WHAT AWARENESS ISN'T: A SENTENTIAL VIEW OF IMPLICIT AND EXPLICIT BELIEF

Kurt Konolige
Artificial Intelligence Center
SRI International
Center for the Study of Language and Information
Stanford University

ABSTRACT

In their attempt to model and reason about the beliefs of agents, artificial intelligence (AI) researchers have borrowed from two different philosophical traditions regarding the folk psychology of belief. In one tradition, belief is a relation between an agent and a proposition, that is, a *propositional attitude*. Formal analyses of propositional attitudes are often given in terms of a possible-worlds semantics. In the other tradition, belief is a relation between an agent and a sentence that expresses a proposition (the *sentential* approach). The arguments for and against these approaches are complicated, confusing, and often obscure and unintelligible (at least to this author). Nevertheless strong supporters exist for both sides, not only in the philosophical arena (where one would expect it), but also in AI.

In the latter field, some proponents of possible-worlds analysis have attempted to remedy what appears to be its biggest drawback, namely the assumption that an agent believes all the logical consequences of his or her beliefs. Drawing on initial work by Levesque, Fagin and Halpern define a *logic of general awareness* that superimposes elements of the sentential approach on a possible-worlds framework. The result, they claim, is an appropriate model for resource-limited believers.

We argue that this is a bad idea: it ends up being equivalent to a more complicated version of the sentential approach. In concluding we cannot refrain from adding to the debate about the utility of possible-worlds analyses of belief.

Introduction

Artificial Intelligence has borrowed from two different philosophical traditions for its formalizations of belief and knowledge. In one tradition, belief is a relation between an agent and a proposition, that is, belief is a *propositional attitude*. In the other tradition, belief is a relation between an agent and a sentence that expresses a proposition. We will call this the *sentential* approach.

The propositional attitude approach has an elegant formalization using in the possible-world semantics developed by Hintikka [4] and Kripke [8]. As is well known, as long as no "impossible" worlds are allowed, this semantics enforces the condition that an agent's beliefs are closed under logical consequence. It is appropriate only for agents who are perfect reasoners with infinite computational capabilities, something which is not realistic for either human or computer agents. Hintikka [5] called this undesirable property *logical omniscience*, a somewhat ambiguous term; here we will use the more precise expression *consequential closure*.

Although it is an idealization, the possible-worlds model is popular because it lends itself to logical analysis. By contrast, the formalization of sentential belief can be relatively complicated, as seen in the work of McCarthy [10], Perlis [12], Konolige [6], and others. Yet because it does not make the assumption of consequential closure, it comes closer to capturing the behavior of real agents.

In a recent paper, Fagin and Halpern [2] attempt to deal with the problem of resource-limited reasoning (among others) by combining features of the two approaches. They are motivated by an idea of Levesque's [9]: that there is a distinction between *explicit beliefs* (those beliefs an agent actually has or professes to) and *implicit beliefs* (all the logical consequences of his implicit beliefs). In their *logic of general awareness*, Fagin and Halpern represent implicit beliefs in the usual fashion with a possible-worlds semantics. To account for explicit belief in this model, they introduce a syntactic filter at each possible world, so that explicit beliefs are implicit beliefs restricted to those sentences allowed by the filter. In the language, they use an *awareness operator* to specify the filter; hence the name of the logic. Awareness is a *syntactic* concept, since its argument is interpreted as a sentence, rather than a proposition.

It seems clear that a logic of general awareness can indeed model resource-limited reasoning from beliefs. However, implicit in the paper is the notion that it is the marriage (unholy though it may be!) of sentential and propositional approaches that is responsible for this result, so that the "elegance and intuitive appeal of the semantic [i.e., possible-worlds] approach" is preserved. Here we take a critical view of this position, and argue that a logic of general awareness is essentially equivalent to the sentential approach, but the addition of possible-worlds elements adds unmotivated and unintuitive complications.

The Argument

A sentential view of explicit and implicit belief

In the sentential view, beliefs are represented as sets of statements or sentences. In its simplest form, the distinction between implicit and explicit belief can be represented by the following diagram:

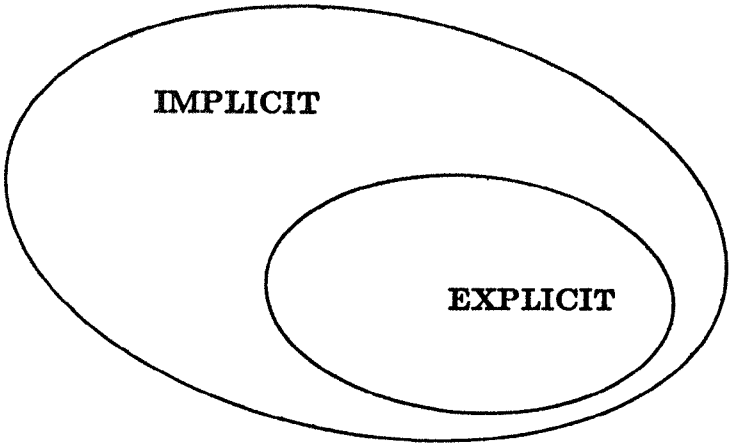

Figure 1. Explicit and implicit belief.

The set of implicit beliefs is the set of logical consequences of the explicit beliefs. If the explicit beliefs are inconsistent, the implicit beliefs are every sentence of the language.

In the sentential approach, the explicit beliefs are normally considered to be primary, that is, they represent the beliefs of the agent, and the implicit beliefs are derived from them by closure under logical consequence. Generally, in giving a formal characterization of explicit beliefs, one attempts to capture how agents syntactically derive one belief from others. One example of this type of system is the *deduction model* (Konolige [7]). This model is based on the observation that AI knowledge bases usually consist of a core set of basic beliefs and a deductive mechanism for deriving some, but usually not all, logical consequences of the core set.

Unhappy with the syntactic characterization of explicit belief, Levesque [9] attempted to characterize explicit belief using a semantics similar to possible-worlds analyses of implicit belief. This analysis has some severe problems in modeling resource-limited reasoning, however, and Fagin and Halpern [2] present an alternative of their own, which we now review.

A logic of general awareness

For our purposes, we can restrict the logic to the simple case of a single believer, Ralph.

We assume a standard propositional language with negation and conjunction as the basic operators, together with a distinguished primitive proposition $\perp$, which is always interpreted as **false**. To this we add three unary modal operators: B (*explicit belief*), L (*implicit belief*), and A (*awareness*). All intermixtures are permitted, e.g., $BAL\phi$.

A *Kripke structure for general awareness* is a tuple $M = \langle S, \pi, \mathcal{A}, \mathcal{B} \rangle$, where S is a set of states, $\pi(s,p)$ is a truth-assignment for each primitive proposition p and state $s \in S$, and B is a transitive, euclidean, and serial binary relation on S (the *accessibility relation*). For each possible world s, $\mathcal{A}(s)$ is a set of sentences of the modal language such that $\perp \in \mathcal{A}(s)$. At least initiallly, this is the *only* restriction on $\mathcal{A}(s)$. The formulas in $\mathcal{A}(s)$ are those Ralph in state s is "aware of," but does not necessarily believe. Fagin and Halpern stress that various restrictions can be placed on awareness to capture a number of different meanings for "aware of." We discuss this in the next section.

The semantics of the language is given by the truth relation $\models$, defined as follows:

$$
\begin{array}{ll}
M, s \not\models \perp & \\
M, s \models p, & \text{if } p \text{ is a primitive proposition and } \pi(s,p) = \textbf{true}. \\
M, s \models \neg\phi & \text{if } M, s \not\models \phi. \\
M, s \models \phi \wedge \psi & \text{if } M, s \models \phi \text{ and } M, s \models \psi. \\
M, s \models L\phi & \text{if } M, t \models \phi \text{ for all } t \text{ such that } t \mathcal{B} s. \\
M, s \models B\phi & \text{if } \phi \in \mathcal{A}(s) \text{ and } M, t \models \phi \text{ for all } t \text{ such that } t \mathcal{B} s. \\
M, s \models A\phi & \text{if } \phi \in \mathcal{A}(s).
\end{array}
\tag{1}
$$

In this semantics, the implicit belief operator L has a standard possible-worlds semantics. It obeys both the positive introspection axiom $L\phi \supset LL\phi$, the negative introspection axioms $\neg L\phi \supset L\neg L\phi$, and the consistency axiom $\neg L\perp$. This is by virtue of the accessibility relation being transitive, euclidean, and serial, respectively. In short, if we restrict the language to just L, we have the belief logic of weak $S5$, plus the consistency axiom (see Halpern and Moses [3]).

As Fagin and Halpern note, explicit belief is defined by the semantics as implicit belief restricted to those sentences permitted by awareness. Thus $B\phi \equiv L\phi \wedge A\phi$. If Ralph is aware of all sentences, then $B\phi \equiv L\phi$ for all ϕ, and explicit and implicit belief are identical.

What is awareness?

Suppose we are trying to represent the behavior of a knowledge base (which we again call Ralph), which responds *yes* to some queries and *no* to others. $B\phi$ then means that the Ralph responds *yes* to the query ϕ, and $\neg B\phi$ is the *no* response. What can we interpret awareness as in this situation? The suggested meaning for $A\phi$ is "Ralph is able to determine whether or not ϕ follows from his initial premises in time T."[1] That is, the sentences Ralph is "aware of" are some class for which it is easy to do deductions, or to show that no deductions exist. We can draw this picture to represent the various sets of sentences and their relationships:

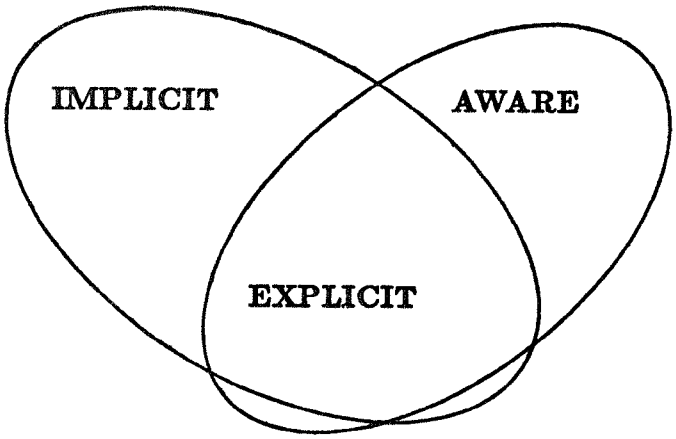

Figure 2. Awareness and belief.

Note that this is the same as our previous picture, except that the intersection of awareness and implicit belief is what defines the explicit beliefs. Logical consequence has taken over the role of deduction from an initial set of beliefs; the awareness filter restricts explicit beliefs to a subset of the consequences. However, unlike Levesque's scheme, the awareness filter is a purely syntactic notion; the set of sentences in the aware set is arbitrary (with the exception of $\perp$).

The logic of general awareness thus characterizes agents as perfect reasoners, restricted in some way to considering a subset of possible sentences to reason about. Now we ask two

[1]The actual phrase that Fagin and Halpern use is "[Ralph] is able to compute the truth of ϕ within time T." However truth is not the appropriate notion for knowledge bases, since it certainly may be the case that they have incorrect or incomplete information about the world.

questions:

1. Does the possible-world semantics play an essential role in this approach?
2. How intuitive and useful is the general notion of an awareness-limited perfect reasoner for representing limited reasoning?

The role of possible-world semantics

In his original work on knowledge and belief, Hintikka [4] used a possible-worlds semantics to systematize various principles he proposed to govern the consistency ("defensibility") of sets of sentences about knowledge and belief. These principles were derived first, from various arguments; it was only fortuitous that conditions on accessibility, such as transitivity and reflexivity, would (as Kripke showed) turn out to be equivalent. For example, consider the axiom of "negative introspection:"

$$\neg B\phi \supset B\neg B\phi \tag{2}$$

This sentence states a condition on Ralph's introspective capabilities, namely, he knows when he doesn't belief something. In Kripke models, this condition is enforced by making the accessibility relation euclidean (if $x\mathcal{B}y$ and $x\mathcal{B}z$, then $y\mathcal{B}z$). The semantical analysis makes the negative introspection rule valid, but does not give us any insight into the nature of introspection. For example, the fact that $\mathcal{B}$ is *euclidean* does not help us to predict what other introspective capabilities Ralph should have, or how negative introspection might break down or degrade in certain situations. Kripke models do not give us the principles upon which we might build and defend a theory of belief, but they are a useful formal tool for analyzing the properties of various modal logical systems.

In the case of awareness, the formal correspondence between accessibility conditions and sets of awareness sentences breaks down; hence the connection between accessibility conditions and belief is ruptured. For example, suppose Ralph believes p; we conclude that all of Bp, Ap, and Lp are true. What is required for BBp to be true? We know $BBp \equiv (LBp \wedge ABp) \equiv (LLp \wedge LAp \wedge ABp)$, and by transitivity LLp is true. We need the further conditions:

1. $A\phi \supset LA\phi$ (semantically, if $\phi \in \mathcal{A}(s)$ and $t\mathcal{B}s$, then $\phi \in \mathcal{A}(t)$), and
2. $B\phi \supset AB\phi$.[2]

[2]Fagin and Halpern suggest that a self-reflective Ralph will be aware what he is aware of, and suggest for this case the condition that if $\phi \in \mathcal{A}(s)$, then $A\phi \in \mathcal{A}(s)$. This makes the sentence BAp true, but says nothing about awareness of *beliefs*, that is, ABp.

Neither (1) nor (2) is affected by the structure of accessibility. Hence the nice formal analysis of introspective properties obtainable in Kripke semantics is not present in the logic of general awareness.

Finally, we note that the logic of general awareness has a natural sentential semantics. Recall that belief is defined as the intersection of those sentences Ralph is aware of with those sentences he implicitly believes:

$$M, s \models B\phi \quad \text{if } \phi \in \mathcal{A}(s) \text{ and } M, t \models \phi \text{ for all } t \text{ such that } t \mathcal{B} s$$

The first half of the conjunction refers explicitly to a set of sentences (those in $\mathcal{A}(s)$); the second half implicitly by the use of possible worlds. It is possible to give this second set a syntactic characterization also. As we have noted, L is axiomatized as weak $S5$ plus consistency. Moore [11] has shown that weak $S5$ characterizes *stable sets*. A set S is stable if it contains all tautologies, is closed under *modus ponens*, and obeys the following conditions:

1. If $\phi \in S$, $L\phi \in S$.
2. If $\phi \notin S$, $\neg L\phi \in S$.

So from a sentential point of view, models of general awareness consist of a stable set intersected with an arbitrary awareness set.

Resource-limited reasoning with awareness

How compelling is the picture of resource-limited reasoning presented by the logic of limited awareness (Figure 2)? What we are asking here is if there is any motivation from our intuitions about folk-psychological notions of belief. For example, the deduction model seems plausible in its general form, because agents learn facts about the world from their observations or from being told, and then go on to deduce further consequences. We might expect to characterize the beliefs of a resource-limited agent by noting what facts he learns, and what rules he uses to infer other facts. Such a model would be useful in predicting the behavior of the agent, given partial information about his beliefs.

On the other hand, the logic of general awareness represents agents as perfect reasoners, restricted to considering some syntactic class of sentences. There don't seem to be any clear intuitions that this is the case for human or computer agents. As an exercise, here are two possible psychological stories that would fit into the awareness framework:

1. Agents compute all logical consequences of their beliefs, throwing away those not in the awareness set, perhaps because of memory limitations. This is not a plausible story, because agents are also affected by time limitations.

2. Agents use a complete logical deduction system to compute consequences of beliefs, but do not pursue those lines of reasoning which require deriving sentences not in the awareness set. This is plausible, because incomplete derivations of this nature could be accomplished with limited space and time resources, given tight enough syntactic restrictions on the derivable sentences. However, this particular story is just the deduction model of belief.

Conclusion: propositional attitudes and possible worlds

It does not seem that there is much to be gained by considering a logic of general awareness, at least as far as modeling resource-limited reasoning from beliefs is concerned. It is no more powerful than current sentential logics, and can be re-expressed in these terms. The practice of mixing sentential and possible-world elements in the semantics does not preserve the elegance of the latter, or offer any essential insight into the psychological nature of explicit belief.

The lesson we should draw from this is not that propositional attitude and sentential views of belief are irreconcilable and should not be intermixed, but rather that the introduction of possible worlds bears rethinking. As Stalnaker [13] notes, in possible-world analyses of belief, possible worlds are simply a formal device for saying what a proposition *is*. If we *define* propositions as sets of possible worlds, then we can easily analyze their properties. However, ease does not mean accuracy: as we have noted, the undesirable property of consequential closure emerges from this analysis.

Yet the idea of possible worlds has a certain intuitive appeal — one speaks of *epistemic alternatives*, as if there were worlds that, for all Ralph knows, could be the actual world. The problem is that epistemic alternatives do not seem to be anything at all like possible worlds. For example, Ralph, being a good computer scientist, doesn't really know whether $P = NP$, so there should be some epistemic alternatives in which it is true, and some in which it is false. Yet because it is a question of mathematics, $P = NP$ is either true in all possible worlds, or false in all of them. Similarly, we might argue that although the laws of physics are obeyed in all possible worlds, Ralph may not have a very good idea of how physical systems behave, and so his epistemic alternatives include worlds that are clearly physically impossible.

If epistemic alternatives are not possible worlds, what are they? Perhaps they are something like the "notional worlds" of Dennett [1]. Whatever they are, it seems that Ralph's inferential capabilities will play a role in their definition. For suppose that Ralph can believe ϕ and $\phi \supset \psi$ without inferring ψ. Then there will be epistemic alternatives in which ϕ and $\phi \supset \psi$ are true, but ψ is not. On the other hand, if Ralph always infers ψ from ϕ and $\phi \supset \psi$, then all epistemic alternatives will be closed under this inference.

Well, this is obviously not a complete or very convincing story about propositional attitudes and belief. Still, fully fleshed out, it may be a reasonable alternative to the normal formalization using possible worlds.

Acknowledgements

I am grateful to David Israel and Joe Halpern for their comments on an earlier draft of this paper. This research was supported in part by Contract N00014–85–C–0251 from the Office of Naval Research, and in part by a gift from the Systems Development Foundation.

References

[1] Dennett, D. C. (1982). Beyond belief. In *Thought and Object*, A. Woodfield editor, Clarendon Press, Oxford, pp. 1–95.

B. Meltzer and D. Michie editors, Edinburgh University Press, Edinburgh, Scotland, pp. 120–147.

[2] Fagin, R. and Halpern, J. Y. (1985). Belief, awareness, and limited reasoning. In Proceedings of the Ninth International Joint Conference on AI, Los Angeles, California, pp. 491–501.

[3] Halpern, J. Y. and Moses, Y. (1985). A guide to the modal logics of knowledge and belief: preliminary draft. In Proceedings of the Ninth International Joint Conference on AI, Los Angeles, California, pp. 479–490.

[4] Hintikka, J. (1962). *Knowledge and Belief.* Cornell University Press, Ithaca, New York.

[5] Hintikka, J. (1975). Impossible possible worlds vindicated. *J. Philosophical Logic* **4**, pp. 475–484.

[6] Konolige, K. (1980). A first-order formalization of knowledge and action for a multiagent planning system. Artificial Intelligence Center Tech Note 232, SRI International, Menlo Park, California.

[7] Konolige, K. (1984). *A Deduction Model of Belief and its Logics.* Doctoral thesis, Stanford University Computer Science Department Stanford, California.

[8] Kripke, S. A. (1963). Semantical considerations on modal logics. *Acta Philsophica Fennica* **16**, pp. 83–94.

[9] Levesque, H. J. (1984). A logic of implicit and explicit belief. In Proceedings of the National Conference on Artificial Intelligence, Houston, Texas, pp. 198–202.

[10] McCarthy, J. (1979). First-order theories of individual concepts and propositions. In *Machine Intelligence 9*, B. Meltzer and D. Michie editors, Edinburgh University Press, Edinburgh, Scotland, pp. 120–147.

[11] Moore, R. C. (1983). Semantical Considerations on Nonmonotonic Logic. Artificial Intelligence Center Technical Note 284, SRI International, Menlo Park, California.

[12] Perlis, D. (1981). Language, computation, and reality. Department of Computer Science Technical Report 95, University of Rochester, Rochester, New York.

[13] Stalnaker, R. C. (1984). *Inquiry.* MIT Press, Cambridge, Massachusetts.