POSSIBLE-WORLD SEMANTICS FOR AUTOEPISTEMIC LOGIC

Technical Note 337

August 1984

By: Robert C. Moore, Staff Scientist
Artificial Intelligence Center
Computer Science and Technology Division

SRI Project 4488

To be presented at the Workshop on Nonmonotonic Reasoning, Mohonk Mountain House, New Paltz, New York, October 17-19, 1984.

The research reported herein was supported by the Air Force Office of Scientific Research under Contract No. F49620-82-K-0031. The views and conclusions expressed in this document are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed of implied, of the Air Force Office of Scientific Research or the U.S. Government.

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

ABSTRACT

In a previous paper [Moore, 1983a, 1983b], we presented a nonmonotonic logic for modeling the beliefs of ideally rational agents who reflect on their own beliefs, which we called "autoepistemic logic." We defined a simple and intuitive semantics for autoepistemic logic and proved the logic sound and complete with respect to that semantics. However, the nonconstructive character of both the logic and its semantics made it difficult to prove the existence of sets of beliefs satisfying all the constraints of autoepistemic logic. This note presents an alternative, possible-world semantics for autoepistemic logic that enables us to construct finite models for autoepistemic theories, as well as to demonstrate the existence of sound and complete autoepistemic theories based on given sets of premises.

I INTRODUCTION

In a previous paper [Moore, 1983a, 1983b], we presented a nonmonotonic logic for modeling the beliefs of ideally rational agents who reflect on their own beliefs, which we called "autoepistemic logic." We defined a simple and intuitive semantics for autoepistemic logic and proved the logic sound and complete with respect to that semantics. However, the nonconstructive character of both the logic and its semantics made it difficult to prove the existence of sets of beliefs satisfying all the constraints of autoepistemic logic. This note presents an alternative, possible-world semantics for autoepistemic logic that enables us to construct finite models for autoepistemic theories, as well as to demonstrate the existence of sound and complete autoepistemic theories based on given sets of premises.

Autoepistemic logic is nonmonotonic, because we can make statements in the logic that allow an agent to draw conclusions about the world from his own lack of information. For example, we can express the belief that "If I do not believe P, then Q is true." If an agent adopts this belief as a premise and he has no means of inferring P, he will be able to derive Q. On the other hand, if we add P to his premises, Q will no longer be derivable. Hence, the logic is nonmonotonic.

Autoepistemic logic is closely related to the nonmonotonic logics of McDermott and Doyle [1980; McDermott, 1982]. In fact, it was designed to be a reconstruction of these logics that avoids some of their peculiarities. This is discussed in detail in our earlier paper [Moore, 1983a, 1983b]. This work is also closely related to that of Halpern and Moses [1984], the chief difference being that theirs is a logic of knowledge rather than belief. Finally, Levesque [1981] has also developed a kind of autoepistemic logic, but in his system the agent's premises are restricted to a sublanguage that makes no reference to what he believes.

II SUMMARY OF AUTOEPISTEMIC LOGIC

The language of autoepistemic logic is that of ordinary propositional logic, augmented by a modal operator L. We want formulas of the form LP to receive the intuitive interpretation "P is believed" or "I believe P." For example, $P \supset LP$ could be interpreted as saying "If P is true, then I believe that P is true."

The type of object that is of primary interest in autoepistemic logic is a set of formulas that can be interpreted as a specification of the beliefs of an agent reflecting upon his own beliefs. We will call such a set of formulas an _autoepistemic theory_. The truth of an agent's beliefs, expressed as an autoepistemic theory, is determined by (1) which propositional constants are true in the external world and (2) which formulas are believed by the agent. A formula of the form LP will be true with respect to an agent if and only if P is in his set of beliefs. To formalize this, we define the notions of autoepistemic interpretation and autoepistemic model. An _autoepistemic interpretation_ I of an autoepistemic theory T is a truth assignment to the formulas of the language of T that satisfies the following conditions:

1. I conforms to the usual truth recursion for propositional logic.
2. A formula LP is true in I if and only if $P \in T$.

An *autoepistemic model* of T is an autoepistemic interpretation of T in which all the formulas of T are true. (Any truth assignment satisfying Condition 1 in which all the formulas of T are true will be called simply a *model* of T.)

We can readily define notions of soundness and completeness relative to this semantics. Soundness of a theory must be defined with respect to some set of premises. Intuitively speaking, an autoepistemic theory T, viewed as a set of beliefs, will be sound with respect to a set of premises A, just in case every formula in T must be true, given that all the formulas in A are true and given that T is, in fact, the set of beliefs under consideration. This is expressed formally by the following definition:

> An autoepistemic theory T is *sound* with respect to a set of premises A if and only if every autoepistemic interpretation of T that is a model of A is also a model of T.

The definition of completeness is equally simple. A semantically complete set of beliefs will be one that contains everything that must be true, given that the entire set of beliefs is true and given that it is the set of beliefs being reasoned about. Stated formally, this becomes

> An autoepistemic theory T is *semantically complete* if and only if T contains every formula that is true in every autoepistemic model of T.

Finally, we can give syntactic characterizations of the autoepistemic theories that conform to these definitions of soundness and completeness [Moore, 1983b, Theorems 3 and 4]. We say that an autoepistemic theory T is _stable_ if and only if (1) it is closed under ordinary tautological consequence, (2) $LP \in T$ whenever $P \in T$, and (3) $\neg LP \in T$ whenever $P \notin T$.

> Theorem: An autoepistemic theory T is semantically complete if and only if T is stable.

We say that an autoepistemic theory T is _grounded_ in a set of premises A if and only if every formula in T is a tautological consequence of $A \cup \{LP \mid P \in T\} \cup \{\neg LP \mid P \notin T\}$.

> Theorem: An autoepistemic theory T is sound with respect to a set of premises A if and only if T is grounded in A.

With these soundness and completeness theorems, we can see that the possible sets of beliefs an ideally rational agent might hold, given A as his premises, would be stable autoepistemic theories that contain A and are grounded in A. We call these theories _stable expansions_ of A.

III AN ALTERNATIVE SEMANTICS FOR AUTOEPISTEMIC LOGIC

The semantics we have provided for autoepistemic logic is simple, intuitive, and allows us to prove a number of important general results, but it requires enumerating an infinite truth assignment if the theory under consideration contains infinitely many formulas. This makes it difficult to exhibit particular models and interpretations we may be interested in. The problem is that, in the general case, there need be no systematic connection between the truth of one formula of the form LP and any other. Autoepistemic logic is designed to characterize the beliefs of ideally rational agents, but we want the semantics to be broader than that. The semantics we have defined is intended to apply to arbitrary sets of beliefs, with the beliefs of ideally rational agents being a special case (just as model theory for standard logic applies to arbitrary sets of formulas, not just to those that are closed under logical consequence). Thus, our semantics makes no necessary connection between the truth of $L(P \land Q)$ and LP or LQ, because it is at least conceivable that an agent might be so logically deficient as to believe $P \land Q$ without believing P or believing Q. In such a case, there is little we can expect the truth definition for an autoepistemic theory to do, other than to list the true formulas of the form LP by brute stipulation.

If we confine our attention to ideally rational agents, however, much more structure emerges. In fact, we can show that stable autoepistemic theories can be simply characterized by Kripke-style possible-world models for modal logic [Kripke, 1971]. For our purposes, what we need to recall about a Kripke structure is that it contains a set of possible worlds and an accessibility relation between pairs of worlds. The truth of a formula is defined relative to a world, and conforms to the usual truth recursion for propositional logic. A formula of the form LP is true in a world W just in case P is true in every world accessible from W. Kripke structures in which the accessibility relation is an equivalence relation are called S5 structures, and the S5 structures that will be of interest to us are those in which every world is accessible from every world. We will call these the *complete* S5 structures. Our major result is that the sets of formulas that are true in every world of some complete S5 structure are exactly the stable autoepistemic theories. (This result has been obtained independently by Halpern and Moses [1984] and by Melvin Fitting [personal communication]).

> Theorem: T is the set of formulas that are true in every world of some complete S5 structure if and only if T is a stable autoepistemic theory.

Proof: Suppose T is the set of formulas true in every world of a complete S5 structure. By the soundness of propositional logic, T is closed under tautological consequence. By the truth rule for L, LP is true in every world just in case P is true in every world; therefore

$LP \in T$ if and only if $P \in T$. Furthermore, by the truth rule for L, LP is false in every world just in case P is false in some world; so $\neg LP \in T$ if and only if $P \notin T$. Therefore T is stable. In the opposite direction, suppose that T is stable. Let T' be the set of formulas of T that contain no occurrences of L. We will call these the _objective_ formulas of T. Since T is closed under tautological consequence, T' will also be closed under tautological consequence. Consider the set of all models of T' and the complete S5 structure in which each of these models defines a possible world. T' will contain exactly the objective formulas true in every world in this model; hence, T' will contain precisely the objective formulas of the stable autoepistemic theory T'' defined by this S5 structure. But by a previous result [Moore 1983b, Theorem 2], stable theories containing the same objective formulas are identical, so T must be the same as T''. Hence, T is the set of formulas true in every world of a complete S5 structure.

Given this result, we can characterize any autoepistemic interpretation of any stable theory by an ordered pair consisting of a complete S5 structure (to specify the agent's beliefs) and a propositional truth assignment (to specify what is actually true in the world). Such a structure (K, V) defines an autoepistemic interpretation of the theory T consisting of all the formulas that are true in every world in K. A formula of T is true in (K, V) if it is true according to the standard truth recursion for propositional logic, where the propositional constants are true in (K, V) if and only if they are true in V, and the formulas of the form LP are true in (K, V) if and only if they are true in every world in K (using the truth rules for Kripke structures). We will say that (K, V) is a _possible-world interpretation_ of T and, if every formula of T is true in (K, V), we will say that (K, V) is also a _possible-world model_ of T. In view of the preceding theorem, it should be obvious that for every autoepistemic interpretation or autoepistemic model of a stable theory there is a corresponding possible-world interpretation or possible-world model, and vice versa.

> Theorem: If (K, V) is a possible-world interpretation of T, then (K, V) will be a possible-world model of T if and only if the truth assignment V is consistent with the truth assignment provided by one of the possible worlds in K (i.e., if the actual world is one of the worlds that are compatible with what the agent believes).

Proof: If V is compatible with one of the worlds in K, then any propositional constant that is true in all worlds in K will be true in V. Therefore, any formula that comes out true in all worlds in K will also come out true in (K, V), and (K, V) will be a possible-world model of T. In the opposite direction, suppose that V is not compatible with any of the worlds in K. Then, for each world W in K, there will be some propositional constant that W and V disagree on. Take that constant or its negation, whichever is true in W, plus the corresponding formulas for all other worlds in K, and form their disjunction. (This will be a finite disjunction, provided there are only finitely many propositional constants in the language.) This disjunction will be true in every world in K, so it will be a formula of T, but it will be false in V. Therefore, (K, V) will not be a possible-world model of T.

IV APPLICATIONS OF POSSIBLE-WORLD SEMANTICS

One of the problems with our original presentation of autoepistemic logic was that, since both the logic and its semantics were defined nonconstructively, we were unable to easily prove the existence of stable expansions of nontrivial sets of premises. With the finite models provided by the possible-world semantics for autoepistemic logic, this becomes quite straightforward. For instance, we claimed [Moore, 1983a, 1983b] that the set of premises $\{\neg LP \supset Q, \neg LQ \supset P\}$ has two stable expansions--one containing P but not Q, and the other containing Q but not P--but we were unable to do more than give a plausibility argument for that assertion. We can now demonstrate this fact quite rigorously.

Consider the stable theory T, generated by the complete S5 structure that contains exactly two worlds, $\{P, Q\}$ and $\{P, \neg Q\}$. (We will represent a possible world by the set of propositional constants and negations of propositional constants that are true in it.) The possible-world interpretations of T will be the ordered pairs consisting of this S5 structure and any propositional truth assignment. Consider all the possible-world interpretations of T in which $\neg LP \supset Q$ and $\neg LQ \supset P$ are both true. By exaustive enumeration, it is easy to see that these are exactly

$(\{\{P, Q\}, \{P, \neg Q\}\}, \{P, Q\})$
$(\{\{P, Q\}, \{P, \neg Q\}\}, \{P, \neg Q\})$

Since, in each case, the actual world is one of the worlds that are compatible with everything the agent believes, each of these is a possible-world model of T. Therefore, T is sound with respect to $\{\neg LP \supset Q, \neg LQ \supset P\}$. Since T is stable and includes $\{\neg LP \supset Q, \neg LQ \supset P\}$ (note that both these formulas are true in all worlds in the S5 structure), T is a stable expansion of A. Moreover, it is easy to see that T contains P but not Q. A similar construction yields a stable expansion of T that contains Q but not P.

On the other hand, if both P and Q are to be in a theory T, the corresponding S5 structure contains only one world, $\{P, Q\}$. But then $\{\{\{P, Q\}\}, \{\neg P, \neg Q\}\}$ is a possible-world interpretation of T in which $\neg LP \supset Q$ and $\neg LQ \supset P$ are both true, but some of the formulas of T are not (P and Q, for instance). Hence, if T contains both P and Q, T is not a stable expansion of $\{\neg LP \supset Q, \neg LQ \supset P\}$.

REFERENCES

Kripke, S. A. [1971] "Semantical Considerations on Modal Logic," in *Reference and Modality*, L. Linsky, ed., pp. 63-72 (Oxford University Press, London, England).

Halpern, J. Y. and Y. Moses [1984] "Towards a Theory of Knowledge and Ignorance," Workshop on Nonmonotonic Reasoning, Mohonk Mountain House, New Paltz, New York (October 17-19, 1984).

Levesque, H. J. [1981] "The Interaction with Incomplete Knowledge Bases: A Formal Treatment," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, University of British Columbia, Vancouver, B.C., Canada, pp. 240-245 (August 24-28, 1981).

McDermott, D. and J. Doyle [1980] "Non-Monotonic Logic I," *Artificial Intelligence*, Vol. 13, Nos. 1, 2, pp. 41-72 (April 1980).

McDermott, D. [1982] "Nonmonotonic Logic II: Nonmonotonic Modal Theories," *Journal of the Association for Computing Machinery*, Vol. 29, No. 1, pp. 33-57 (January 1982).

Moore R. C. [1983a] "Semantical Considerations on Nonmonotonic Logic," *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, pp. 272-279 (August 8-12, 1983).

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