SRI International

SEMANTICAL CONSIDERATIONS ON NONMONOTONIC LOGIC

Technical Note 284

June 1983

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

SRI Project 4488

This is a revised and expanded version of a paper to appear in Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, August 8-12, 1983.

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.

SRI International®

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

ABSTRACT

Commonsense reasoning is "nonmonotonic" in the sense that we often draw, on the basis of partial information, conclusions that we later retract when we are given more complete information. Some of the most interesting products of recent attempts to formalize nonmonotonic reasoning are the nonmonotonic logics of McDermott and Doyle [McDermott and Doyle, 1980; McDermott, 1982]. These logics, however, all have peculiarities that suggest they do not quite succeed in capturing the intuitions that prompted their development. In this paper we reconstruct nonmonotonic logic as a model of an ideally rational agent's reasoning about his own beliefs. For the resulting system, called autoepistemic logic, we define an intuitively based semantics for which we can show autoepistemic logic to be both sound and complete. We then compare autoepistemic logic with the approach of McDermott and Doyle, showing how it avoids the peculiarities of their nonmonotonic logic.

I INTRODUCTION

It has been generally acknowledged in recent years that one important feature of ordinary commonsense reasoning that standard logics fail to capture is its nonmonotonicity. An example frequently given to illustrate the point is the following. If we know that Tweety is a bird, we will normally assume, in the absence of evidence to the contrary, that Tweety can fly. If, however, we later learn that Tweety is a penguin, we will withdraw our prior assumption. If we try to model this in a formal system, we seem to have a situation in which a theorem P is derivable from a set of axioms S, but is not derivable from some set S′ that is a superset of S. The set of theorems, therefore, does not increase monotonically with the set of axioms; hence this sort of reasoning is said to be "nonmonotonic." As Minsky [1974] has pointed out, standard logics are always monotonic, because their inference rules make every axiom permissive. That is, the inference rules are always of the form "P is a theorem if $Q_1,\ldots,Q_n$ are theorems," so that new axioms can only make more theorems derivable; they can never invalidate a previous theorem.

Recently there have been a number of attempts to formalize this type of nonmonotonic reasoning. The general idea is to allow axioms to be restrictive as well as permissive, by employing inference rules of the form "P is a theorem if $Q_1,\ldots,Q_n$ are not theorems." The inference that birds can fly is handled by having, in effect, a rule that says that, for any X, "X can fly" is a theorem if "X is a bird" is a theorem and "X cannot fly" is not a theorem. If all we are told about Tweety is that he is a bird, we will not be able to derive "Tweety cannot fly"; consequently, "Tweety can fly" will be inferable. If we are told that Tweety is a penguin and we already know that no penquin can fly, we will be able to derive the fact that Tweety cannot fly, and so the inference that Tweety can fly will be blocked.

One of the most interesting embodiments of this approach to nonmonotonic reasoning is McDermott and Doyle's "nonmonotonic logic" [McDermott and Doyle, 1980; McDermott, 1982]. McDermott and Doyle modify a standard first-order logic by introducing a sentential operator "M," whose informal interpretation is "is consistent." Nonmonotonic inferences about birds being able to fly would be sanctioned in their system by the axiom [McDermott, 1982, p. 33]

(ALL X)(BIRD(X) /\ M(CAN-FLY(X)) -> CAN-FLY(X)).

This formula can be read informally as "for all X, if X is a bird and it is consistent to assert that X can fly, then X can fly." McDermott and Doyle can then have a single general nonmonotonic inference rule, whose intuitive content is "MP is derivable if ~P is not derivable."

McDermott and Doyle's approach to nonmonotonic reasoning seems more interesting and ambitious than some other approaches in two respects. First, since the principles that lead to nonmonotonic inferences are explicitly represented in the logic, those very principles can be reasoned about. That is, if P is such a principle, we could start out believing Q -> P or even MP -> P, and come to hold P by drawing inferences, either monotonic or nonmonotonic. So, if we use McDermott and Doyle's representation of the belief that birds can fly, we could also represent various inferences that would lead us to adopt that belief. Second, since they use only general inference rules, they are able to provide a formal semantic interpretation with soundness and completeness proofs for each of the logics they define. In formalisms that use content-specific nonmonotonic inference rules dealing with contingent aspects of the world (i.e., it might have been the case that birds could not fly), it is difficult to see how this could be done. The effect is that nonmonotonic inferences in McDermott and Doyle's logics are justified by the meaning of the premises of the inferences.

There are a number of problems with McDermott and Doyle's nonmonotonic logics, however. The first logic they define [McDermott and Doyle, 1980] gives such a weak notion of consistency that, as they

point out, MP is not inconsistent with ~P. That is, it is possible for a theory to assert simultaneously that P is consistent with the theory and that P is false. McDermott subsequently [1982] tried basing nonmonotonic logics on the standard modal logics T, S4, and S5. He discovered, however, that the most plausible candidate for formalizing the notion of consistency that he wanted, nonmonotonic S5, collapses to ordinary S5 and is therefore monotonic. In the rest of this paper we develop an alternative formalization of nonmonotonic logic that shows why these problems arise in McDermott and Doyle's logics and how they can be avoided.

II NONMONOTONIC LOGIC AND AUTOEPISTEMIC REASONING

The first step in analyzing nonmonotonic logic is to determine what sort of nonmonotonic reasoning it is meant to model. After all, nonmonotonicity is a rather abstract syntactic property of an inference system, and there is no a priori reason to believe that all forms of nonmonotonic reasoning should have the same logical basis. In fact, McDermott and Doyle seem to confuse two quite distinct forms of nonmonotonic reasoning, which we will call default reasoning and autoepistemic reasoning. They talk as though their systems were intended to model the former, but they actually seem much better suited to modeling the latter.

By default reasoning we mean the drawing of plausible inferences from less-than-conclusive evidence in the absence of information to the contrary. The examples about birds being able to fly are of this type. If we know that Tweety is a bird, that gives us some evidence that Tweety can fly, but it is not conclusive. In the absence of information to the contrary, however, we are willing to go ahead and tentatively conclude that Tweety can fly. Now even before we do any detailed analysis of nonmonotonic logic, we can see that there will be problems in interpreting it as a model of default reasoning: In the formal semantics McDermott and Doyle provide for nonmonotonic logic, all the nonmonotonic inferences are valid. Default reasoning, however, is clearly not a form of valid inference.[1]

Consider the belief that lies behind our willingness to infer that Tweety can fly from the fact that Tweety is a bird. It is probably something like most birds can fly, or almost all birds can fly, or a typical bird can fly. To model this kind of reasoning, in a theory whose only axioms are "Tweety is a bird" and "Most birds can fly," we ought to be able to infer (nonmonotonically) "Tweety can fly." Now if

this were a form of valid inference, we would be guaranteed that the conclusion is true if the premises are true. This is manifestly not the case. The premises of this inference give us a good reason to draw the conclusion, but not the ironclad guarantee that validity demands.

Now reconsider McDermott's formula that yields nonmonotonic inferences about birds being able to fly:

(ALL X)(BIRD(X) /\ M(CAN-FLY(X)) -> CAN-FLY(X))

McDermott suggests as a gloss of this formula "Most birds can fly," which would indicate that he thinks of the inferences it sanctions as default inferences. But if we read M as "is consistent" as McDermott and Doyle repeatedly tell us to do elsewhere, the formula actually says something quite different: "For all X, if X is a bird and it is consistent to assert that X can fly, then X can fly." Since the inference rule for M is intended to convey "MP is derivable if ~P is not derivable," the notion of consistency McDermott and Doyle have in mind seems to be that it is consistent to assert P if ~P is not derivable. McDermott's formula, then, says that the only birds that cannot fly are the ones that can be inferred not to fly. If we have a theory whose only axioms are this one and an assertion to the effect that Tweety is a bird, then the conclusion that Tweety can fly would be a valid inference. That is, if it is true that Tweety is a bird, and it is true that only birds inferred not to fly are in fact unable to fly, and Tweety is not inferred not to fly, then it must be true that Tweety can fly.

This type of reasoning is not a form of default reasoning at all; it rather seems to be more like reasoning about one's own knowledge or belief. Hence, we will refer to it as autoepistemic reasoning. Autoepistemic reasoning, while different from default reasoning, is an important form of commonsense reasoning in its own right. Consider my reason for believing that I do not have an older brother. It is surely not that one of my parents once casually remarked, "You know, you don't have any older brothers," nor have I pieced it together by carefully

sifting other evidence. I simply believe that if I did have an older brother I would know about it; therefore, since I don't know of any older brothers, I must not have any. This is quite different from a default inference based on the belief, say, that most MIT graduates are eldest sons, and that, since I am an MIT graduate, I am probably an eldest son.

Default reasoning and autoepistemic reasoning are both nonmonotonic, but for different reasons. Default reasoning is nonmonotonic because, to use a term from philosophy, it is defeasible: its conclusions are tentative, so, given better information, they may be withdrawn. Purely autoepistemic reasoning, however, is not defeasible. If you really believe that you already know all the instances of birds that cannot fly, you cannot consistently hold to that belief and at the same time accept new instances of birds that cannot fly.[2]

As Stalnaker [1980] has observed, autoepistemic reasoning is nonmonotonic because the meaning of an autoepistemic statement is context-sensitive; it depends on the theory in which the statement is embedded.[3] If we have a theory whose only two axioms are

```
BIRD(TWEETY)
(ALL X)(BIRD(X) /\ M(CAN-FLY(X)) -> CAN-FLY(X)),
```

then MP does not merely mean that P is consistent--it means that P is consistent with the nonmonotonic theory that contains only those two axioms. We would expect CAN-FLY(TWEETY) to be a theorem of this theory. If we change the theory by adding ~CAN-FLY(TWEETY) as an axiom, we then change the meaning of MP to be that P is consistent with the nonmonotonic theory that contains only the axioms

```
~CAN-FLY(TWEETY)
BIRD(TWEETY)
(ALL X)(BIRD(X) /\ M(CAN-FLY(X)) -> CAN-FLY(X)),
```

and we would not expect CAN-FLY(TWEETY) to be a theorem. The operator M changes its meaning with context just as do indexical words in natural language, such as "I," "here," and "now." The nonmonotonicity

associated with autoepistemic statements should therefore be no more puzzling than the fact that "I am hungry" can be true when uttered by a particular speaker at a particular time, but false when uttered by a different speaker at the same time or the same speaker at a different time. So we might say that, whereas default reasoning is nonmonotonic because it is defeasible, autoepistemic reasoning is nonmonotonic because it is indexical.

III THE FORMALIZATION OF AUTOEPISTEMIC LOGIC

Rather than try directly to analyze McDermott and Doyle's nonmonotonic logic as a model of autoepistemic reasoning, we will first define a logic that demonstrably does model certain aspects of autoepistemic reasoning and then compare nonmonotonic logic with that. We will call our logic, naturally enough, autoepistemic logic. The language will be much like McDermott and Doyle's, an ordinary logical language augmented by autoepistemic modal operators. McDermott and Doyle treat consistency as their fundamental notion, so they take M as the basic modal operator and define its dual L to be ~M~. Our logic, however, will be based on the notion of belief, so we will take L to mean "is believed," treat it as primitive, and define M as ~L~. In any case, this gives us the same notion of consistency as theirs: a formula is consistent if its negation is not believed. Since there are some problems with regard to the meaning of quantifying into the scope of an autoepistemic operator that are not relevant to the main point of this paper, we will limit our attention to propositional autoepistemic logic.

Autoepistemic logic is intended to model the beliefs of an agent reflecting upon his own beliefs. The primary objects of interest are sets of autoepistemic logic formulas that are interpreted as the total beliefs of such agents. We will call such a set of formulas an autoepistemic theory. The truth of an agent's beliefs, expressed as a propositional autoepistemic theory, will be determined by (1) which propositional constants are true in the external world and (2) which formulas the agent believes. 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 notions of interpretation and model as follows:

We proceed in two stages. First we define a propositional interpretation of an autoepistemic theory T to be an assignment of truth-values to the formulas of the language of T that is consistent with the usual truth recursion for propositional logic and with any arbitrary assignment of truth-values to propositional constants and formulas of the form LP. A propositional model of an autoepistemic theory T is a propositional interpretation of T in which all the formulas of T are true. The propositional interpretations and models of an autoepistemic theory are, therefore, precisely those we would get in ordinary propositional logic by treating all formulas of the form LP as propositional constants. We therefore inherit the soundness and completeness theorems of propositional logic; i.e., a formula P is true in all the propositional models of an autoepistemic theory T if and only if it is a tautological consequence of T (i.e., derivable from T by the usual rules of propositional logic).

Next we define an autoepistemic interpretation of an autoepistemic theory T to be a propositional interpretation of T in which, for every formula P, LP is true if and only if P is in T. It should be noted that the theory T itself completely determines the truth of any formula of the form LP in all the autoepistemic interpretations of T, independently of the truth assignment to the propositional constants. Hence, for every truth assignment to the propositional constants of T, there is exactly one corresponding autoepistemic interpretation of T. Finally, an autoepistemic model of T is an autoepistemic interpretation of T in which all the formulas of T are true. So the autoepistemic interpretations and models of T are just the propositional interpretations and models of T that conform to the intended meaning of the modal operator L.

This gives us a formal semantics for autoepistemic logic that matches its intuitive interpretation. Suppose that the beliefs of an agent situated in a particular world are characterized by the autoepistemic theory T. The world in question will provide an assignment of truth-values for the propositional constants of T, and any

formula of the form LP will be true relative to the agent just in case he believes P. In this way, the agent and the world in which he is situated directly determine an autoepistemic interpretation of T. That interpretation will be an autoepistemic model of T, just in case all the agent's beliefs are true in his world.

Given this semantics for autoepistemic logic, what do we want from a notion of inference for the logic? From an epistemological perspective, the problem of inference is the problem of what set of beliefs (theorems) an ideally rational agent would adopt on the basis of his initial premises (axioms). Since we are trying to model the beliefs of a rational agent, the beliefs should be sound with respect the premises; we want a guarantee that the beliefs are true provided that the premises are true. Moreover, since we assume that the agent is ideally rational, the beliefs should be semantically complete; we want them to contain everything that the agent would be semantically justified in concluding from his beliefs and from the knowledge that they are his beliefs. An autoepistemic logic that meets these conditions can be viewed as a competence model of reflection upon one's own beliefs. Like competence models generally, it assumes unbounded resources of time and memory, and is therefore not a plausible model of any finite agent. It is, however, the model upon which the behavior of rational agents ought to converge as their time and memory resources increase.

Formally, we will say an autoepistemic theory T is sound with respect to an initial set of premises A if and only if every autoepistemic interpretation of T in which all the formulas of A are true is an autoepistemic model of T. This notion of soundness is the weakest condition that guarantees that all of the agent's beliefs are true whenever all his premises are true. Let I be the autoepistemic interpretation of T that is determined by what is true in the actual world (including what the agent actually believes). If all the formulas of T are true in every autoepistemic interpretation of T in which all the formulas of A are true, then all the formulas of T will be true in I

if all the formulas of A are true in I; hence, all the agent's beliefs will be true in the world if all the agent's premises are true in the world. However, if there is an autoepistemic interpretation of T in which all the formulas of A are true but some formulas of T are false, then it is possible that I is that interpretation, and that all the agent's premises will be true in the world, but some of his beliefs will not.

Our formal notion of completeness is that an autoepistemic theory T is semantically complete if and only if T contains every formula that is true in every autoepistemic model of T. If a formula P is true in every autoepistemic model of an agent's beliefs, then it must be true if all the agent's beliefs are true, and an ideally rational agent should be able to recognize that and infer P. On the other hand, if P is false in some autoepistemic model of the agent's beliefs, then that model, for all he can tell, might be the way the world actually is; he is therefore justified in not believing P.

The next problem is to give a syntactic characterization of the autoepistemic theories that satisfy these conditions. With a monotonic logic, the usual procedure is to define a collection of inference rules to apply to the axioms. For a nonmonotonic logic this is a nontrivial matter. Much of the technical ingenuity of McDermott and Doyle's systems lies simply in their formulation of a coherent notion of nonmonotonic derivability. The problem is that nonmonotonic inference rules do not yield a simple iterative notion of derivability the way monotonic inference rules do. We can view a monotonic inference process as applying the inference rules in all possible ways to the axioms, generating additional formulas to which the inference rules are applied in all possible ways, and so forth. Since monotonic inference rules are monotonic, once a formula has been generated at a given stage, it remains in the generated set of formulas at every subsequent stage. Thus the theorems of a theory in a monotonic system can be defined simply as all the formulas that are generated at any stage. The problem with attempting to follow this pattern with nonmonotonic inference rules

is that we cannot draw nonmonotonic inferences reliably at any particular stage, since something inferred at a later stage may invalidate them. Lacking such an iterative structure, nonmonotonic systems often use nonconstructive "fixed point" definitions, which do not directly yield algorithms for enumerating the "derivable" formulas, but do define sets of formulas that respect the intent of the nonmonotonic inference rules (e.g., in McDermott and Doyle's fixed points, MP is included whenever ~P is not included.)

For our logic, it is easiest to proceed by first specifying the closure conditions that we would expect the beliefs of an ideally rational agent to possess. Viewed informally, the beliefs should include whatever the agent could infer either by ordinary logic or by reflecting on what he believes. Stalnaker [1980] has put this formally by suggesting that a set of formulas T that represents the beliefs of an ideally rational agent should satisfy the following conditions:

1. If $P_1,\ldots,P_n$ are in T, and $P_1,\ldots,P_n \vdash Q$, then Q is in T (where "$\vdash$" means ordinary tautological consequence).

2. If P is in T, then LP is in T.

3. If P is not in T, then ~LP is in T.

Stalknaker [1980, p. 6] describes the state of belief characterized by such a theory as <u>stable</u> "in the sense that no further conclusions could be drawn by an ideally rational agent in such a state." We will therefore describe the theories themselves as <u>stable autoepistemic theories</u>.

There are a number of interesting observations we can make about stable autoepistemic theories. First we note that, if a stable autoepistemic theory T is consistent, it will satisfy two more intuitively sound conditions:

4. If LP is in T, then P is in T.

5. If ~LP is in T, then P is not in T.

Condition 4 holds because, if LP were in T and P were not, ~LP would be in T (by Condition 3) and T would be inconsistent.[4] Condition 5 holds because, if ~LP and P were both in T, LP would be in T (by Condition 2) and T would be inconsistent.

Conditions 2-5 imply that any consistent stable autoepistemic theory will be both sound and semantically complete with respect to formulas of the form LP and ~LP: If T is such a theory, then LP will be in T if and only if P is in T, and ~LP will be in T if and only if P is not in T. Thus, all the propositional models of a stable autoepistemic theory are autoepistemic models. Stability implies a soundness result even stronger than this, however. We can show that the truth of any formula of a stable autoepistemic theory depends only on the truth of the formulas of the theory that contain no autoepistemic operators. (We will call these formulas "objective.")

> Theorem 1. If T is a stable autoepistemic theory, then any autoepistemic interpretation of T that is a propositional model of the objective formulas of T is an autoepistemic model of T.

(The proofs of all theorems are given in the appendix.)

In other words, if all the objective formulas in an autoepistemic theory are true, then all the formulas in that theory are true. Given that the objective formulas of a stable autoepistemic theory determine whether the theory is true, it is not surprising that they also determine what all the formulas of the theory are.

> Theorem 2. If two stable autoepistemic theories contain the same objective formulas, then they contain exactly the same formulas.[5]

Finally, with these characterization theorems, we can prove that the syntactic property of stability is equivalent the semantic property of completeness.

> Theorem 3. An autoepistemic theory T is semantically complete if and only if T is stable.

By Theorem 3, we know that stability of an agent's beliefs guarantees that they are semantically complete, but stability alone does not tell us whether they are sound with respect to his initial premises. That is because the stability conditions say nothing about what an agent should not believe. They leave open the possibility of an agent's believing propositions that are not in any way grounded in his initial premises. What we need to add is a constraint specifying that the only propositions the agent believes are his initial premises and those required by the stability conditions. To satisfy the stability conditions and include a set of premises A, an autoepistemic theory T must include all the tautological consequences of A U {LP | P is in T} U {¬LP | P is not in T}. Conversely, we will say that an autoepistemic theory T is grounded in a set of premises A if and only if every formula of T is included in the tautological consequences of A U {LP | P is in T} U {¬LP | P is not in T}. The following theorem shows that this syntactic constraint on T and A captures the semantic notion of soundness.

> Theorem 4. An autoepistemic theory T is sound with respect to an initial set of premises A if and only if T is grounded in A.

From Theorems 3 and 4, we can see that the possible sets of beliefs that an ideally rational agent might hold, given A as his premises, ought to be just the extensions of A that are grounded in A and stable. We will call these the stable expansions of A. Note that we say "sets", because there may be more than one stable expansion of a given set of premises. For example, consider {¬LP -> Q, ¬LQ -> P} as an initial set of premises.[6] The first formula asserts that, if P is not believed, then Q is true; the second asserts that, if Q is not believed, then P is true. In any stable autoepistemic theory that includes these premises, if P is not in the theory, Q will be, and vice versa. But if the theory is grounded in these premises, if P is in the theory there will be no basis for including Q, and vice versa. Consequently, a stable expansion of {¬LP -> Q, ¬LQ -> P} will contain either P or Q, but not both.

It can also happen that there are no stable expansions of a given set of premises. Consider, for instance, {~LP -> P}.[7] If T is a stable autoepistemic theory that contains ~LP -> P, it must also contain P. If P were not in T, ~LP would have to be in the T, but then P would be in T--a contradiction. On the other hand, if P is in T, then T is not grounded in {~LP -> P}. Therefore no stable autoepistemic theory can be grounded in {~LP -> P}.

This seemingly strange behavior results from the indexicality of the autoepistemic operator L. Since L is interpreted relative to an entire set of beliefs, its interpretation will change with the various ways of completing a set of beliefs. In each acceptable completion of a set of beliefs, the interpretation of L will change to make that set stable and grounded in the premises. Sometimes, though, no matter how we try to form a complete a set of beliefs, the result never coincides with the interpretation of L in a way that gives us a stable set of beliefs grounded in the premises.

This raises the question of how to view autoepistemic logic as a logic. If we consider a set of premises A as axioms, what do we consider the theorems of A to be? If there is a unique stable expansion of A, it seems clear that we want this expansion to be the set of theorems of A. But what if there are several stable expansions of A--or none at all? If we take the point of view of the agent, we have to say that there can be alternative sets of theorems, or no set of theorems of A. This may be a strange property for a logic to possess, but, given our semantics, it is clear why this happens. An alternative (adopted by McDermott and Doyle with regard to their fixed points) is to take the theorems of A to be the intersection of the set of all formulas of the language with all the stable expansions of A. This yields the formulas that are in all stable expansions of A if there is more than one, and it makes the theory inconsistent if there is no stable expansion of A. This too is reasonable, but it has a different interpretation. It represents what an outside observer would know, given only knowledge of the agent's premises and that he is ideally rational.

IV ANALYSIS OF NONMONOTONIC LOGIC

Now we are in a position to provide an analysis of nonmonotonic logic that will explain its peculiarities in terms of autoepistemic logic. Briefly, our conclusions will be that the original nonmonotonic logic of McDermott and Doyle [1980] is simply too weak to capture the notions they wanted, and that McDermott's [1982] attempt to strengthen the logic does so in the wrong way.

McDermott and Doyle's first logic is very similar to our autoepistemic logic with one glaring exception; its specification includes nothing corresponding to our Condition 2 (if P is in T, then LP is in T). McDermott and Doyle define the nonmonotonic fixed points of a set of premises A, corresponding to our stable expansions of A. In the propositional case, their definition is equivalent to the following:

> T is a fixed point of A just in case T is the set of tautological consequences of A U {~LP | P is not in T}.

Our definition of a stable expansion of A, on the other hand, could be stated as

> T is a stable expansion of A just in case T is the set of tautological consequences of A U {LP | P is in T} U {~LP | P is not in T}.

In nonmonotonic logic, {LP | P is in T} is missing from the "base" of the fixed points. This makes it possible for there to be nonmonotonic theories with fixed points that contain P but not LP. So, under an autoepistemic interpretation of L, McDermott and Doyle's agents are omniscient as to what they do not believe, but they may know nothing as to what they do believe.

This explains essentially all the peculiarities of McDermott and Doyle's original logic. For instance, they note [1980, p. 69] that MC does not follow from M(C /\ D). Changing the modality to L, this is

equivalent to saying that ~LP does not follow from ~L(P \/ Q). The problem is that, lacking the ability to infer LP from P, nonmonotonic logic permits interpretations of L that are more restricted than simple belief. Suppose we interpret L as "inferable in n or fewer steps" for some particular n. P might be inferable in exactly n steps, and P \/ Q in n+1. According to this interpretation ~L(P \/ Q) would be true and ~LP would be false. Since this interpretation of L is consistent with McDermott and Doyle's definition of a fixed point, ~LP does not follow from ~L(P \/ Q). The other example of this kind noted by McDermott and Doyle is that {MC, ~C} has a consistent fixed point, which amounts to saying simultaneously that P is consistent with everything asserted and that P is false. But this set of premises is equivalent to {~LP, P}, which would have no consistent fixed points if LP were forced to be in every fixed point that contains P.

On the other hand, McDermott and Doyle consider it to be a problem that {MC -> D, ~D} has no consistent fixed point in their theory. Restated in terms of L, this set of premises is equivalent to {P -> LQ, P}. Since a stable autoepistemic theory containing these premises will also contain LQ, it must also contain Q to be consistent. (Otherwise it would contain ~LQ.) But Q is not contained in any theory grounded in the premises {P -> LQ, P}; it is possible for P -> LQ and P both to be true with respect to an agent while Q is false. So there is no consistent stable expansion of {P -> LQ, P} in autoepistemic logic; hence, this set of premises cannot be the foundation of an appropriate set of beliefs for an ideally rational agent. Thus, our analysis justifies nonmonotonic logic in this case, contrary to the intuition of McDermott and Doyle.

McDermott and Doyle recognized the weakness of the original formulation of nonmonotonic logic, and McDermott [1982] has gone on to develop a group of theories that are stronger because they are based on modal rather than classical logic. McDermott's nonmonotonic modal theories alter the logic in two ways. First, the definition of fixed point is changed to be equivalent to

> T is a fixed point of A just in case T is the set of modal consequences of A U {~LP | P is not in T},

where "modal consequence" means that P |- LP is used as an additional inference rule. Second, McDermott considers only theories that include as premises the axioms of one of the standard modal logics "T," "S4," and "S5."

Merely changing the definition of fixed point brings McDermott's logic much closer to autoepistemic logic. In particular, adding P |- LP as an inference rule means that all modal fixed points of A are stable expansions of A. However, adding P |- LP as an inference rule, rather than adding {LP | P is in T} to the base of T, has as a consequence that not all stable expansions of A are modal fixed points of A. The difference is that, in autoepistemic logic, if P can be derived from LP, then both can be in a stable expansion of the premises, whereas in McDermott's logic there must be a derivation of P that does not rely on LP. Thus, although in autoepistemic logic there is a stable expansion of {LP -> P} that includes P, in McDermott's logic there is no modal fixed point of {LP -> P} that includes P. It is as if, in autoepistemic logic, one can acquire the belief that P and justify it later by the premise that, if P is believed, then it is true. In nonmonotonic logic, however, the justification of P has to precede belief in LP. This makes the interpretation of L in nonmonotonic modal logic more like "justified belief" than simple belief.

Since we have already shown that autoepistemic logic requires no specific axioms to capture a competence model of autoepistemic reasoning, we might wonder what purpose is served by McDermott's second modification of nonmonotonic logic, the addition of the axioms of various modal logics. The most plausible answer is that, besides behaving in accordance with the principles of autoepistemic logic, an ideally rational agent might well be expected to know what some of those principles are. For instance, the modal logic T has all instances of the schema L(P -> Q) -> (LP -> LQ) as axioms. This says that the agent's beliefs are closed under modus ponens--which is true for an

ideally rational agent, so he might as well believe it. S4 adds the schema LP -> LLP, which means that, if the agent believes P, he believes that he believes it (Condition 2). S5 adds the schema ~LP -> L~LP, which means that, if the agent does not believe P, he believes that he does not believe it (Condition 3). Since all these formulas are always true with respect to any ideally rational agent, it seems plausible to expect him to adopt them as premises. Thus, S5 seems to be the most plausible candidate of the nonmonotonic logics as a model of autoepistemic reasoning.

The problem is that all of these logics also contain the schema LP -> P, which means that, if the agent believes P, then P is true--but this is not generally true, even for ideally rational agents.[8] It turns out that LP -> P will always be contained in any stable autoepistemic theory (that is, ideally rational agents always believe that their beliefs are true), but making it a premise allows beliefs to be grounded that otherwise would not be. As a premise the schema LP -> P can itself be justification for believing P, while as a "theorem" it must be derived from ~LP, in which case P is not believed, or from P, in which case P must be independently justified, or from some other grounded formulas. In any case, as a premise schema, LP -> P can sanction <u>any</u> belief whatsoever in autoepistemic logic. This is not generally true in modal nonmonotonic logic, as we have also seen, but it is true in nonmonotonic S5. The S5 axiom schema ~LP -> L~LP embodies enough of the model theory of autoepistemic logic to allow LP to be "self grounding": The schema ~LP -> L~LP is equivalent to the schema ~L~LP -> LP, which allows LP to be justified by the fact that its negation is not believed. This inference is never in danger of being falsified, but, from this and LP -> P, we obtain an unwarranted justification for believing P.

The collapse of nonmonotonic S5 into monotonic S5 follows immediately. Since LP -> P can be used to justify belief in any formula at all, there are no formulas that are absent from every fixed point of theories based on nonmonotonic S5. It follows that there are no formulas of the form ~LP that are contained in every fixed point of

theories based on nonmonotonic S5; hence there are no theorems of the form ~LP in any theory based on nonmonotonic S5. (Recall that the theorems are the intersection of all the fixed points.) Since these formulas are just the ones that would be produced by nonmonotonic inference, nonmonotonic S5 collapses to monotonic S5. In more informal terms, an agent who assumes that he is infallible is liable to believe anything, so an outside observer can conclude nothing about what he does not believe.

The real problem with nonmonotonic S5, then, is not the S5 schema; therefore McDermott's rather unmotivated suggestion to drop back to nonmonotonic S4 [1982, p. 45] is not the answer. The S5 schema merely makes explicit the consequences of adopting LP -> P as a premise schema that are implicit in the logic's natural semantics. If we want to base nonmonotonic logic on a modal logic, the obvious solution is to drop back, not to S4, but to what Stalnaker [1980] calls "weak S5"--S5 without LP -> P. It is much better motivated and, moreover, has the advantage of actually being nonmonotonic.

In autoepistemic logic, however, even this much is unneccessary. Adopting any of the axioms of weak S5 as premises makes no difference to what can be derived. The key fact is the following theorem:

> Theorem 5. If P is true in every autoepistemic interpretation of T, then T is grounded in A U {P} if and only if T is grounded in A.

An immediate corollary of this result is that, if P is true in every autoepistemic interpretation of T, then T is a stable expansion of A U {P} if and only if T is a stable expansion of A.

The modal axiom schemata of weak S5,

$$L(P \rightarrow Q) \rightarrow (LP \rightarrow LQ)$$
$$LP \rightarrow LLP$$
$$\sim LP \rightarrow L\sim LP,$$

simply state Conditions 1-3, so all their instances are true in every autoepistemic interpretation of any stable autoepistemic theory. The

nonmodal axioms of weak S5 are just the tautologies of propositional logic, so they are true in every interpretation (autoepistemic or otherwise) of any autoepistemic theory (stable or otherwise). It immediately follows by Theorem 5, therefore, that a set of premises containing any of the axioms of weak S5 will have exactly the same stable expansions as the corresponding set of premises without any weak-S5 axioms.

V CONCLUSION

McDermott and Doyle recognized that their original nonmonotonic logic was too weak; when McDermott tried to strengthen it, however, he misdiagnosed the problem. Because he was thinking of nonmonotonic logic as a logic of provability rather than belief, he apparently thought the problem was the lack of any connection between provability and truth. At one point he says "Even though ~M~P (abbreviated LP) might plausibly be expected to mean 'P is provable,' there was not actually any relation between the truth values of P and LP," [1982, p. 34], and later he acknowledges the questionability of the schema LP -> P, but says that "it is difficult to visualize any other way of relating provability and truth," [1982, p. 35]. If one interprets nonmonotonic logic as a logic of belief, however, there is no reason to expect any connection between the truth of LP and the truth of P. And, as we have seen, the real problem with the original nonmonotonic logic was that the "if" half of the semantic definition of L--that LP is true if and only if P is believed--was not expressed in the logic.

NOTES

[1] In their informal exposition, McDermott and Doyle [1980, pp. 44-46] emphasize that their notion of nonmonontonic inference is not to be taken as a form of valid inference. If this is the case, their formal semantics cannot be regarded as the "real" semantics of their nonmonotonic logic. At best, it would provide the conditions that would have to hold for the inferences to be valid, but this leaves unanswered the question of what formulas of nonmonotonic logic actually mean.

[2] Of course, autoepistemic reasoning can be combined with default reasoning; we might believe that we know about most of the birds that cannot fly. This could lead to defeasible autoepistemic inferences, but their defeasibility would be the result of their also being default inferences.

[3] Stalnaker's note, which to my knowledge remains unpublished, grew out of his comments as a respondent to McDermott at a Conference on Artificial Intelligence and Philosophy, held in March 1980 at the Center for Advanced Study in the Behavioral Sciences. N.B., the term "autoepistemic reasoning" is ours, not his.

[4] Condition 4 will, of course, also be satisfied by an inconsistent stable autoepistemic theory, since such a theory would include all formulas of autoepistemic logic.

[5] This theorem implies that our autoepistemic logic does not contain any "nongrounded" self-referential formulas, such as one finds in what are usually called "syntactical" treatments of belief. If, instead of a belief operator, we had a belief predicate, Bel, there might be a term p that denotes the formula Bel(p). Whether Bel(p) is believed or not is clearly independent of any objective beliefs. The lack of such formulas constitutes a characteristic difference between sentence-operator and predicate treatments of propositional attitudes and modalities.

[6] McDermott and Doyle [1980, p. 51] present this example as {MC -> ~D, MD -> ~C}.

[7] McDermott and Doyle [1980, p. 51] present this example as {MC -> ~C}.

[8] LP -> P would be an appropriate axiom schema if the interpretation of LP were "P is known" rather than "P is believed," but that notion is not nonmonotonic. An agent cannot, in general, know when he does not know P, because he might believe P--leading him to believe that he knows P--while P is in fact false. Since agents are unable to reflect directly on what they do not know (only on what they do not believe), an autoepistemic logic of knowledge would not be a nonmonotonic logic; rather, the appropriate logic would seem to be monotonic S5.

REFERENCES

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).

Minsky, M. [1974] "A Framework for Representing Knowledge," MIT Artificial Intelligence Laboratory, AIM-306, Massachusetts Institute of Technology, Cambridge, Massachusetts (June 1974).

Stalnaker, R. [1980] "A Note on Non-monotonic Modal Logic," Dept. of Philosophy, Cornell University, unpublished manuscript.

APPENDIX: PROOFS OF THEOREMS

Theorem 1. If T is a stable autoepistemic theory, then any autoepistemic interpretation of T that is a propositional model of the objective formulas of T is an autoepistemic model of T.

Proof. Suppose that T is a stable autoepistemic theory and I is an autoepistemic interpretation of T that is a propositional model of the objective formulas of T. All the objective formulas of T are true in I. T must be consistent because an inconsistent stable autoepistemic theory would contain all formulas of the language, which would include many objective formulas that are not true in I. Let P be an arbitrary formula in T. Since stable autoepistemic theories are closed under tautological consequence, T must also contain a set of formulas $P_1,\ldots,P_k$ that taken together entail P, where, for each i between 1 and k, there exist n and m such that P_i is of the form

$$P_{i,1} \vee LP_{i,2} \vee\ldots\vee LP_{i,n} \vee \neg LP_{i,n+1} \vee\ldots\vee \neg LP_{i,m}$$

and $P_{i,1}$ is an objective formula. (Any formula is interderivable with a set of such formulas by propositional logic alone.) There are two cases to be considered:

(1) Suppose at least one of $LP_{i,2},\ldots,LP_{i,n}$, $\neg LP_{i,n+1},\ldots,\neg LP_{i,m}$ is in T. By Conditions 4 and 5, we know that, if any such formula is in T, it must be true in I, since T is consistent and I is an autoepistemic interpretation of T. But, since each of these formulas entails P_i, it follows that P_i is also true in I.

(2) Suppose the first case does not hold. Conditions 2 and 3 guarantee that in every stable autoepistemic theory, for every formula P, either LP or ¬LP will be in the theory. Hence, if T does not contain any of $LP_{i,2},\ldots,LP_{i,n}$, $\neg LP_{i,n+1},\ldots,\neg LP_{i,m}$, it must contain all of $\neg LP_{i,2},\ldots,\neg LP_{i,n}$, $LP_{i,n+1},\ldots,LP_{i,m}$. But $P_{i,1}$ is a tautological consequence of P_i and these formulas (imagine repeated applications of

the resolution principle); so $P_{i,1}$ must be in T. But $P_{i,1}$ is objective, and so, by hypothesis, must be true in I. Since $P_{i,1}$ entails P_i, it must be the case that P_i is true in I.

In either case, P_i will be true in I. All the P_i taken together entail P, so P must also be true in I. Since P was chosen arbitrarily, every formula of T must be true in I; hence I is an autoepistemic model of T.

Theorem 2. If two stable autoepistemic theories contain the same objective formulas, then they contain exactly the same formulas.

Proof. Suppose that T_1 and T_2 contain the same objective formulas and T_1 contains P. We prove by induction on the depth of nesting of autoepistemic operators in P (the "L-depth" of P) that T_2 also contains P. If the L-depth of P is 0, the theorem is trivially true, since P will be an objective formula. Now suppose that P has an L-depth of d greater than 0, and that, if two stable autoepistemic theories contain the same objective formulas, then they contain exactly the same formulas whose L-depth is less than d.

Since stable autoepistemic theories are closed under tautological consequence, T_1 must also contain a set of formulas $P_1,\ldots,P_k$ that are interderivable with P by propositional logic, where, for each i between 1 and k, there exist n and m such that P_i is of the form

$$P_{i,1} \vee LP_{i,2} \vee \ldots \vee LP_{i,n} \vee \neg LP_{i,n+1} \vee \ldots \vee \neg LP_{i,m}$$

and $P_{i,1}$ is an objective formula. Note that, since propositional logic will treat all the formulas of the form $LP_{i,j}$ as propositional constants, it is impossible to increase the L-depth of a formula by propositional inference, so each of these formulas will have an L-depth of not more than d.

We can also assume that T_1 and T_2 are consistent. If one of these theories were inconsistent, it would contain all formulas of the language. Since, by hypothesis, the two theories contain the same

objective formulas, the other theory would contain all the objective formulas of the language and, since these formulas are inconsistent, it would also contain all the formulas of the language. For each P_i, there are three cases to be considered:

(1) T_1 contains $LP_{i,j}$ for some j between 2 and n. Since T_1 is consistent, by Condition 4 it must also contain $P_{i,j}$. Since the L-depth of $P_{i,j}$ is one less than that of $LP_{i,j}$, it must be less than d; so, by hypothesis, T_2 must contain $P_{i,j}$ and, by Condition 2, it must contain $LP_{i,j}$. But P_i is a tautological consequence of $LP_{i,j}$, so T_2 contains P_i.

(2) T_1 contains $\sim LP_{i,j}$ for some j between n+1 and m. Since T_1 is consistent, by Condition 5 it must not contain $P_{i,j}$. Since the L-depth of $P_{i,j}$ is one less than that of $\sim LP_{i,j}$, it must be less than d; therefore, by hypothesis, T_2 must not contain $P_{i,j}$ and, by Condition 3, it must contain $\sim LP_{i,j}$. But P_i is a tautological consequence of $\sim LP_{i,j}$, so T_2 contains P_i.

(3) Suppose neither of the first two cases holds. Conditions 2 and 3 guarantee that in every stable autoepistemic theory, for every formula P, either LP or $\sim LP$ will be in the theory. Hence, if T_1 does not contain any of $LP_{i,2},\ldots,LP_{i,n}$, $\sim LP_{i,n+1},\ldots,\sim LP_{i,m}$, it must contain all of $\sim LP_{i,2},\ldots,\sim LP_{i,n}$, $LP_{i,n+1},\ldots,LP_{i,m}$. But $P_{i,1}$ is a tautological consequence of P_i and these formulas; so $P_{i,1}$ must be in T_1. $P_{i,1}$ is objective, however, so $P_{i,1}$ must also be in T_2. Since P_i is a tautological consequence of $P_{i,1}$, T_2 contains P_i.

Thus, all of $P_1,\ldots,P_k$ are in T_2. Since P is a tautological consequence of these formulas, P is also in T_2. Since P was chosen arbitrarily, every formula in T_1 is also in T_2. The same argument can be used to show that every formula in T_2 is also in T_1, so T_1 and T_2 contain exactly the same formulas.

Theorem 3. An autoepistemic theory T is semantically complete if and only if T is stable.

Proof. "If" direction: we show that, if T is a stable autoepistemic theory, then T contains every formula that is true in every autoepistemic model of T. Let T be a stable autoepistemic theory and let P be an arbitrary formula that is not in T. We show that there is an autoepistemic model of T in which P is false.

We know from propositional logic that P is propositionally equivalent to (i.e., true in the same propositional models as) the conjunction of a set of formulas $P_1,\ldots,P_k$, where, for each i between 1 and k, there exist n and m such that P_i is of the form

$$P_{i,1} \vee LP_{i,2} \vee\ldots\vee LP_{i,n} \vee \neg LP_{i,n+1} \vee\ldots\vee \neg LP_{i,m}$$

and $P_{i,1}$ is an objective formula. Since P will be a tautological consequence of $P_1,\ldots,P_k$ and T is stable, Condition 1 guarantees that, if P is not in T, at least one of $P_1,\ldots,P_k$ must not be in T. Let P_i be such a formula. P_i is a tautological consequence of each of its disjuncts, so none of them can be in T. We show that there is an autoepistemic model of T in which all of these disjuncts are false.

Since $P_{i,1}$ is not in T, it must not be a tautological consequence of the objective formulas of T. Given this and the fact that $P_{i,1}$ is objective, it follows from the completeness theorem for propositional logic that there must be a truth assignment to the propositional constants of T in which $P_{i,1}$ is false and all the objective formulas of T are true. But, we can extend this truth assigment (or any truth assignment to the propositional constants of T—see Section III) to an autoepistemic interpretation of T. Call this interpretation I and note that $P_{i,1}$ is false in I. I will be a propositional model of the objective formulas of T; so, by Theorem 1, I is an autoepistemic model of T in which $P_{i,1}$ is false.

Now consider the other disjuncts of P_i. Note that Conditions 2 and 3 require that a stable theory contain all the formulas of the form LP or $\neg$LP that are true in the autoepistemic interpretations of the theory. Since none of $LP_{i,2},\ldots,LP_{i,n}$, $\neg LP_{i,n+1},\ldots,\neg LP_{i,m}$ are in T, none of $LP_{i,2},\ldots,LP_{i,n}$, $\neg LP_{i,n+1},\ldots,\neg LP_{i,m}$ are true in any autoepistemic

interpretation of T. In particular, none of $LP_{i,2},\ldots,LP_{i,n}$, $\neg LP_{i,n+1},\ldots,\neg LP_{i,m}$ are true in I. Therefore, I is an autoepistemic model of T in which, since all of the disjuncts of P_i are false, P_i itself is false. But P is propositionally equivalent to a conjunction that includes P_i, so I is an autoepistemic model of T in which P is false.

"Only if" direction: we show that, if T is semantically complete, then T is stable. Suppose T is semantically complete. For any formula P, if P is true in every autoepistemic model of T, then P is in T. Let I be an arbitrary autoepistemic model of T. If we can show that some formula P is true in I, P must be true in every autoepistemic model of T (since I is arbitrarily chosen) and, thus, P must be in T. We now show that T satisfies Conditions 1-3.

(1) Suppose $P_1,\ldots,P_n$ are in T and $P_1,\ldots,P_n \vdash Q$. Since I is a model of T, $P_1,\ldots,P_n$ will be true in I. Since $P_1,\ldots,P_n$ will is true in I and Q is a tautological consequence of $P_1,\ldots,P_n$, Q will also be true in I. Therefore, Q will be in T. (2) Suppose P is in T. Since I is an autoepistemic model of T, LP will be true in I. Therefore, LP will be in T. (3) Suppose P is not in T. Since I is an autoepistemic model of T, $\neg LP$ will be true in I. Therefore, $\neg LP$ will be in T.

Conditions 1-3 are all satisfied, so T is stable.

Theorem 4. An autoepistemic theory T is sound with respect to an initial set of premises A if and only if T is grounded in A.

Proof. "If" direction: suppose T is grounded in A. Every formula of T is therefore included in the tautological consequences of A U {LP | P is in T} U {$\neg LP$ | P is not in T}. We show that T is sound with respect to A--i.e., that every autoepistemic interpretation of T in which all the formulas of A are true is an autoepistemic model of T.

Let I be an autoepistemic interpretation of T in which all the formulas in A are true. We show that I is an autoepistemic model of T. If P is in A, then, trivially, P is true in I. If P is of the form LQ

and Q is in T, or if P is of the form ¬LQ and Q is not in T, then P is true in I because I is an autoepistemic interpretation of T. We have now shown that all the formulas in A U {LP | P is in T} U {¬LP | P is not in T} are true in I, so all their tautological consequences are true in I. But all the formulas of T are included in this set, so I is an autoepistemic model of T. Since I was an arbitrarily chosen autoepistemic interpretation of T in which all the formulas of A are true, every autoepistemic interpretation of T in which all the formulas of A are true is an autoepistemic model of T.

"Only if" direction: suppose T is sound with respect to A. Every autoepistemic interpretation of T in which all the formulas of A are true is therefore an autoepistemic model of T. We show that T is grounded in A--i.e., every formula of T is a tautological consequence of A U {LP | P is in T} U {¬LP | P is not in T}.

Let A′ = A U {LP | P is in T} U {¬LP | P is not in T}. Note that, for all P, if P is in T, LP will be in A′, so LP will be true in every propositional model of A′; however, if P is not in T, ¬LP will be in A′ and LP will not be true in any propositional model of A′. Therefore, in every propositional model of A′, LP is true if and only if P is in T, so every propositional model of A′ is an autoepistemic interpretation of T. Since every autoepistemic interpretation of T in which all the formulas of A are true is an autoepistemic model of T, every propositional model of A′ is an autoepistemic model of T. Since every formula in T is true in in every autoepistemic model of T, every formula in T is true in every propositional model of A′. By the completeness theorem for propositional logic, every formula of T is therefore a tautological consequence of A′. Hence T is grounded in A.

> Theorem 5. If P is true in every autoepistemic interpretation of T, then T is grounded in A U {P} if and only if T is grounded in A.

Proof. Suppose that P is true in every autoepistemic interpretation of T. For any set of premises A, the set of autoepistemic interpretations of T in which every formula of A U {P} is true is therefore the same as

the set of autoepistemic interpretations of T in which every formula of A is true. Thus, every autoepistemic interpretation of T in which every formula of A U {P} is true is an autoepistemic model of T if and only if every autoepistemic interpretation of T in which every formula of A is true is an autoepistemic model of T. Hence, T is sound with respect to A U {P} if and only if T is sound with respect to A. By Theorem 4, therefore, T is grounded in A U {P} if and only if T is grounded in A.