reverse direction is completely symmetric. ■

Lemma 3.6 *Let $\forall\vec{x}(\sigma \vee \gamma)$ be an existential-free ordinary sentence in SPCNF, where σ is subjective and γ objective. Then $\models \mathbf{B}\forall\vec{x}(\sigma \vee \gamma) \equiv \forall\vec{x}(\sigma \vee \mathbf{B}\gamma)$.*

Theorem 2 *Epistemic states are uniquely determined by their objective sentences.*

Proof: (Sketch) By Lemma 3.5, it suffices to show that two epistemic states Γ and Γ' that agree on their objective sentences also agree on their existential-free sentences.

Let Γ and Γ' be epistemic states with corresponding sets of worlds M and M'. Let α be an existential-free ordinary sentence. By Lemma 3.2, we can assume, without loss of generality, that α is in SPCNF. Let $\alpha = \forall\vec{x}\bigwedge\alpha_i$. Since $\forall\vec{x}\bigwedge\alpha_i$ is existential-free, it is easy to see that $\models \mathbf{B}\forall\vec{x}\bigwedge\alpha_i \equiv \bigwedge\mathbf{B}\forall\vec{x}\alpha_i$. Thus it suffices to treat the case $\alpha = \forall\vec{x}\beta$, where $\beta = \mathbf{B}\beta_1 \vee \ldots \vee \mathbf{B}\beta_m \vee \neg\mathbf{B}\beta_{m+1} \vee \ldots \vee \neg\mathbf{B}\beta_n \vee \gamma$ and γ is a disjunction of literals.
The main proof proceeds by induction on the the nesting of $\mathbf{B}$'s in α using Lemma 3.6. ■

Corollary 3.7 *Epistemic states are uniquely determined by their objective and existential-free sentences.*

Theorem 2 can be shown to apply to the logic of [Lak91] as well. The main difference between KL^- and the logic of [Lak91] is that the latter uses four-valued worlds (called *situations*) instead of the two-valued ones defined in this paper. Intuitively, Theorem 2 holds in such a case as well because nothing in the proof depends on whether worlds are two- or four-valued. (A proof can be found in [Lak92b].)

Theorem 3 *Epistemic states in the logic of [Lak91] are uniquely determined by their objective sentences.*

It is perhaps interesting to note that Theorem 2 hinges on the availability of equality as a built-in predicate. More formally, let KL^{--} be the logic obtained from KL^- by removing the =-predicate from the language.

Theorem 4 *Epistemic states in KL^{--} are in general not uniquely determined by the objective sentences they contain.*

Proof: (Sketch)
Just as in the case of Levesque's KL (Theorem 1), it suffices to construct two sets of worlds M_1 and M_2 that agree on all objective beliefs yet disagree on believing $\delta = \exists x(\mathrm{P}(x) \wedge \neg\mathbf{B}\mathrm{P}(x))$.

As in the proof of Theorem 1, let $n_1, n_2, n_3 \ldots$ be an ordering of the standard names. Let

$$W_0 = \{w \mid w \models \mathrm{P}(n) \text{ iff } n = n_1;\ w[a] = n_1\}.$$

$$W_i = \{w \mid w \models \mathrm{P}(n) \text{ iff } n = n_1 \text{ or } n = n_{2i};\ w[a] = n_{2i}\}.$$

Using these sets of worlds, M_1 and M_2 can now be defined as follows:

$$M_1 = \bigcup_{i=0}^{\infty} W_i.$$

$M_2 = \bigcup_{i=1}^{\infty} W_i$.

It can be shown that both M_1 and M_2 have the same objective beliefs, which are those that follow from believing $\mathtt{P}(n_1) \wedge \mathtt{P}(a)$. On the other hand, $M_1 \not\models \mathbf{B}\delta$ and $M_2 \models \mathbf{B}\delta$. ■

As a final note, while epistemic states are uniquely determined by their objective sentences, it is not at all clear whether or not KL^- itself is reducible, that is, whether or not every sentence is equivalent to a sentence without nested beliefs.[15] Note that, in contrast, the irreducibility of KL follows immediately from the fact that epistemic states in KL are not uniquely determined by their objective sentences.

4 Summary

In this paper we presented a first-order belief logic with quantifying-in, where an agent's epistemic state is uniquely determined by its objective sentences. The logic was derived from Levesque's logic KL, which does not have this property, by giving up the requirement that beliefs are closed under first-order logical consequence. Our result also extends to an existing logic with an even more limited notion of belief.

Acknowledgements

I would like to thank Anil Gupta for pointers to additional literature and Jürgen Kalinski for comments on an earlier draft of the paper.

References

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

[HM85] Halpern, J. Y. and Moses, Y. O., A Guide to the Modal Logics of Knowledge and Belief, in *Proc. of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA, 1985, pp. 480–490.

[HM90] Halpern, J. Y. and Moses, Y. O., Knowledge and Common Knowledge in a Distributed Environment, JACM, **37**(3), 1990, pp. 549–587.

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

[Hin69] Hintikka, J., *Models for Modalities*, D. Reidel, 1969.

[Kap71] Kaplan, D., Quantifying In, in [Lin71], pp. 112–144.

[15] In fact, we believe that KL^- is not reducible, but we have not been able to find a proof one way or the other.

[Kon88] Konolige, K., On the Relation between Default Logic and Autoepistemic Theories, *Artificial Intelligence* **35**(3), 1987, pp. 343-382.

[Kon89] Konolige, K., On the Relation between Autoepistemic Logic and Circumscription (Preliminary Report), in *Proc. of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989.

[Kon92] Konolige, K., Quantification in Autoepistemic Logic, to appear in: *Fundamenta Informaticae.*

[Kri63] Kripke, S. A., Semantical considerations on modal logic, *Acta Philosophica Fennica* **16**, pp. 83–94, 1963.

[Kri80] Kripke, S. A., *Naming and Necessity*, Harvard University Press, Cambridge, MA, 1980.

[LL88] Lakemeyer, G. and Levesque, H. J., A Tractable Knowledge Representation Service with Full Introspection, in *Proc. of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA, 1988, pp. 145–159.

[Lak91] Lakemeyer, G., Decidable Reasoning in First-Order Knowledge Bases with Perfect Introspection, to appear in *Proc. of the Twelth International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991.

[Lak92a] Lakemeyer, G., On Perfect Introspection with Quantifying-in, in preparation.

[Lak92b] Lakemeyer, G., *Models of Belief for Decidable Reasoning in Incomplete Knowledge Bases*, Technical Report, Department of Computer Science, University of Toronto, Toronto, Ontario, 1992.

[Lev81] Levesque, H. J., *A Formal Treatment of Incomplete Knowledge Bases*, Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Ontario, 1981.

[Lev84] Levesque, H. J., Foundations of a Functional Approach to Knowledge Representation, *Artificial Intelligence*, **23**, 1984, pp. 155-212.

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

[Lin71] Linsky, L. (ed.), *Reference and Modality*, Oxford University Press, Oxford, 1971.

[MT89] Marek, W. and Truszczyński, M., Relating Autoepistemic and Default Logics. in *Proc. of the First International Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann, San Mateo, CA, 1989, pp. 276–288.

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

[Pat87] Patel-Schneider, P. F., *Decidable, Logic-Based Knowledge Representation*, Ph.D thesis, University of Toronto, 1987.

[vDa86] van Dalen, D., Intuitionistic Logic, in D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, Vol. III, D. Reidel, 1986, pp. 225–339.