Noname manuscript No.
(will be inserted by the editor)

Löb-safe Logic for Human-like Agents

Seth Ahrenbach · Second Author

Received: date / Accepted: date

Abstract Löb's Obstacle is a problem facing agents who can reason with certain powers of self-reference. When epistemic logics model the reasoning of human-like agents, they model agents that have such powers. This paper shows Löb's Obstacle in various formal systems of epistemic logic, and presents an epistemic logic that avoids it by relaxing assumptions about agents' knowledge and belief.

Keywords Löb's Theorem · Epistemic Logic · Agent Foundations

1 Introduction

Epistemic logic emerged from attempts to formalize philosophers' theories of knowledge. It has since found usefulness in areas of computer science ranging from database theory to artificial intelligence and information security. In economics, it serves as a foundation for the ideally rational agents of game theory. These applications tend to treat knowledge as a metaphorical concept relating information to things that can be modeled as agents, even if they are not very complex. While recent advances concerning the combination of dynamic logics with epistemic notions demonstrate the liveliness of the field, the static foundation of these dynamic logics has not advanced beyond the elegant S5 modal operator in quite some time.

This paper argues that philosophical logicians must return their attention to the static base. Lying at the core of epistemic logic is an obstacle emerging

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

from the depths of mathematical logic that threatens to undermine the enterprise. This obstacle blocks the way forward of any epistemic logic intended to model reasonably sophisticated agents, and certainly those for modeling humans and advanced artificial intelligence. It has many names, but researchers in the field of Agent Foundations call it Löb's Obstacle. This paper describes how Löb's Obstacle prevents epistemic logics of various sorts from modeling human-like agents, and then describes a logic that is Löb Safe which combines weakened knowledge and belief operators into a multi-modal system.

Section 2 reminds readers of the basics of modal logic and its epistemic interpretation. In section 3 we present the modal version of Löb's Theorem and describe its importance for mathematical logic. Section 4 identifies a number of logics from the epistemic modal family and shows how each one collides with Löb's Obstacle, rendering them inapplicable to reasonably complex agents. Section **??** presents an epistemic logic that avoids Löb's Obstacle by weakening the knowledge and belief operators. We mount a defense of this logic on philosophical grounds while also identifying its flaws. We hope to spark a conversation that will improve the state of epistemic logic's foundation so that it can model cognitive attitudes of humans and advanced machine intelligence. Section 8 concludes with suggested criteria for assessing potential formal epistemic solutions.

2 Modal Logic

Modal logic extends propositional logic with simple syntax, but the increase in expressive power is quite dramatic. Initially explored by Aristotle, picked up again by Medieval philosophers, and finally revived in the early 20th century, modal logic's intended inperpretation is as a logic of necessary and possible truths. But this is not the only interpretation of the symbols. By fiddling with the semantics of the modal operators, one can create a modal logic for reasoning in a wide variety of aspects. In the late 1950's and 60's, renewed interest in modal logic for reasoning about temporal modalities led to alternative interpretations, like logic of oblication (deontic logic), logic of knowledge and belief (epistemic and doxastic logics), and provability logic. Each of these logics is a modal logic with different interpretations of the modal operator, and different constraints on the underlying possibility relations.

This paper explores the way modal logics for belief, knowledge, and provability run into a fundamental mathematical obstacle of self-referential reasoning, or even just reasoning with fixed points, which inevitably leads to corrupted inferences. Ironically, the modal formula capturing the obstacle was first identified in deontic logic, illustrating the cross-cutting nature of modal logic.

3 Löb's Theorem

Löb's Obstacle takes its name from Martin Hugo Löb, who answered one of L. Henkin's follow up questions to Gödel's striking incompleteness theorems. The question was, "what about formulas that assert their own *provability*, as opposed to unprovability?"[1] Löb's response was as follows. A sufficiently powerful system, *e.g.* Peano arithmetic, can prove that formulas asserting that their own provability implies that they are true only when Peano arithmetic actually proves the formula. It seems a little trivial. For formal systems capturing the reasoning of agents, it can lead to surprising results.

Löb's Theorem in provability logic is,

$$\Box(\Box\varphi \Rightarrow \varphi) \Rightarrow \Box\varphi. \tag{1}$$

The $\Box$ is interpreted as "provability" in some formal system, particularly one at least as powerful as Peano arithmetic. If a modal operator involves the reasoning abilities of a human-like agent, then *a fortiori* it is a formal system at least as powerful as Peano arithmetic. This presents the following problem. If that agent, in its own reasoning system, can deduce the soundness of its own system, then it's reasoning system is unsound. This is because φ can be any formula, including $\bot$. The particular modal logic at risk, in our mind, is epistemic logic, or any doxastic logic with accurate reasoners.

Here we give a template derivation of Löb's Theorem, which we shall refer to below when describing how Löb's Obstacle corrupts various epistemic logics.

(3.1) $\Box(\Box\varphi \Rightarrow \varphi)$ ..... Assumption
(3.2) $\Box(\psi \Leftrightarrow (\Box\psi \Rightarrow \varphi))$ ..... Löb Sentence[2]
(3.3) $\Box(\Box\psi \Leftrightarrow \Box(\Box\psi \Rightarrow \varphi))$ ..... Axiom K
(3.4) $\Box(\Box\psi \Rightarrow \Box(\Box\psi \Rightarrow \varphi)))$ ..... (4.3) Simplification of $\Leftrightarrow$
(3.5) $\Box(\Box\psi \Rightarrow (\Box\Box\psi \Rightarrow \Box\varphi))$ ..... (4.4) Axiom K
(3.6) $\Box(\Box\psi \Rightarrow \Box\Box\psi)$ ..... Axiom 4
(3.7) $\Box(\Box\psi \Rightarrow \Box\varphi)$ ..... (4.5), (4.6)
(3.8) $\Box(\Box\psi \Rightarrow \varphi)$ ..... (4.7), (4.1)
(3.9) $\Box\psi$ ..... (4.3), (4.8)
(3.10) $\Box\Box\psi$ ..... (4.9), Axiom 4
(3.11) $\Box\Box\psi \Rightarrow \Box\varphi$ ..... (4.8), Axiom K
(3.12) $\Box\varphi$ ..... (4.10), (4.11)

$\mathcal{QED}$

Mathematical and Provability logicians refer to the key components of this proof as Löb Conditions. Identifying them in the proof above helps us identify which epistemic logics collide with Löb's Obstacle. Conversely, understanding how the Löb Conditions interact helps us construct epistemic logics that avoid the Obstacle.

The Conditions are:

[1] Not a direct quote.
[2] Sometimes referred as a Curry sentence after Logician Haskell Curry.

1. The Löb Sentence. A self-referential sentence, also formalizable as a modal fixed point.
2. Axiom K. The standard distribution axiom of normal modal logics.
3. Axiom 4. The axiom corresponding to a transitive frame relation.
4. The rule of necessitation. Likewise a standard feature of normal modal logics.

The Löb Sentence is sometimes not mentioned as a Condition, because Löb's Theorem is typically studied in the context of mathematical logic or provability logic, where such self-referential expressiveness is known to exist. We point out, however, that humans are capable of reasoning about self-referential sentences, as well, and any advanced artificial agent will be able to do so, as well.

Finally, we note the importance of Löb's Theorem's antecedent: $\Box(\Box\varphi \Rightarrow \varphi)$. Epistemic logics typically include the antecedent as a theorem, in which case Löb's Theorem will allow us to derive $\Box\varphi$ for all φ. This is why consistent mathematical systems at least as expressive as Peano arithmetic cannot prove their own consistency.

We identify some candidate epistemic logics and, on the assumption that they capture human-like reasoning, show how they crash into Löb's Obstacle.

4 Epistemic Logics that Crash

4.1 S5 Epistemic Logic

The most prominent epistemic logic in the literature, by far, is S5 epistemic logic. S5 epistemic logic is routinely presented as the logic of knowledge, and often serves as a static base for dynamic extensions to epistemic logic involving action and communication. Its characteristic axioms are:

$$\mathbf{K_i}\,(\varphi \Rightarrow \psi) \Rightarrow (\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\psi) \tag{2}$$

$$\mathbf{K_i}\,\varphi \Rightarrow \varphi \tag{3}$$

$$\neg\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\neg\mathbf{K_i}\,\varphi \tag{4}$$

Clearly (2) is Axiom K, and (3), troublingly, is the antecedent of Löb's Theorem, known as Axiom T. (4) is called the Negative Introspection axiom, or sometimes in philosophy circles, the Wisdom Axiom. Logicians call it Axiom 5. It is read, "If i does not know that φ, then she knows that she doesn't know it". Other than being clearly invalid for humans, this axiom and (3) allows us to derive,

$$\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\mathbf{K_i}\,\varphi$$

Proof

(4.1) $\neg\mathbf{K_i}\,\neg\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\varphi$ — Contrapositive of Axiom 5

(4.2) $\mathbf{K_i}\,\neg\mathbf{K_i}\,\neg\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\mathbf{K_i}\,\varphi$ — Rule of Necessitation on (5.1), Axiom K

(4.3) $\varphi \Rightarrow \neg\mathbf{K_i}\neg\varphi$	Axiom T, Contrapositive
(4.4) $\neg\mathbf{K_i}\neg\varphi \Rightarrow \mathbf{K_i}\neg\mathbf{K_i}\neg\varphi$	Axiom 5
(4.5) $\varphi \Rightarrow \mathbf{K_i}\neg\mathbf{K_i}\neg\varphi$	(4.3), (4.4)
(4.6) $\mathbf{K_i}\varphi \Rightarrow \mathbf{K_i}\neg\mathbf{K_i}\neg\mathbf{K_i}\varphi$	$\mathbf{K_i}\varphi/\varphi$, (4.5)
(4.7) $\mathbf{K_i}\varphi \Rightarrow \mathbf{K_i}\mathbf{K_i}\varphi$	(4.2), (4.6)

$\mathcal{QED}$

Thus, S5 satisfies Löb's three conditions, if we assume the presence of self-referential sentences possible, which we should. Therefore, with $\mathbf{K_i}$ instead of $\Box$, the proof of Löb's Theorem is possible in this brand of S5. However, to make matters worse, the antecedent of Löb's Theorem is itself an axiom of S5. Therefore, $\mathbf{K_i}\varphi$ is a theorem, for all φ.

We take this as a *reductio ad absurdum* that S5 epistemic logic cannot be a logic for reasoning about the knowledge of agents with expressive power beyond Peano arithmetic. Therefore, it cannot be a logic of knowledge for humans, or human-like agents.

4.2 Hintikka's S4 Epistemic Logic

In Hintikka's 1967 *Knowledge and Belief: A logic of the two notions*, he presented an epistemic logic for determining the validity and consistency of claims people make about knowledge and belief. He rejected out of hand the negative introspection axiom for knowledge, but chose to include positive introspection, which is formalized as $\mathbf{K_i}\varphi \Rightarrow \mathbf{K_i}\mathbf{K_i}\varphi$. Clearly then, if Hintikka's epistemic system is meant for human-like reasoners who can express sentences like, "If I know this sentence is true, then 1 + 1 = 2," then it crashes into Löb's Obstacle, with the extra bite of having the antecedent of Löb's Theorem as a theorem itself, and therefore, $\mathbf{K_i}\varphi$ is also a theorem.

5 Bi-Modal Synthesis of Knowledge and Belief

5.1 Hintikka's Bi-Modal System

5.2 Kraus and Lehman System

5.3 Voorbraak's Objective Knowledge and Rational Belief: OKRIB

6 CRAP

$\mathbf{K_i}\,(\varphi \Rightarrow \psi) \Rightarrow (\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\psi)$	Distribution of $\mathbf{K_i}$
$\mathbf{K_i}\,\varphi \Rightarrow \varphi$	Truth
$\mathbf{B_i}\,(\varphi \Rightarrow \psi) \Rightarrow (\mathbf{B_i}\,\varphi \Rightarrow \mathbf{B_i}\,\psi)$	Distribution of $\mathbf{B_i}$
$\mathbf{B_i}\,\varphi \Rightarrow \langle \mathbf{B}_i \rangle\,\varphi$	Belief Consistency
$\mathbf{K_i}\,\varphi \Rightarrow \mathbf{B_i}\,\varphi$	Knowledge implies Belief
$\mathbf{B_i}\,\varphi \Rightarrow \mathbf{B_i}\,\mathbf{K_i}\,\varphi$	Evidential Restraint
From $\vdash \varphi$ and $\vdash \varphi \Rightarrow \psi$, infer $\vdash \psi$	Modus Ponens
From $\vdash \varphi$, infer $\vdash \mathbf{K_i}\,\varphi$	Necessitation of $\mathbf{K_i}$

Table 1 Logic of Grounded-Coherent Epistemic Agents

Theorem 1 (Positive Belief Introspection) $\mathbf{B_i}\,\varphi \Rightarrow \mathbf{B_i}\,\mathbf{B_i}\,\varphi$

Proof

(6.1) $\mathbf{B_i}\,\varphi \Rightarrow \mathbf{B_i}\,\mathbf{K_i}\,\varphi$	ER Axiom
(6.2) $\mathbf{B_i}\,\mathbf{K_i}\,\varphi \Rightarrow \mathbf{B_i}\,\mathbf{B_i}\,\varphi$	KiB Axiom + Necessitation of $\mathbf{B_i}$ [3]
(6.3) $\mathbf{B_i}\,\varphi \Rightarrow \mathbf{B_i}\,\mathbf{B_i}\,\varphi$	(5.1), (5.2)

$\mathcal{QED}$

The logic defined by these axiom schemas and inference rules avoids Löb's Obstacle for $\mathbf{K_i}$, as it is no longer has the positive introspection property.

[3] This rule can be derived from Necessitation of $\mathbf{K_i}$ and KiB Axiom.

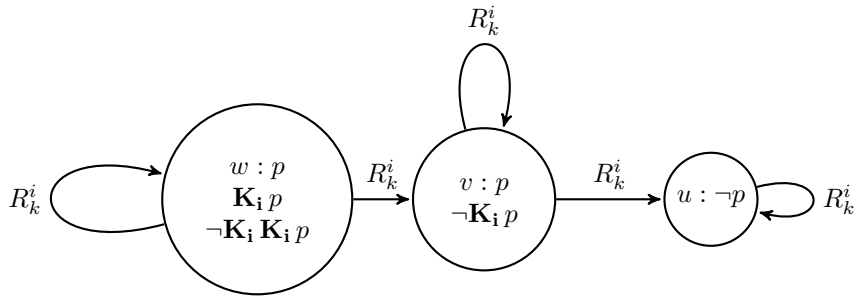

Fig. 1 A counterexample to $\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\mathbf{K_i}\,\varphi$.

Doxastic logic typically includes as an axiom of Belief Consistency, which corresponds to a serial doxastic possibility relation. Crucially, this axiom $\mathbf{B_i}\,\varphi \Rightarrow \langle \mathbf{B}_i \rangle\,]\varphi$ is equivalent to $\neg(\mathbf{B_i}\,\varphi \wedge \mathbf{B_i}\,\neg\varphi)$, not $\neg\mathbf{B_i}\,(\varphi \wedge \neg\varphi)$. The former does not run into Löb's Obstacle, while the latter ends in disaster.

Theorem 2 (Consistency Disaster) *If* $\neg\mathbf{B_i}\,(\varphi \wedge \neg\varphi)$ *and* $\mathbf{B_i}\,(\mathbf{B_i}\,\varphi \Rightarrow \varphi) \Rightarrow \mathbf{B_i}\,\varphi$ *are theorems, then* $\mathbf{B_i}\,(\varphi \wedge \neg\varphi)$.

Proof

(6.1)	$\neg\mathbf{B_i}\,(\varphi \wedge \neg\varphi)$	Belief is Consistent
(6.2)	$\mathbf{B_i}\,(\varphi \wedge \neg\varphi) \Rightarrow (\varphi \wedge \neg\varphi)$	(6.1)
(6.3)	$\mathbf{B_i}\,(\mathbf{B_i}\,(\varphi \wedge \neg\varphi) \Rightarrow (\varphi \wedge \neg\varphi))$	Necessitation of $\mathbf{B_i}$
(6.4)	$\mathbf{B_i}\,(\mathbf{B_i}\,(\varphi \wedge \neg\varphi) \Rightarrow (\varphi \wedge \neg\varphi)) \Rightarrow \mathbf{B_i}\,(\varphi \wedge \neg\varphi)$	Löb's Theorem
(6.5)	$\mathbf{B_i}\,(\varphi \wedge \neg\varphi)$	(6.3), (6.4)

But of course we have the equivalence $\mathbf{B_i}\,(\varphi \wedge \psi) \equiv (\mathbf{B_i}\,\varphi \wedge \mathbf{B_i}\,\psi)$, so from the axiom of Belief Consistency and Theorem 2, it follows that:

Theorem 3 (Beliefs Inconsistent) *For all* φ, $\mathbf{B_i}\,\varphi$.

Proof This follows from Theorem 2 and $\mathbf{B_i}\,(\varphi \wedge \psi) \equiv (\mathbf{B_i}\,\varphi \wedge \mathbf{B_i}\,\psi)$.

$\mathcal{QED}$

Therefore, the logic cannot apply to humans.

One might wonder whether it would be acceptable to abandon the Truth Axiom for knowledge and allow Löb's Theorem to hold for it. This would introduce more modesty to the notion of knowledge, where a human-like agent knows that her knowledge is true only for those propositions that she actually knows, but not in the general sense. What would this mean for epistemology? A false proposition would no longer imply a lack of knowledge.

7 Avoiding Löb

$\mathbf{K_i}\,(\varphi \Rightarrow \psi) \Rightarrow (\mathbf{K_i}\,\varphi \Rightarrow \mathbf{K_i}\,\psi)$	Distribution of $\mathbf{K_i}$
$\mathbf{K_i}\,\varphi \Rightarrow \varphi$	Truth
$\mathbf{B_i}\,(\varphi \Rightarrow \psi) \Rightarrow (\mathbf{B_i}\,\varphi \Rightarrow \mathbf{B_i}\,\psi)$	Distribution of $\mathbf{B_i}$
$\mathbf{K_i}\,\varphi \Rightarrow \mathbf{B_i}\,\varphi$	Knowledge implies Belief
$\mathbf{B_i}\,\varphi \Rightarrow \langle \mathbf{K}_i \rangle\,\mathbf{K_i}\,\varphi$	Weak Evidential Restraint
From $\vdash \varphi$ and $\vdash \varphi \Rightarrow \psi$, infer $\vdash \psi$	Modus Ponens
From $\vdash \varphi$, infer $\vdash \mathbf{K_i}\,\varphi$	Necessitation of $\mathbf{K_i}$

Table 2 Logic of Grounded-Coherent Epistemic Agents

Does this epistemic logic avoid Löb's Obstacle?

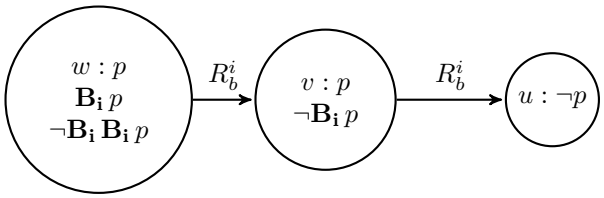

Fig. 2 A counterexample to $\mathbf{B_i}\,\varphi \Rightarrow \mathbf{B_i}\,\mathbf{B_i}\,\varphi$.

Without positive belief introspection, Löb's Theorem for belief is no longer derivable, and therefore this (very weak) epistemic logic avoids Löb's Obstacle.

8 Concluding Remarks

References

1. Author, Article title, Journal, Volume, page numbers (year)
2. Author, Book title, page numbers. Publisher, place (year)