WHAT'S WRONG WITH NON-MONOTONIC LOGIC?

David J. Israel

Bolt Beranek and Newman Inc.
50 Moulton St.
Cambridge, Mass. 02238

ABSTRACT

In this paper, I ask, and attempt to answer, the following question: What's Wrong with Non-Monotonic Logic? The answer, briefly, is that the motivation behind the wonderfully impressive work involved in its development is based on a confusion of proof-theoretic with epistemological issues.

- - - - - - - - - - - - -

What's wrong with non-monotonic logic (and for that matter, with the logic of default reasoning)?*

The first question we should ask is: What's supposed to be wrong with "standard", monotonic logic? In recent - and extremely impressive - work, Doyle and McDermott [1], McDermott [2], and Reiter [3] have argued that classical logic - in virtue of its monotonicity - is incapable of adequately capturing or representing certain crucial features of real live reasoning and inference. In particular, they note that our knowledge is always incomplete, and is almost always known to be so; that, in pursuing our goals - both practical and theoretical - we are forced to make assumptions or to draw conclusions on the basis of incomplete evidence; conclusions and assumptions which we may have to withdraw in the light of either new evidence or further cogitation on what we already believe. An essential point here is that new evidence or new inference may lead us to reject previously held beliefs, especially those that we knew to be inadequately supported or merely presumptively assumed. In sum, our theories of the world are revisable; and thus our attitudes towards at least some our beliefs must likewise be revisable.

Now what has all this to do with logic and its monotonicity? Both Reiter and Doyle-McDermott characterize the monotonicity of standard logic in syntactic or proof-theoretic terms. If A and B are two theories, and A is a subset of B, then the theorems of A (the set of sentences p such that p is a syntactic consequence of A) is a subset of the theorems of B. (If one keeps in mind that in standard logical theory, a theory is identified with a set of sentences closed under syntactic consequence, that is with its theorems, one can see how unsurprising "syntactic monotonicity" is.) A slightly more interesting property is "semantic monotonicity", that is the monotonicity of the (functional) relation of semantic consequence or entailment. A set of sentences A entails a sentence p iff every model of A is a model of p; that is, iff every interpretation under which every sentence in A is true is an interpretation under which p is true also. If this relation holds for some A and some p, it holds between any set of sentences which properly includes A and that same p. Roughly speaking, embedding A in a larger set cannot alter the fact that if all the sentences in A were true, p would have to be true too. Thus Doyle and McDermott: "Monotonic logics lack the phenomenon of new information leading to a revision of old conclusions". [1]

To remedy this lack, Doyle and McDermott introduce into an otherwise standard first order language a modal operator "M" which, they say, is to be read as "It is consistent with everything that is believed that..." (Reiter's "M", which is not a symbol of the object language, is also supposed to be read "It is consistent to assume that..". I think there is some unclarity on Reiter's part about his "M". He speaks of it in ways conducive to interpreting it as a metalinguistic predicate on sentences of the object language; and hence not as an operator at all, either object-language or metalanguage. So his default rules are expressed in a language whose object-language contains sentences of the form "Mp", i.e., in a language which, relative to the original first-order object language, is a meta-meta-language.) Now in fact this reading isn't quite right.** The suggested reading doesn't capture the notion Doyle-McDermott and Reiter seem to have in mind. What they have in mind is, to put it non-linguistically (and hence, of course, non-syntactically): that property that a belief has just in case it is both compatible with everything a given subject believes at a given time and remains so when the subject's belief set undergoes certain kinds of changes under the pressure of both new information and further thought, and where those changes are the result of rational epistemic policies.

*The research reported in this paper was supported in part by the Advanced Research Projects Agency, and was monitored by ONR under Contract No. N00014-77-C-0378.

I've put the notion in this very epistemologically oriented way precisely to hone in on what I take to be the basic misconception underlying the work on non-monotonic logic and the logic of default reasoning. The researchers in question seem to believe that logic - deductive logic, for there is no other kind - is centrally and crucially involved in the fixation and revision of belief. Or to put it more poignantly, they mistake so-called deductive rules of inference for real, honest-to-goodness rules of inference. Real rules of inference are precisely rules of belief fixation and revision; deductive rules of transformation are precisely not. Consider that old favorite: modus (ponendo) ponens. It is not a rule that should be understood as enjoining us as follows: whenever you believe that p and believe that if p then q, then believe that q. This, after all, is one lousy policy. What if you have overwhelmingly good reasons for rejecting the belief that q? All logic tells you is that you had best reconsider your belief that p and/or your belief that if p then q (or, to be fair, your previously settled beliefs on the basis of which you were convinced that not-q); it is perforce silent on how to revise your set of beliefs so as to .. to what? Surely, to come up with a good theory that fits the evidence, is coherent, simple, of general applicability, reliable, fruitful of further testable hypotheses, etc. Nor is it the case that if one is justified in believing that p and justified in believing that if p then q (or even justified in believing that p entails q), is one justified in believing (inferring) that q. Unless, of course, one has no other relevant beliefs. But one always does.

The rule of modus ponens is, first and foremost, a rule that permits one to perform certain kinds of syntactical transformations on (sets of) formally characterized syntactic entities. (Actually, first and foremost, it is not really a rule at all; it is "really" just a two-place relation between on the one hand an ordered pair of wffs., and on the other, a wff.) It is an important fact about it that, relative to any one of a family of interpretations of the conditional, the rule is provably sound, that is truth (in an interpretation)-preserving. The crucial point here, though, is that adherence to a set of deductive rules of transformation is not a sufficient condition for rational belief; it is sufficient (and necessary) only for producing derivations in some formal system or other. Real rules of inference are rules (better: policies) guiding belief fixation and revision. Indeed, if one is sufficiently simple-minded, one can even substitute for the phrase "good rules of inference", the phrase "(rules of) scientific procedure" or even "scientific method". And, of course, there is no clear sense to the phrase "good rules of transformation". (Unless "good" here means "complete" - but with respect to what? Truth?)

Given this conception of the problem to which Doyle-McDermott and Reiter are addressing themselves, certain of the strange properties of, on the one hand, non-monotonic logic and on the other, the logic of default reasoning, are only to be expected. In particular, the fact that the proof relation is not in general decidable. The way the "M" operator is understood, we believers are represented as follows: to make an assumption that p or to put forth a presumption that p is to believe a proposition to the effect that p is consistent with everything that is presently believed and that it will remain so even as my beliefs undergo certain kinds of revisions. And in general we can prove that p only if we can prove at least that p is consistent with everything we now believe. But, of course, by Church's theorem there is no uniform decision procedure for settling the question of the consistency of a set of first-order formulae. (Never mind that the problem of determining the consistency of arbitrary sets of formulae of the sentential calculus is NP-complete.) This is surely wrong-headed: assumptions or hypotheses or presumptions are not propositions we accept only after deciding that they are compatible with everything else we believe, not to speak of having to establish that they won't be discredited by future evidence or further reasoning. When we assume p, it is just p that we assume, not some complicated proposition about the semantic relations in which it stands to all our other beliefs, and certainly not some complicated belief about the syntactic relations any one of its linguistic expressions has to the sentences which express all those other beliefs. (Indeed, there is a problem with respect to the consistency requirement, especially if we allow beliefs about beliefs. Surely, any rational subject will believe that s/he has some false beliefs, or more to the point, any such subject will be disposed to accept that belief upon reflection. By doing so, however, the subject guarantees itself an inconsistent belief-set; there is no possible interpretation under which all of its beliefs are true. Should this fact by itself worry it (or us?).)

After Reiter has proved that the problem of determining whether an arbitrary sentence is in an extension for a given default theory is undecidable, he comments:

**Nor is it quite clear. By "consistent" are we to mean syntactically consistent in the standard monotonic sense of syntactic derivability or in the to-be-explicated non-monotonic sense? Or is it semantic consistency of one brand or another that is in question? This unclarity is fairly quickly remedied. We are to understand by "consistency" standard syntactic consistency, which in standard systems can be understood either as follows: A theory is syntactically consistent iff there is no formula p of its language such that both p and its negation are theorems, or as follows: iff there is at least one sentence of its language which is not a theorem. There are otherwise standard, that is, monotonic, systems for which the equivalence of these two notions does not hold; and note that the first applies only to a theory whose language includes a negation operator.

> (A)ny proof theory whatever for... default theories must somehow appeal to some inherently non semi-decidable process. [That is, the proof-relation, not just the proof predicate, is non recursive; the proofs, not just the theorems, are not recursively enumerable. Why such a beast is to be called a logic is somewhat beyond me - DI.] This extremely pessimistic result forces the conclusion that any computational treatment of defaults must necessarily have an heuristic component and will, on occasion, lead to mistaken beliefs. Given the faulty nature of human common sense reasoning, this is perhaps the best one could hope for in any event.

Now once again substitute in the above "(scientific or common sense) reasoning" for "default(s)" and then reflect on how odd it is to think that there could be a purely proof-theoretic treatment of scientific reasoning. A heuristic treatment, that is a treatment in terms of rational epistemic policies, is not just the best we could hope for. It is the only thing that makes sense. (Of course, if we are very fortunate, we may be able to develop a "syntactic" encoding of these policies; but we certainly mustn't expect to come up with rules for rational belief fixation that are actually provably truth-preserving. Once again, the only thing that makes sense is to hope to formulate a set of rules which, from within our current theory of the world and of ourselves as both objects within and inquirers about that world, can be argued to embody rational policies for extending our admittedly imperfect grasp of things.)

Inference (reasoning) is non-monotonic: New information (evidence) and further reasoning on old beliefs (including, but by no means limited to, reasoning about the semantic relationships - e.g., of entailment - among beliefs) can and does lead to the revision of our theories and, of course, to revision by "subtraction" as well as by "addition". Entailment and derivability are monotonic. That is, logic - the logic we have, know, and - if we understand its place in the scheme of things - have every reason to love, is monotonic.

BRIEF POSTSCRIPT

I've been told that the tone of this paper is overly critical; or rather, that it lacks constructive content. A brief postscript is not the appropriate locus for correcting this defect; but it may be an appropriate place for casting my vote for a suggestion made by John McCarthy. In his "Epistemological Problems of Artificial Intelligence" [4], McCarthy characterizes the epistemological part of "the AI problem" as follows: "(it) studies what kinds of facts about the world are available to an observer with given opportunities to observe, how these facts can be represented in the memory of a computer, and what rules permit legitimate conclusions to be drawn from these facts." [Emphasis added.] This, though brief, is just about right, except for a perhaps studied ambiguity in that final clause. Are the conclusions legitimate because they are entailed by the facts? (Are the rules provably sound rules of transformation?) Or are the conclusions legitimate because they constitute essential (non-redundant) parts of the best of the competing explanatory accounts of the original data; the best by our own, no doubt somewhat dim, lights? (Are the rules arguably rules of rational acceptance?) At the conclusion of his paper, McCarthy disambiguates and opts for the right reading. In the context of an imaginative discussion of the Game of Life cellular automaton, he notes that "the program in such a computer could study the physics of its world by making theories and experiments to test them and might eventually come up with the theory that its fundamental physics is that of the Life cellular automaton. We can test our theories of epistemology and common sense reasoning by asking if they would permit the Life-world computer to conclude, on the basis of its experiments, that its physics was that of Life." McCarthy continues:

> More generally, we can imagine a metaphilosophy that has the same relation to philosophy that metamathematics has to mathematics. Metaphilosophy would study mathematical (? - D.I.) systems consisting of an "epistemologist" seeking knowledge in accordance with the epistemology to be tested and interacting with a "world". It would study what information about the world a given philosophy would obtain. This would depend also on the structure of the world and the "epistemologist's" opportunities to interact. AI could benefit from building some very simple systems of this kind, and so might philosophy.

Amen; but might I note that such a metaphilosophy does exist. Do some substituting again: for "philosophy" (except in its last occurrence), substitute "science"; for "epistemologist", "scientist"; for "epistemology", either "philosophy of science" or "scientific methodology". The moral is, I hope, clear. Here is my constructive proposal: AI researchers interested in "the epistemological problem" should look, neither to formal semantics nor to proof-theory; but to - of all things - the philosophy of science and epistemology.

REFERENCES

[1]McDermott, D., Doyle, J. "Non-Monotonic Logic I", AI Memo 486, MIT Artificial Intelligence Laboratory, Cambridge, Mass., August 1978.

[2]McDermott, D. "Non-Monotonic Logic II", Research Report 174, Yale University Department of Computer Science, New Haven, Conn., February 1980.

[3]Reiter, R. "A Logic for Default Reasoning", Technical Report 79-8, University of British Columbia Department of Computer Science, Vancouver, B.C., July 1979.

[4]McCarthy, J. "Epistemological Problems of Artificial Intelligence", In Proc. IJCAI-77. Cambridge, Mass., August, 1977, pp. 1038-1044.