

AN INTRODUCTION TO NON-MONOTONIC LOGIC^{* **}

Drew McDermott
Department of Computer Science
Yale University
New Haven, Connecticut 06520

Jon Doyle
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

"Non-monotonic" logical systems are logics in which the introduction of new axioms can invalidate old theorems. Such logics are very important in modeling the beliefs of active processes which, acting in the presence of incomplete information, must make and subsequently revise predictions in light of new observations. We present the motivation and history of such logics, develop a model theory for one important non-monotonic logic, and prove the completeness of the first-order non-monotonic predicate calculus.

1. THE PROBLEM OF INCOMPLETE KNOWLEDGE

The relation between formal logic and the operation of the mind has always been unclear. Some of the more striking differences between properties of formal logics and mental phenomenology occur in situations dealing with perception, ambiguity, common-sense, causality and prediction. One common feature of these problems is that they seem to involve working with incomplete knowledge. Perception must account for the noticing of overlooked features, common-sense ignores myriad special exceptions, assigners of blame can be misled, and plans for the future must consider never-to-be-realized contingencies. It is this apparently unavoidable making of mistakes in these cases that leads to some of the deepest problems of the formal analysis of mind.

* Drew McDermott acknowledges the support of a Josiah Willard Gibbs instructorship. Jon Doyle thanks the Fannie and John Hertz Foundation for supporting his research with a graduate fellowship. This research was conducted in part at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract number N00014-75-C-0643, and in part by NSF Grant MCS77-04828.

** This paper condenses the first part of "Non-Monotonic Logic I" (appearing as MIT AI Memo 486, and to appear in *Artificial Intelligence*), which discusses proof theory for non-monotonic logic, the evolution of logical theories, the relationship of non-monotonic logic to stronger logics, logics of incomplete information, and how non-monotonic logic models the logic of Doyle's Truth Maintenance System.

Some studies of these problems occur in the philosophical literature, the most relevant here being Rescher's [15] analysis of counterfactual conditionals and belief-contravening hypotheses. In artificial intelligence, studies of perception, ambiguity and common-sense have led to knowledge representations which explicitly and implicitly embody much information about typical cases, defaults, and methods for handling mistakes. [11,14] Studies of problem-solving and acting have attempted representing predictive and causal knowledge so that decisions to act require only limited contemplation, and that actions, their variations, and their effects can be conveniently described and computed. [1,2] Indeed, one of the original names applied to these efforts, "heuristic programming", stems from efficiency requirements forcing the use of methods which occasionally are wrong or which fail. The possibility of failure means that formalizations of reasoning in these areas must capture the process of revisions of perceptions, predictions, deductions and other beliefs.

Classical symbolic logic lacks tools for describing how to revise a formal theory to deal with inconsistencies caused by new information. This lack is due to a recognition that the general problem of finding and selecting among alternate revisions is very hard. (For an attack on this problem, see [15]. [12] surveys the complexities.) Although logicians have been able to ignore this problem, philosophers and researchers in artificial intelligence have been forced to face it because humans and computational models are subject to a continuous flow of new information. One important insight gained through computational experience is that there are at least two different problems involved, what might be called "routine revision" and "world-model reorganization".

World-model reorganization is the very hard problem of revising a complex model of a situation when it turns out to be wrong. Routine revision, on the other hand, is the problem of maintaining a set of facts which, although expressed as universally true, have exceptions. For example, a program may have the belief that all animals with beaks are birds. Telling this program about a platypus will cause a contradiction, but intuitively not as serious a contradiction as those requiring total revision. Classical logics, by lumping all contradictions together, have overlooked the possibility of handling the easy ones by expanding the notation in which rules are stated. That is, we could have avoided this problem by stating the belief as "If something is an animal with a beak, then *unless proven otherwise*, it is a bird." If we allow statements of this kind, the problem becomes how to coordinate sets of such rules. Each such statement may be seen as providing a piece of advice about belief revision; for our approach to make sense, all the little pieces of advice must determine a unique revision. This is the subject of this paper. Of course, even if we are successful, the world-model reorganization problem will still be unsolved. But we hope factoring out the routine revision problem will make the more difficult problem clearer.

2. APPROACHES TO NON-MONOTONIC LOGIC AND THE SEMANTICAL DIFFICULTIES

Traditional logics are called *monotonic* because the theorems of a theory are always a subset of the theorems of any extension of the theory. ([11] introduced this name for this property of classical logics. [2] terms it the "extension" property.) In this paper, by *theory* we will mean a set of axioms. A more precise statement of monotonicity is this: If A and B are two theories, and $A \subseteq B$, then $Th(A) \subseteq Th(B)$, where $Th(S) = \{p: S \vdash p\}$ is the set of theorems of S . We will be even more precise about the definition of $\vdash$ later.

Monotonic logics lack the phenomenon of new information leading to a revision of old conclusions. We obtain *non-monotonic logics* from classical logics by extending them with a modality ("consistent") well-known in artificial intelligence circles, and show that the resulting logics have well-founded, if unusual, model and proof theories. We introduce the proposition-forming modality M (read "consistent"). Informally, Mp is to mean that p is consistent with everything believed. (See [8].) One small theory employing this modality is

- (1) $\text{noon} \wedge M[\text{sun-shining}] \supset \text{sun-shining}$
- (2) noon
- (3) $\text{eclipse} \supset \neg \text{sun-shining},$

in which we can prove

- (4) $\text{sun-shining}.$

If we add the axiom

- (5) eclipse

then (4) is inconsistent, so (4) is not a theorem of the extended theory.

The use of non-monotonic techniques has some history, but until recently the intuitions underlying these techniques were inadequate and led to difficulties involving the semantics of non-monotonic inference rules in certain cases. We mention some of the guises in which non-monotonic reasoning methods and belief revising processes have appeared.

In PLANNER [3], a programming language based on a negationless calculus, the THNOT primitive formed the basis of such reasoning. THNOT, as a goal, succeeded only if its argument failed, and failed otherwise. Thus if the argument to THNOT was a formula to be proved, the THNOT would succeed only if the attempt to prove the embedded formula failed.

Two related forms of non-monotonic deductive systems are those described in [8] and [16]. McCarthy and Hayes give some indications of how actions might be described using modal operators like "normally" and "consistent", but present no detailed guidelines on how such operators might be carefully defined. Sandewall, in a deductive system applied to the frame problem used a deductive representation of non-monotonic rules based on a primitive called UNLESS. This was used to deduce conditions of situations resulting from actions except in those cases where properties of the action changed the extant conditions. Thus one might say that things retain their color unless painted.

Sandewall's interpretation of UNLESS was in accord with then current intuitions: UNLESS(p) is true if p is not deducible from the axioms using the classical first-order inference rules. Unfortunately, this definition has several problems, as pointed out by Sandewall. One problem is that it can happen that both p and UNLESS(p) are deducible, since from a rule like "from UNLESS(C) infer D " D can be inferred, but at the same time UNLESS(D) is also deducible since D is not deducible by classical rules. These

problems are partly due to the dependence of the notion of "deducible" on the intention of deduction rules based on "not deducible". This question-begging definition leads to perplexing questions of beliefs when complicated relations between UNLESS statements are present. For example, given the axioms

A
 $A \wedge \text{Unless}(B) \supset C$
 $A \wedge \text{Unless}(C) \supset B,$

we are faced with the somewhat paradoxical situation that either B or C can be deduced, but not both simultaneously. On the other hand, in the axiom system

A
 $A \wedge \text{Unless}(B) \supset C$
 $A \wedge \text{Unless}(C) \supset D$
 $A \wedge \text{Unless}(D) \supset E,$

one would expect to see A, C and E believed, and B and D not believed. One might be tempted to dismiss these anomalous cases as uninteresting. In fact, such cases are not perverse; rather, they occur naturally and are very important in many applications.

Spurred by Sandewall's presentation of the problems arising through such non-monotonic inference rules, Kramosil [5] considered sets of inference rules of the form

"From $\vdash p, \vdash q$, infer $\vdash r$ ",

where $\vdash$ and $\vdash$ are tokens of the meta-language and the number of antecedents can be arbitrary. ([19] presents a working system organized around a theory-metatheory distinction in which such rules can be written.) Kramosil defined the set of theorems in such a system as the intersection of all subsets of the language closed under the inference rules. He noted that this set may not itself be closed under the inference rules, and showed that in the special case in which the inference rules preserve truth values (that is, are effectively monotonic) that if the set of theorems of the monotonic inference rules alone is also closed with respect to the non-monotonic inference rules, then this set is the set of non-monotonic theorems. Kramosil's conclusion was that a set of inference rules defines a formalized theory (one in which all formulas have a well-defined truth value) if and only if this same theory is that of the monotonic inference rules alone, which he interprets to mean that the non-monotonic rules are either useless or meaningless.

As we will show in this paper, Kramosil's interpretation was too pessimistic with regard to the possibility of formalizing such rules and their unusual properties. As we have argued above, the purpose of non-monotonic inference rules is not to add certain knowledge where there is none, but rather to guide the selection of tentatively held beliefs in the hope that fruitful investigations and good guesses will result. This means that one should not *a priori* expect non-monotonic rules to derive valid conclusions independent of the monotonic rules. Rather one should expect to be led to a set of beliefs which while perhaps eventually shown incorrect will meanwhile coherently guide investigations.

One class of non-monotonic inferences consist of what might be called "minimal" inferences, in which a minimal model for some set of beliefs is assumed by assuming the set of beliefs to be a complete description of a state of affairs. Such techniques are discussed in [4,9,13], but none of these discuss the problem of revising beliefs in the face of new information or contradictions.

The semantical problems and routine belief revisions problems were in large part resolved by Doyle in his Truth Maintenance System (TMS) [1], and in related systems [6,7,17,18]. In the TMS, each statement has an associated set of justifications, each of which represents a reason for holding the statements as a belief. These justifications are used to determine the set of current beliefs by examining the recorded justifications to find well-founded support (non-circular proofs) whenever possible for each belief. When hypotheses change, these justifications are again examined to update the set of current beliefs. This scheme provides a more accurate version of antecedent and erasing procedures of PLANNER without the need to explicitly check for circular proofs. The non-monotonic capability appears as so-called non-monotonic justifications, which support belief just in case some specified statements are believed, and other specified statements are not believed. This allows, for example, belief in a statement to be justified whenever no proof of the negation of the statement is known. This representation of non-monotonic justifications, in combination with the belief revision algorithms, produced the first system capable of performing the routine revision of apparently inconsistent theories into consistent theories.

3. LINGUISTIC PRELIMINARIES

We settle on a language L which will be the language of all theories mentioned in the following. L is the usual language of predicate calculus augmented by the unary modality M . In this paper, the letters C, D, E and F will be used as syntactic variables ranging over propositional constant

letters. The letters p , q and r will be used for formulas.

The inferential system used defines a first-order theory to be a set of axioms including the following infinite class of axioms: For all formulas p , q and r :

- (6) (i) $p \supset [q \supset p]$
- (ii) $[p \supset [q \supset r]] \supset [[p \supset q] \supset [q \supset r]]$
- (iii) $[\neg q \supset \neg p] \supset [[\neg q \supset p] \supset q]$
- (iv) $\forall x p(x) \supset p(t)$

where $p(x)$ is a formula and t is a constant or a variable free for x in $p(x)$ and $p(t)$ denotes the result of substituting t for every free occurrence of x in $p(x)$, and

- (v) $\forall x [p \supset q] \supset [p \supset \forall x q]$

if p is a formula containing no free occurrence of x . (These axioms are from [10].) These are the *logical* axioms. All other axioms are called *proper*, or non-logical axioms. The theory with no proper axioms is called the *predicate calculus* (PC). (Note that this theory also contains strings containing the letter M , so it is actually not strict PC.) The *sentential calculus* (SC) consists of axioms which are instances of (i), (ii) and (iii) only. A theory consisting only of the sentential calculus plus a finite number of statements is called a *statement theory*. In this paper, the letters A and B will be used to stand for theories.

4. PROOF-THEORETIC OPERATORS

The monotonic rules of inference we will use (also from [10]) are

- (7) Modus Ponens: from p and $p \supset q$, infer q
- Generalization: from p , infer $\forall x p$.

If S is a set of formulas, and p follows from S and the axioms of A by the rules (7), we say $S \vdash_A p$. We abbreviate $\vdash_{PC}$ by $\vdash$ alone. We define $Th(S) = \{p: S \vdash p\}$.

The particular inference rules (7) are not very important. Later in the paper, when we concentrate on statement theories, the rule of generalization will be dropped without much fanfare. All that is important is that the operator Th have the following properties, which together are called *monotonicity*:

- (8) (i) $A \subseteq Th(A)$
- (ii) If $A \subseteq B$, then $Th(A) \subseteq Th(B)$,

and the property (9) of *idempotence*

$$(9) \quad Th(Th(A)) = Th(A).$$

Clearly, any classical inference system satisfies these conditions. Condition (9) can also be viewed as a fixed point equation, stating that the set of theorems monotonically derivable from a theory is a fixed point of the operator which computes the closure of a set of formulas under the monotonic inference rules. A well-known property of the monotonic inference rules is that $Th(A)$ is the smallest fixed point of this closing process; in fact, that $Th(A)$ is the intersection of all S such that $A \subseteq S$ and $Th(S) = S$.

In order to deal with non-monotonic logic, we need a new inference rule like this one (which we will take back immediately):

$$(10) \quad \text{"If } \nvdash_A \neg p, \text{ then } \vdash_A Mp."$$

That is, if a formula's negation is not derivable, it may be inferred to be consistent. As it stands, however, this rule is of no value because it is circular. "Derivable" means "derivable from axioms by inference rules", so we cannot define an inference rule in terms of derivability so casually.

Instead, we retain the definition of $\vdash$ as meaning monotonic derivability, and define the operator NM as follows: for any first-order theory A and any set of formulas $S \subseteq L$ (L , recall, is the entire language), let

$$(11) \quad NM_A(S) = Th(A \cup As_A(S)),$$

where $As_A(S)$, the set of *assumptions* from S , is given by

$$(12) \quad As_A(S) = \{Mq: q \in L \text{ and } \neg q \notin S\} - Th(A).$$

Notice that theorems of A of the form Mq are never counted as assumptions. NM_A takes a set S and produces a new set which includes $Th(A)$ but also includes much more: everything provable from the enlarged set of axioms and assumptions which is the original theory together with all assumptions not ruled out by S . We would like to define $TH(A)$, the set of theorems non-monotonically derivable from A , by analogy with the monotonic case as

$$(13) \quad "TH(A) = \text{the smallest fixed point of } NM_A."$$

This "definition" tries to capture the idea of adding the non-monotonic inference rule (10) to a first-order theory A . This is plausible, since it demands a set such that all of its elements may be proven from axioms and assumptions not wiped out by the proofs. Unfortunately, there is in general

no appropriate fixed point of NM_A . It can happen that a theory has no fixed point under the operator NM_A . Even if there are fixed points, there need not be a smallest fixed point.

For example, consider the theory $T1$ obtained as

$$(14) \quad T1 = PC \cup \{ MC \supset \neg D, MD \supset \neg C \},$$

where C and D are propositional constants. NM_{T1} has two fixed points, which can be called $F1$ and $F2$. $F1$ contains $\neg C$ but not $\neg D$, and $F2$ contains $\neg D$ but not $\neg C$. Since $\neg D$ is not in $F1$, MD is in $F1$, and so $\neg C$ is in $F1$. Similarly, the presence of $\neg D$ in $F2$ keeps $\neg C$ out and MC in $F2$. The problem is that neither $F1 \cap F2$ nor $F1 \cup F2$ is a fixed point of NM_{T1} . Since neither $\neg C$ nor $\neg D$ is in $F1 \cap F2$, MC and MD are both in $NM_{T1}(F1 \cap F2)$, so $\neg C$ and $\neg D$ are in $NM_{T1}(F1 \cap F2)$, so $F1 \cap F2 \neq NM_{T1}(F1 \cap F2)$. Similarly, both $\neg C$ and $\neg D$ are in $NM_{T1}(F1 \cup F2)$, so applying NM_{T1} to the union results in a smaller set. So in this case there is no natural status for $\neg C$ and $\neg D$.

An example of a theory with no fixed point of the corresponding operator is the theory $T2$ obtained as

$$(15) \quad T2 = PC \cup \{ MC \supset \neg C \}.$$

In this case, NM_{T2} has no fixed point, since alternate applications of the operator to any set produce new sets in which either both MC and $\neg C$ exist or neither exist.

Therefore, we must accept a somewhat less elegant definition of TH. Let us define TH as follows:

$$(16) \quad TH(A) = \bigcap (\{L\} \cup \{S: NM_A(S) = S\}).$$

That is, the set of provable formulas is the intersection of all fixed points of NM_A , or the entire language if there are no fixed points. We will use the abbreviation $A \vdash p$ to indicate that $p \in TH(A)$. With this definition, neither MC nor MD is a theorem of $T1$ in (14), but $MC \vee MD$ is. In the following, we will abbreviate $\{S: NM_A(S) = S\}$ as $FP(A)$, and (somewhat abusing the terms) call the elements of this set *fixed points* of the theory A .

5. MODEL THEORY

An *interpretation* V of formulas over a language L is a pair $\langle X, U \rangle$, where X is a nonempty set, and U is a function which associates relations and values over the domain X with each predicate, variable, constant and propositional constant letter in the usual fashion. That is, for each n -ary

predicate letter P , $U(P) \subseteq X^n$; for each variable or constant x , $U(x) \in X$; and for each propositional constant letter C , $U(C) \in \{0, 1\}$. Using this mapping function U we define the value $V(p)$ of a formula p in the interpretation V to be an element of $\{0, 1\}$ satisfying the following conditions: For an atomic formula $p(x_1, \dots, x_n)$, the value is 1 if $\langle U(x_1), \dots, U(x_n) \rangle \in U(p)$, and is 0 otherwise. $V(\neg p) = 1$ if $V(p) = 0$, and is 0 otherwise. $V(p \supset q) = 1$ if either $V(p) = 0$ or $V(q) = 1$, and is 0 otherwise. $V(\forall x p) = 1$ if for all $y \in X$, $V'(p) = 1$, where $V' = \langle X [y/x]U \rangle$, where $[y/x]U$ is the mapping derived from U by changing its value at the point x to the value y . $V(\forall x p) = 0$ otherwise. If $V(p) = 1$, we say that V *satisfies* p , and write $V \models p$.

A *monotonic model* of a set of formulas $S \subseteq L$ is an interpretation V which satisfies each formula in S , that is, $V(p) = 1$ for each formula $p \in S$. A *non-monotonic model* of a theory A is a pair $\langle V, S \rangle$, where V is a monotonic model of S , and $S \in FP(A)$. When the context makes the intended meaning clear, we will use the term *model* of A to mean either a non-monotonic model, a monotonic model, or an element of $FP(A)$ for the theory A .

Although unorthodox, this definition provides a meaning for formulas Mp which reflects the proof-theoretic property that " p is consistent with what is believed". This notion is made precise by including in the model a set of "current assumptions" (namely, $As_A(S)$). A model for a theory must assign 1 to all of these assumptions, so the effect is that Mp is assigned 1 in a model if $\neg p$ is not derivable and $\neg Mp$ is not derivable from the current assumptions and the original theory, that is, if p is consistent with what is "believed" in the model. Unfortunately, Mp may be assigned 1 in some model even when $\neg p$ is derivable (for example, when no axiom mentions Mp at all). This indicates that the logic is too weak. We discuss this problem in the full paper. We present a stronger logic, with a more elegant model theory, in a forthcoming paper.

With this definition of model, we can justify the definition of provability.

Theorem 1. (Soundness) If $A \vdash p$, then $V \models p$ for all models $\langle V, S \rangle$ of A .

Proof: Assume $A \vdash p$. If there are no models of A , the theorem follows trivially. Otherwise, p is a member of every fixed point of A . But since every model of A is a monotonic model of a fixed point of A , every model assigns 1 to p . ■

Theorem 2. (Completeness) If $V \models p$ for all models $\langle V, S \rangle$ of

A, then $A \vdash p$.

Proof: Assume that it is not true that $A \vdash p$. Thus there is a fixed point S of NMA which does not contain p. Now $Th(S) = S$ by idempotence, so $S \not\vdash p$. But the predicate calculus is complete, so some monotonic model V of S has $V(p) = 0$. ■

It is not surprising that we have completeness, since the definition of truth makes reference to provability. The proof was for first-order theories, but it can easily be generalized to any complete formal logic. For example, if we take care not to confuse M with the SS operator "possibly", we can easily get a complete non-monotonic extension of SS.

ACKNOWLEDGEMENTS

We wish to thank Gerald Jay Sussman, Rohit Parikh, David Harel, Mitch Marcus, Lucia Vaina, and Vaughan Pratt for helpful criticisms and comments.

REFERENCES

[1] Doyle, J., "Truth Maintenance Systems for Problem Solving," MIT MS Thesis, 1977. See "A Truth Maintenance System", MIT AI Memo 521, 1979.

[2] Hayes, P. J., "The Frame Problem and Related Problems in Artificial Intelligence," in A. Elithorn and D. Jones, editors, *Artificial and Human Thinking*, San Francisco: Josey-Bass, 1973.

[3] Hewitt, C. E., "Description and theoretical analysis (using schemata) of PLANNER: a language for proving theorems and manipulating models in a robot," MIT AI Laboratory TR-258, 1972.

[4] Joshi, A. K., and S. J. Rosenschein, "A formalism for relating lexical and pragmatic information: its relevance to recognition and generation," *Proc. Workshop on Theoretical Issues in Natural Language Processing*, Cambridge, Massachusetts, 1975, pp. 79-83.

[5] Kramosil, I., "A Note on Deduction Rules with Negative Premises," *Proc. Fourth IJCAI*, pp. 53-56, 1975.

[6] London, P. E., "Dependency Networks as a Representation for Modelling in General Problem Solvers," Department of Computer Science TR-698, University of Maryland, 1978.

[7] McAllester, D. A., "A Three-Valued Truth Maintenance

System," MIT AI Lab, Memo 473, 1978.

[8] McCarthy, J. and P. J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in B. Meltzer and D. Michie, *Machine Intelligence 4*, New York: American Elsevier 1969, pp. 463-502.

[9] McCarthy, J., "Epistemological Problems of Artificial Intelligence," *Proc. Fifth IJCAI*, pp. 1038-1044, 1977.

[10] Mendelson, E., *Introduction to Mathematical Logic*, New York: Van Nostrand Reinhold, 1964.

[11] Minsky, M., "A Framework for Representing Knowledge," MIT AI Lab, AI Memo 306, 1974.

[12] Quine, W. V., and J. S. Ullian, *The Web of Belief*, second edition, New York: Random House, 1978.

[13] Reiter, R., "On Closed World Data Bases," Department of Computer Science, University of British Columbia, TR 77-14, 1977.

[14] Reiter, R., "On Reasoning by Default," *Proc. Second Symp. on TINLAP*, Urbana, Illinois, 1978.

[15] Rescher, N., *Hypothetical Reasoning*, Amsterdam: North Holland 1964.

[16] Sandewall, E., "An Approach to the Frame Problem, and its Implementation," in B. Meltzer and D. Michie, editors, *Machine Intelligence 7*, New York: John Wiley and Sons, 1972, pp. 195-204.

[17] Stallman, R. M., and C. J. Sussman, "Forward Reasoning and Dependency-Directed Backtracking in a System for Computer-Aided Circuit Analysis," *Artificial Intelligence*, Vol. 9, No. 2, (1977), pp. 135-196.

[18] Thompson, A., "Network Truth Maintenance for Deduction and Modelling," Jet Propulsion Lab, 1979.

[19] Weyhrauch, R. "Prolegomena to a Theory of Formal Reasoning," Stanford AI Lab, Memo AIM-315, 1978.