SEMANTICAL CONSIDERATIONS ON NONMONOTONIC LOGIC

Robert C. Moore
Artificial Intelligence Center
SRI International, Menlo Park, CA 94025

ABSTRACT

Commonsense reasoning is "nonmonotonic" in the sense that we often draw conclusions on the basis of partial information that we retract when we are given more complete information. Some of the most interesting products of the 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 give a reconstruction of nonmonotonic logic as a model of an Ideally rational agent's reasoning about his own beliefs. For the resulting system, called autoeplstemlc logic, we define an Intuitively based semantics for which we can show autoeplstemlc logic to be both sound and complete. We then compare the McDermott and Doyle logics to autoeplstemlc logic, showing how it avoids their peculiarities.

INTRODUCTION

It has become generally acknowledged In recent years that one important feature of ordinary commonsense reasoning that standard logics fall to capture is Its nonmonotonicity. An example frequently given to illustrate the point is the following: If we know that Tweety Is a bird, in the absence of evidence to the contrary, we will normally assume 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 where 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. Thus the set of theorems 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 monotonlc, 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,....Q_n are theorems," so new axioms can only make more theorems derivable; they can never result in a previous theorem being invalidated.

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'"»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, then we will not be able to derive "Tweety cannot fly," and the inference to "Tweety can fly" will go through. If we are told that Tweety is a penguin and we know that no penquin can fly, we will be able to derive the fact that Tweety cannot fly, and 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 licensed 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 in the logic. 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 making inferences, either monotonic or nonmonotonic. So, if we represent the belief that birds can fly in McDermott and Doyle's way, 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 the nonmonotonic inferences in McDermott and Doyle's logics are justified by the meaning of the formulas involved.

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. Subsequently McDermott [1982] tried basing nonmonotonic logics on the standard modal logics T, S4, and S5. McDermott discovered, however, that the most plausible candidate for capturing the notion of consistency that he wanted, nonmonotonic S5, collapses to ordinary S5 and is therefore monotonic. In the rest of this paper we show why these problems arise and how to avoid them.

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 are intended to model the former, but they actually seem much better suited to modeling the latter.

By default reasoning, we mean drawing plausible inferences from less than conclusive evidence in the absence of any 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 make the assumption 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.

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, the conclusion would be guaranteed to be true provided that 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 iron-clad guarantee that validity demands.

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

(ALL X)(BIRD(X) A 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 licenses 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 P is consistent if ~P is not asserted. McDermott's formula, then, says that the only birds than cannot fly are the ones that are asserted not to fly. If we have a theory whose only axioms are this and an assertion to the effect than 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 asserted to be unable to fly are in fact unable to fly, and Tweety is not asserted to be unable 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 surely know about it, and 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 oldest sons, and that, since I am an MIT graduate, I am probably an oldest 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.

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. 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. In this theory, we would expect CAN-FLY(TWEETY) to be a theorem. If we change the theory by adding ~CAN-FLY(TWEETY) as an axiom, then we 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 English words such as "I," "here," and "now." The nonmonotonicity of autoepistemic theories 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 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 we can show does capture what we want to model about autoepistemic reasoning and then compare nonmonotonic logic to 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 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. There are some problems about the meaning of quantifying into the scope of an autoepistemic operator that are not relevant to the main point of this paper, so we will limit our attention to propositional autoepistemic logic.

Autoepistemic logic is intended to model the beliefs of an Ideally rational agent reflecting upon his own beliefs. The primary objects of interest are sets of formulas of autoepistemic logic which 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 for an agent just in case 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 any arbitrary assignment of truth-values to propositional constants and formulas of the form LP. A propositional model (or simply model) of an autoepistemic theory is a propositional interpretation of T in which all the formulas of T are true. So the propositional interpretations and models of an autoepistemic theory are just those we would get In ordinary propositional logic by considering all formulas of the form LP to be 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. 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 models and interpretations of T that conform to the intended meaning of the modal operator L.

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 agent ought to adopt on the basis of his initial premises (axioms). Since we are trying to model the beliefs of a rational agent, we want the agent's beliefs to be sound with respect to his premises: We want a guarantee

that the beliefs are true provided that the premises are true. Moreover, as an ideally rational agent, we want the agent's beliefs to be semantically complete: We want the beliefs to contain everything that the agent would be semantically justified in concluding from the assumption that his beliefs are true and 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 being. It is, however, the model to 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 which is a propositional model of A is a model of T. This notion of soundness guarantees that all of the agent's beliefs are true whenever all his premises are true. Given our semantic definitions, the world will always be an autoepistemic interpretation of an autoepistemic theory T. So if all the formulas of T are true in every autoepistemic interpretation of T where all the formulas of A are true, it follows that if all the formulas in A are true in the world then all the formulas in T will be 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, If that model is the way the world actually is, then all the formulas of A will be true in the world and some formulas of T 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, for all he can tell that model might be the way the world actually is, so he should not believe P.

The next problem is to give a syntactic characterization of the autoepistemic theories that satisfy these conditions. In a monotonic logic, the usual procedure is to define a collection of inference rules to apply to the axioms. In a nonmonotonic logic this is a nontrivial matter. Much of the technical ingenuity of McDermott and Doyle's system lies in simply coming up with 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 is generated at a given stage, it stays in at every subsequent stage. Thus the theorems of a theory in a monotonic system can be defined simply as all the formulas that get generated at any stage. The problem with attempting to follow this pattern with nonmonotonic inference rules is that we cannot reliably draw nonmonotonic inferences at any particular stage, since something inferred at a later stage may invalidate them. Lacking such an iterative structure, nonmonotonic systems 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 giving the closure conditions that we would expect the beliefs of an ideally rational agent to possess. Informally, the beliefs should include whatever the agent could infer by ordinary logic, and whatever he could infer by reflecting on what he believes. To put this formally, an autoepisteraic theory T that represents the beliefs of an ideally rational agent should satisfy the following conditions:

I. 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).

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

III. 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 stable "in the sense that no further conclusions could be drawn by an ideally rational agent in such a state."[3] We will therefore describe the theories themselves as stable autoepistemic theories.

There are a number of interesting observations that 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:

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

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

If LP is in T and P were not in T, then by condition III ~LP would be in T, and T would be inconsistent.[4] If ~LP is in T and P were in T, then by condition II LP would be in T, and T would be inconsistent.

Conditions II - V 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. Stability implies a soundness result even stronger than this, however. We can show than the truth 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 "objective" formulas.)

> Theorem 1. If T is a stable autoepistemic theory, then any autoepistemic interpretation of T which is a proposltional model of the objective formulas of T is an autoepistemic model of T[5].

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.[6]

Finally, with these characterization theorems, we can prove that the syntactic criterion of stability captures the semantic criterion 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 tells us nothing about 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 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 (J {LP I 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 just in case every formula of T is included in the tautological consequences of A U {LP I 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.

We have shown, then, that the stable autoepistemic theories grounded In a set of premises A exactly characterize the possible sets of beliefs that an Ideally rational agent might be expected to hold given A as his premises. We will call these the stable expansions of A. Notice that we say "sets" (plural), because there may be more than one stable expansion of a given set of premises. Consider {~LP -> Q, ~LQ -> P} as an Initial set of premises. The first formula says that, if P is not believed, then Q is true, and the second says 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 then 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. So, 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}. No stable autoepistemic theory that contains this formula can fail to contain P. If P were not In the theory, ~LP would have to be In the theory, but then P would be in the theory—a contradiction. On the other hand, if P is In the theory, then the theory will not be grounded In the premises. So no stable autoepistemic theory can be grounded in {~LP -> P}.

This seemingly strange behavior Is the result of 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 way of completing 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 complete a set of beliefs, the set of beliefs and the interpretation of L never coincide in a way that gives us a stable set of beliefs grounded in the premises.

This raises a 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, then It is clear that we want this expansion to be the set of theorems of A. But what if there are several, or no, stable expansions of A? If we take the point of view of the agent, we have to say that there can be alternate sets of theorems, or no set of theorems of A. This may be a strange property for a logic to have, 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 entire language with all the stable expansions of A. This gives us 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 an ideally rational agent.

IV ANALYSIS OF NONMONOTONIC LOGIC

Now we are in a position to give an analysis of nonmonotonic logic that will explain It6 peculiarities in terms of autoepistemic logic. Briefly, our conclusion will be that the original nonmonotonic logic of McDermott and Doyle [1980] Is simply too weak to capture the notions they wanted, but 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; they have nothing in their logic corresponding to our condition II (if P Is in T, then LP is in T). Analogous to our stable expansions of a set of premises A, McDermott and Doyle define the nonmonotonic fixed points of A. In the proposltlonal 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 I P is not in T}.

In nonmonotonic logic, {LP I P is In T} is missing from the "base" of the fixed points. This admits fixed points of nonmonotonic logic 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. 69J that MC does not follow from M(C A 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 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. On this Interpretation ~L(P \/ Q) would be true and "LP would be false. Since this interpretation of L is consistent with the definition of a fixed point, "LP does not follow from ~L(P \/ Q). The other example of this kind they note 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. This set of premises is equivalent to {"LP, P}, but this would be inconsistent 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. In autoepistemic logic, there are consistent stable theories that contain these premises, but none of them are grounded In the premises, so they do not characterize appropriate sets of beliefs for a rational agent. Thus, our analysis justifies nonmonotonic logic in this case, against the Intuitions of McDermott and Doyle.

McDermott and Doyle recognized the weakness of the original formulation of nonmonotonic logic, and McDermott [1982] has developed a group of theories that are stronger because they are based on modal logic 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 I P is not In T},

where "modal consequence" means that P I- LP is Included as an Inference rule. Second, McDermott considers only theories that have the axioms of the modal logics T, SA, or S5 among the premises.

Changing the definition of fixed point by Itself brings McDermott's logic much closer to autoepistemic logic- In particular, adlng P |- LP as an inference rule means that all modal fixed points of A are stable expansions of A. However, adding P I- LP as an Inference rule rather than adding {LP I P is in T} to the base of T, means that not all stable expansions of A are modal fixed points of A. The difference Is that, In autoepistemic logic, LP only has to be true to be in a set of beliefs; in McDermott's logic it must be grounded in a derivation of P that does not rely on LP. So, in autoepistemic logic there Is a stable expansion of {LP -> P} that includes P, but 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 its belief. From the point of view of the autoepistemic interpretation of L, modal nonmonotonic logic is more conservative than it has to be.

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 to nonmonotonic logic, the addition of the axioms of various modal logics. The most plausible answer is that, in addition to 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 certainly true, so the agent 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 II. 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 III. Since all these formulas always are true for 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 autoeplstemic 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, and that is not generally true for Ideally rational agents. It turns out that LP -> P will always be contained in any stable autoeplstemic 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 we have seen, as a premise the LP -> P can itself be justification for believing P, while, as a "theorem," it must either 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 license <u>any</u> belief whatsoever in autoeplstemic 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 autoeplstemic 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 get an unwarranted Justification for believing P.

The collapse of nonmonotonic S5 into monotonic S5 immediately follows. 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 S5, so there are no formulas of the form ~LP that are contained in every fixed point of 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 takes It as a premise 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, so 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 has the advantage of actually being nonmonotonic.

In autoeplstemic 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 autoeplstemic interpretation of T, then T is a stable expansion of A U {P} Iff T is a stable expansion of A.

The modal axiom schemata of weak S5,

L(P -> Q) -> (LP -> LQ)
LP -> LLP
~LP -> L~LP,

simply state conditions I - III, so all their instances are true in every autoeplstemic interpretation of any stable autoeplstemic theory. The nonmodal axioms of weak S5 are just the tautologies of propositional logic, so they are true in every interpretation (autoeplstemic or otherwise) of any autoeplstemic theory (stable or otherwise). It immediately follows by Theorem 5, then, 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 S5 axioms.

V CONCLUSION

McDermott and Doyle recognized that their original nonmonotonic logic was too weak, but, when he tried to strengthen it, McDermott misdiagnosed the problem. It appears that, because he was thinking of nonmonotonic logic as a logic of provability rather than belief, he thought the problem was the lack of a 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 questlonablllty 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—LP Is true If and only If L is believed—was not expressed In the logic.

ACKNOWLEDGMENTS

This research was supported by the Air Force Office of Scientific Research under Contract No. F49629-82-K-0031.

NOTES

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, then their formal semantics cannot be considered to be 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 mean.

2 Of course, autoepistemlc reasoning can be mixed with default reasoning; we might believe that we know about most of the birds that cannot fly. This could lead to defeasible autoepistemlc Inferences, but their defeasiblllty would be the result of their also being default Inferences.

Stalnaker s note, which to ray 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.

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

Space does not permit the inclusion of of the proofs of the theorems. They are given In full In elsewhere [Moore, 1983].

This theorem implies that our autoepistemlc logic does not contain any purely self-referential formulas as one finds in what are usually called "syntactical" treatments of belief. If, Instead of a belief operator, we had a belief predicate, 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 is a characteristic difference between sentence-operator and predicate treatments of proposltlonal attitudes and modalities.

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

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

REFERENCES

[1] McDermott, D. and J. Doyle [1980] "Non-Monotonlc Logic I," Artificial Intelligence, Vol. 13, Nos. 1, 2, pp. 4l-72 (April 1980).

[2] 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)."

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

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

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