Levesque's Axiomatization of Only Knowing is Incomplete*

Joseph Y. Halpern†
IBM Research Division
Almaden Research Center, Dept. K53/802
650 Harry Road
San Jose, CA 95120-6099
halpern@almaden.ibm.com

Gerhard Lakemeyer
Institute of Computer Science
University of Bonn
Römerstr. 164
D-53117 Bonn, Germany
gerhard@cs.uni-bonn.de

Abstract

We show that the axiomatization given by Levesque for his logic of "only knowing" (Levesque 1990), which he showed to be sound and complete for the unquantified version of the logic and conjectured to be complete for the full logic, is in fact incomplete.

1 Introduction

Levesque (1990) introduced a first-order modal logic $\mathcal{OL}$ with a modal operator for "only knowing", which was taken to be the conjunction of "knowing at least" and "knowing at most".[1] He provided a collection of axioms for this logic which he showed were sound and complete in the unquantified case. He conjectured that the axiomatization was complete for the full logic. As we show here, it is not.

In the next section of this note we review the syntax and semantics of $\mathcal{OL}$, and Levesque's axiomatization of it. In Section 3, we show that Levesque's axiomatization is incomplete. We conclude in Section 4 with some further discussion of the problem of axiomatizing $\mathcal{OL}$.

2 A review of $\mathcal{OL}$

We briefly review enough of $\mathcal{OL}$ here to make this paper self-contained. The reader is encouraged to consult (Levesque 1990) for further details and intuition.

*This paper is essentially identical to one that appears in *Artificial Intelligence* **74**:2, 1995, pp. 381–387.

†Research sponsored in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.

[1] Although we have used the word "knowledge" here, we actually allow the agent's knowledge to be false, so that "belief" may be more appropriate. Since the distinction between knowledge and belief is irrelevant in this paper, following Levesque we use the words "knowledge" and "belief" interchangeably in this paper to denote belief.

The non-modal part of $\mathcal{OL}$ consists of a standard first-order logic with $=$ and a countably infinite set of *standard names*, which are treated syntactically like constants, but have a special semantics (see below). There are neither regular constants nor function symbols. This base language is extended by two modal operators, B and N, where $B\alpha$ can be read as "the agent believes (at least) α" and $N\alpha$ can be read as "the agent believes at most that α is false." $O\alpha$ is taken to be an abbreviation of $L\alpha \wedge N\neg\alpha$. An *atomic formula* is a predicate other than $=$ applied to standard names. The formula $\alpha[x/n]$ denotes α with all occurrences of the free variable x replaced by n. A formula is said to be *basic* if it does not involve the N (or O) operator, *objective* if it does not involve any modal operators, and *subjective* if all predicate symbols occur within the scope of a modal operator.

The semantics is based on the notion of possible worlds, where a world is an interpretation of the predicate symbols over the domain consisting of the standard names. Thus, the standard names are *rigid designators*, denoting the same element of the domain, namely themselves, in every world. A world w can be identified with the set of atomic formulas that are true at w (using the standard semantics of first-order logic). We call the set of all worlds W_0. Belief (B) is modeled in a standard possible-world fashion in terms of a set W of worlds. The beliefs of the agent are those sentences that are true in all worlds of W. It is well known that this simple model of belief yields the modal logic *K45*, that is, beliefs are closed under logical consequence and positive as well as negative introspection. As we said, given a set W of worlds, B denotes truth in all the worlds of W. N, on the other hand, denotes truth in all worlds *not* in W, that is, all the worlds in $W_0 - W$. By itself, N is just another ordinary belief operator like B. However, as we shall see later, the interaction between B and N turns out to be surprisingly subtle.

Given a pair W, w, which we call a *situation*, an arbitrary sentence of $\mathcal{OL}$ is interpreted according to the following recursive rules.

$W, w \models p$ if $p \in w$, where p is an atomic formula
$W, w \models (n = m)$ if n and m are identical standard names
$W, w \models \neg\alpha$ if $W, w \not\models \alpha$
$W, w \models \alpha \vee \beta$ if $W, w \models \alpha$ or $W, w \models \beta$
$W, w \models \exists x\alpha$ if $W, w \models \alpha[x/n]$ for some standard name n
$W, w \models B\alpha$ if for all $w' \in W, W, w' \models \alpha$
$W, w \models N\alpha$ if for all $w' \notin W, W, w' \models \alpha$

Since the semantics of a subjective sentence σ for a given situation W, w does not depend on w, we often write $W \models \sigma$ instead of $W, w \models \sigma$ in this case. Analogously, we write $w \models \alpha$ instead of $W, w \models \alpha$ for objective α.

Actually, the semantics we have just described is not quite Levesque's semantics. Rather than allowing W to be an arbitrary set of worlds, Levesque requires that W be *maximal* in a sense we now describe.

Two sets of worlds are said to be equivalent if they represent the same set of basic beliefs. More precisely, we say that sets W and W' are *equivalent* if for every basic formula α, we have $W \models B\alpha$ iff $W' \models B\alpha$. Levesque shows that there is a unique way to extend each set of worlds to a largest set which is equivalent to it. These largest sets of worlds are called *maximal* sets. For technical reasons, Levesque uses only maximal sets in his semantics for $\mathcal{OL}$. Thus, Levesque defines a formula α to be *valid* if $W, w \models \alpha$ for all situations such that W is maximal.

We also use Levesque's version of validity, but notice that his definitions make perfect sense even if we do not restrict to maximal sets. We define a formula α to be *strongly valid* if $W, w \models \alpha$ for all situations W, w (including ones where W is non-maximal). Clearly a formula that is strongly valid must be valid. It follows immediately from the definition of maximality that validity and strong validity coincide if we restrict to basic formulas. On the other hand, there may be non-basic formulas that are valid but not strongly valid.

We next review Levesque's axiomatization.

The axiom system AX

Let L stand for both B and N.

Axioms:

A1: All instances of theorems of first-order logic.
A2: $L\alpha$, where α is an instance of a theorem of first-order logic.
A3: $(n_i = n_i) \wedge (n_i \neq n_j)$, where n_i and n_j are distinct standard names.
A4: $L(\alpha \Rightarrow \beta) \Rightarrow (L\alpha \Rightarrow L\beta)$.
A5: $\forall x L\alpha \Rightarrow L\forall x\alpha$.
A6: $\sigma \Rightarrow L\sigma$, if σ is a subjective sentence.
A7: $N\alpha \Rightarrow \neg B\alpha$, if α is a falsifiable objective sentence.

Inference rules:

MP: From α and $\alpha \Rightarrow \beta$ 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 in α and one not in α.

Levesque showed that AX is sound with respect to his notion of validity, where only maximal sets of worlds are considered, and complete with respect to unquantified sentences, so that any valid sentence without quantifiers is provable from these axioms. It is easy to see that AX is also sound with respect to strong validity, where we allow arbitrary sets of worlds.

3 Incompleteness of the axiom system

In this section we prove that Levesque's axiom system is incomplete with respect to the full language. In fact, we show that there is a formula that is strongly valid that is not provable in his system.

Consider the two sets W_1 and W_2 defined in the proof of Theorem 3.6 of (Levesque 1990). W_1 consists of all worlds in which at least the odd-numbered standard names satisfy P, and let $W_2 = W_1 - \{w_0\}$, where w_0 is the world where the standard names that satisfy P are precisely the odd-numbered standard names. It is easy to check that the only standard names believed to satisfy P in both W_1 and W_2 are the odd-numbered names, that is,

$$W_j \models B(P(\mathbf{n}_i)) \text{ iff } i \text{ is odd, for } j = 1, 2.$$

Levesque shows

Lemma 3.1: (Levesque 1990, Lemma 3.6.2) *For any objective formula α, we have $W_1 \models B\alpha$ iff $W_2 \models B\alpha$.*

We next define a slightly nonstandard notion of satisfaction $\models_{NS}$. Actually, $\models_{NS}$ agrees with $\models$ except on situations of the form W_2, w. Formally, all clauses in the definition of $\models_{NS}$ are identical to the corresponding clause in the definition of $\models$, except for formulas of the form $N\alpha$ if we are considering the set W_2. In this case, we define:

$$W_2, w \models_{NS} N\alpha \text{ iff } W_2, w' \models_{NS} \alpha \text{ for all } w' \notin W_1.$$

Notice that for $\models$, the corresponding definition would have as its last clause "for all $w' \notin W_2$". In particular this means we do not consider the world w_0 when evaluating the truth of $N\alpha$ in W_2 according to $\models_{NS}$.

Lemma 3.2: *For all objective formulas α, we have $W_1 \models_{NS} N\alpha$ iff $W_2 \models_{NS} N\alpha$.*

Proof: This is immediate from the definitions, since in both cases, to evaluate the truth of $N\alpha$, we consider the worlds not in W_1. ∎

Lemma 3.3: *Everything provable from AX is strongly valid with respect to $\models_{NS}$.*

Proof: We must check that all the axioms of AX are strongly valid with respect to $\models_{NS}$ and that all the rules of inference preserve strong validity with respect to $\models_{NS}$. The result then follows by a straightforward induction on the length of the proof. All the cases are completely straightforward except possibly the axiom A7. Since $\models$ and $\models_{NS}$ agree on all sets of worlds except possibly W_2, we must only check what that this axiom holds in W_2.

Suppose that for some falsifiable objective formula α, we have $W_2 \models_{NS} N\alpha \wedge B\alpha$. By Lemma 3.2, we have that $W_1 \models_{NS} N\alpha$. Since $\models_{NS}$ and $\models$ agree with respect to W_1, we must have $W_1 \models N\alpha$. Since $\models_{NS}$ and $\models$ agree with respect to formulas of the form $B\alpha$ where α is objective, we must also have $W_2 \models B\alpha$. By Lemma 3.1, we have $W_1 \models B\alpha$. Thus, $W_1 \models B\alpha \wedge N\alpha$. But this contradicts the strong validity of axiom A7 with respect to $\models$. ∎

Our goal is now to construct a formula that is strongly valid with respect to $\models$ but not with respect to $\models_{NS}$. By Lemma 3.3, such a formula cannot be provable from AX, thus showing that AX is incomplete.

Let ψ_1 be the sentence $\exists x(P(x) \wedge \neg BP(x))$ and let ψ_2 be the sentence $\exists x(\neg P(x) \wedge BP(x))$.[2] Thus, ψ_1 is true if there is a standard name satisfying P that is not one of the standard names believed to satisfy P; ψ_2 is true if there is a standard name satisfying $\neg P$ which is one of the standard names believed to satisfy P. Let $\psi = \psi_1 \vee \psi_2$. Notice that ψ is true at every world with respect to W_1 or W_2 except w_0, the world where the standard names that satisfy P are precisely those believed to satisfy P. Thus,

$$W_i, w \models \psi \text{ for } i = 1, 2, \text{ unless } w = w_0. \tag{1}$$

Since ψ does not mention N, it is easy to see that (1) also holds if we replace $\models$ by $\models_{NS}$.

[2] These sentences were used (for a different purpose) in (Levesque 1984).

Lemma 3.4: *$N\psi \Rightarrow \neg B\psi$ is strongly valid (with respect to $\models$).*

Proof: Suppose $W \models N\psi$. Let $A = \{n : n \text{ is a standard name and } W \models BP(n)\}$ and let w be a world such that $w \models P(n)$ iff $n \in A$. It is easy to see that $W, w \models \neg\psi$. Since $W \models N\psi$, it must be the case that $w \in W$. Thus, $W \models \neg B\psi$. This proves the strong validity of $N\psi \Rightarrow \neg B\psi$. (Notice that this does not follow from axiom A7, since ψ is not an objective formula, although the proof of its validity follows along the same general lines as the corresponding proof for objective formulas.) ∎

The following lemma shows that, although it is valid with respect to $\models$, $N\psi \Rightarrow \neg B\psi$ is not valid with respect to $\models_{NS}$:

Lemma 3.5: *$W_2 \models_{NS} N\psi \wedge B\psi$.*

Proof: This follows from observation (1) above, which says that ψ is satisfied (with respect to $\models$ or $\models_{NS}$) by every world except w_0. However, we do not consider w_0 for either B (since $w_0 \notin W_2$) or N (because of our nonstandard semantics). ∎

Thus, there is a formula that is strongly valid (with respect to $\models$) that is not provable, namely $N\psi \Rightarrow B\psi$. We conclude, as desired, that:

Theorem 3.6: *AX is not a complete axiomatization for $\mathcal{OL}$.*

4 Discussion

Having shown that Levesque's axiomatization is incomplete, the question remains what a complete axiomatization would look like.

Typically, we expect an axiomatization to be recursive. As Levesque already noted, his axiom system is not recursive. In particular, A7 is not recursive, since it involves checking whether a formula is first-order formula is falsifiable, which is known to be a co-r.e. problem (see (Rogers 1967)). This is not an artifact of Levesque's framework. It is easy to show that there cannot be a recursive complete axiomatization of $\mathcal{OL}$, since the validity problem for the language is not r.e.

Lemma 4.1: *Every complete axiomatization of $\mathcal{OL}$ is non-recursive.*

Proof: Suppose there were a recursive complete axiomatization AX' of $\mathcal{OL}$. Then the set of falsifiable objective formulas would be r.e., since we could generate them by generating all the objective formulas α such that $N\alpha \Rightarrow \neg B\alpha$ is provable from AX'. Since the set of falsifiable objective formulas is co-r.e., this is a contradiction. ∎

If we are willing give up recursiveness, then finding a non-recursive axiomatization is, in a sense, trivial: simply declare every valid sentence an axiom. Of course, for an axiomatization to be instructive, it should not have to appeal to the very notion which it tries to capture. We would hope that the axioms would be "natural", and give insight into the logic.

We do not know whether there is a "natural" proof-theoretic account of the logic (whatever that may mean), but, as the following results suggest, if there is one, it will be hard to find.

Recall that our incompleteness proof proceeds by showing that, for a particular basic formula ψ, the formula $N\psi \Rightarrow \neg B\psi$ is strongly valid yet not provable from the axioms. The latter formula almost looks like an instance of axiom A7. It is not, of course, since A7 would apply only if the formula ψ were objective. The obvious idea, namely to strengthen axiom schema A7 by allowing it to range over all falsifiable basic sentences, can easily be dismissed. For example, consider the subjective sentence $BP(n)$ for some predicate P and standard name n. $BP(n)$ is obviously falsifiable, yet $NBP(n) \Rightarrow \neg BBP(n)$ is not valid. In fact, $NBP(n) \equiv BBP(n)$ is easily derivable from the axioms (using A6) and is therefore valid.

But what about basic sentences that are not subjective like the sentence ψ used in the previous section? In other words, do we obtain a complete axiomatization if we replace axiom A7 by the following axiom A7$'$?

A7$'$: $N\alpha \Rightarrow \neg B\alpha$, if α is a falsifiable basic non-subjective sentence.

Since ψ is basic, non-subjective, and falsifiable, the offending sentence $N\psi \Rightarrow \neg B\psi$ would now come out trivially as a theorem. Unfortunately, A7$'$ does not solve the problem either, since restricting the axiom schema to non-subjective basic sentences is still unsound. To see this, consider the formula

$$\varphi = \forall x(P(x) \Rightarrow BP(x)),$$

which is obviously falsifiable. However,

Lemma 4.2: *$N\varphi \Rightarrow \neg B\varphi$ is not valid.*

Proof: Let W_P consist of all worlds w such that $w \models \forall x P(x)$. Clearly W_P is maximal. For suppose W and W_P are equivalent. Then, in particular, $W \models B(\forall x(P(x))$, so we must have $W \subseteq W_P$. We next show that $W_P \models B\varphi \wedge N\varphi$. It is easy to see that $W_P, w \models B(\forall x P(x)) \Rightarrow \varphi$ for all worlds w. Since $W_P \models B(\forall x P(x))$, it follows that $W_P, w \models \varphi$ for all worlds w. This means that $W_P \models B\varphi \wedge N\varphi$. Hence $N\varphi \Rightarrow \neg B\varphi$ is not valid (and, *a fortiori*, not strongly valid). ∎

Although, as we just showed, $N\varphi \Rightarrow \neg B\varphi$ is not valid, there is a sense in which it just misses being valid. As we now show, the only time it fails to be valid is when every standard name is known not to satisfy P (as was the case for the set W_P of worlds considered in Lemma 4.2).

Lemma 4.3: *$\neg B(\forall x P(x)) \Rightarrow (N\varphi \Rightarrow \neg B\varphi)$ is strongly valid.*

Proof: Let W be any set of worlds such that $W \models \neg B(\forall x P(x)) \wedge N\varphi$. Since $W \models \neg B(\forall x P(x))$, there is a standard name n^* such that $W \models \neg BP(n^*)$. Since $W \models N\varphi$, it follows that for all $w' \notin W$, we have $W, w' \models \forall x(P(x) \Rightarrow BP(x))$. In particular, this means that for all $w' \notin W$, we must have $w' \models \neg P(n^*)$. Thus, there must be some $w \in W$ such that $w \models P(n^*)$. Clearly $W, w \models \neg\varphi$, so $W \models \neg B\varphi$, as desired. ∎

These lemmas show that finding a relatively natural extension of axiom A7 that would cover the counterexample is a subtle matter. Nor is there any guarantee that such an extension would give us a complete axiomatization. For example, notice that all the sound axioms we have considered so far are not only valid, but strongly valid. It may well be that there are formulas that are valid but not strongly valid. If so, we need to find an axiom that is valid but not strongly valid. The formulas that we have considered in this paper do not have this property. We leave further exploration of these issues to future work.

References

Levesque, H. J. (1984). Foundations of a functional approach to knowledge representation. *Artificial Intelligence 23*, 155–212.

Levesque, H. J. (1990). All I know: a study in autoepistemic logic. *Artificial Intelligence 42*(3), 263–309.

Rogers, Jr., H. (1967). *Theory of Recursive Functions and Effective Computability*. New York: McGraw-Hill.