Reset reproduction of article published in *Computational Intelligence*, Vol. 3, No. 3 (August 1987), pp. 175-176. Written August 28, 1986. Reprinted July 1994. Current address: Massachusetts Institute of Technology, Laboratory for Computer Science, Cambridge, Massachusetts

Logic, Rationality, and Rational Psychology
A commentary on Drew McDermott's *Critique of Pure Reason*

Jon Doyle
Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213

McDermott has written a fine critique of the logicist views that most reasoning is deductive and that logic is the proper theory of thinking, and I largely agree with his observations. But he laments abandoning the attractions of the logicist method along with its flaws. I believe this disappointment is unnecessary if the flaws and attractions are separated as follows.

The underlying flaw in the logicist view is the misperception that reasoning is an aspect of the agent's structure rather than one of its activities. (See [Harman 1986] and §1.1 of [Doyle 1979].) When inferences are viewed as filling in part of a Platonic logical structure, it is natural to hope them deductive. When inferences are actions carrying the agent from initial states of mind and sets of attitudes to new states and attitudes, it is instead natural to hope them rational—in particular, rational according to the agent's own attitudes at the time—or intentional, according to rationally adopted plans for reasoning. But as sketched below, rational inferences are typically non-deductive.

Everyone in AI knows that most possible enlargements of the set of beliefs with conclusions that follow deductively are senseless—uninformative or unhelpful wastes of time and effort—and that designers of automated deduction systems work hard to minimize the number of patently useless steps they take. But the value of possible inferences from axioms cannot be judged purely on the basis of the axioms themselves. These judgments also require expectations about the likely consequences of possible inferences, and the relative merits of these consequences. To think that most reasoning is deductive is to think that beliefs alone largely determine reasoning actions. But this isolation of action from value is not possible. Starting with the same set of beliefs, two agents might be either intelligent or stupid, competent or comatose, depending on their motivations.

Even rational deductive inferences play only a small role in reasoning. Rational and intentional changes of assumptions play much larger, though often unacknowledged, roles. For example, McDermott views non-monotonic logic and reason maintenance as logicist enterprises, but properly formulated they are better viewed as limited theories of rational and intentional selections of attitudes. As [Doyle 1980, 1983a, 1985, 1986] first suggest and then show, by separating the notions of belief revision and reasoned assumptions from each other and from the irrelevant (in this context) notion of logical consistency, it is readily seen that (A) defaults are preferences in favor of making specific assumptions, (B) reasons (or justifications) are conditional intentions to draw specific conclusions, and (C) both are eminently computable, in contrast to the uncomputability of non-monotonic logic that McDermott deplores. These are merely two simple cases of rational but non-deductive reasoning, closely connected with Pascal's wager and James' will to believe. Other examples involving revisions of belief include rational selection of sources of inconsistencies, conserva-

tive accommodations to new information, and rational selection among alternative theories in learning. (See especially [Doyle 1980, 1986] for details.)

If reasoning is an activity, knowledge is best thought of as the grounds of reasoning: not merely the agent's factual information (its beliefs and subjective probabilities, or "content theory"), but its evaluative and procedural information (its preferences and plans, desires and intentions, or "process model") as well. But the logicist aim of writing down knowledge before implementing it still applies here. McDermott is too pessimistic in thinking that one can only start programming immediately. Put crudely, if Lisp data-structures can represent or encode beliefs typically formulated and expressed more clearly in natural or logical languages, Lisp procedures for efficiency's sake typically represent or encode preferences and plans very opaquely compared to the explanations or specifications of program behavior supplied by the programmer. Indeed, analysis of programs to find their underlying plans and construction of programs from plans are central to current work in automatic programming. The aim of formalization or specification prior to implementation is the enduring attraction of the logicist method, but should apply to all components of knowledge, factual, evaluative, and procedural. As with beliefs, the agent's preferences and plans need not be represented explicitly anywhere in the implementation program. (HACKER's [Sussman 1975] self-programming offers an excellent example of this.) Their method of embodiment is a choice of the programmer, influenced by considerations of convenience and computational efficiency.

However, formalization of intended behaviors crucially depends on a good vocabulary of formal concepts for describing the behaviors and structures of possible psychological organizations, and at present, most known formal concepts are apparently inadequate, while many seemingly interesting concepts remain informal. The mathematical search for good vocabularies is the enterprise of *rational psychology* [Doyle 1983b]. Here the term "rational" refers not to any rationality of the agents under study, but instead to the method of investigation, that of finding the best formal concepts through mathematical analysis. In this investigation, one considers all aspects of psychologies—not just beliefs—and one does not make the restrictive assumption (as logicists mistakenly do) that the appropriate vocabulary is purely logical (or purely decision-theoretic, neurological, behavioral, or Freudian, for that matter).

Reasoning is an activity. Logic and deduction may offer some concepts useful for formalizing and specifying the internal structure of this activity, but certainly are inadequate by themselves. Similarly, programming may offer a way of constructing reasoning agents, but programming alone is certainly inadequate for understanding what is to be implemented. To understand and design interesting psychological agents we must find the proper concepts for formulating psychological ideas—and that means continued effort in rational psychology.

References

Doyle, J., 1979. A truth maintenance system, *Artificial Intelligence* **12**, 231-272.

Doyle, J., 1980. A model for deliberation, action, and introspection, Cambridge: MIT Artificial Intelligence Laboratory, TR-581.

Doyle, J., 1983a. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Carnegie-Mellon University, Department of Computer Science, report 83-125.

Doyle, J., 1983b. What is Rational Psychology? Toward a modern mental philosophy, *AI Magazine*, V. 4, No. 3, 50-53.

Doyle, J., 1985. Reasoned assumptions and Pareto optimality, *Ninth International Joint Conference on Artificial Intelligence* 87-90.

Doyle, J., 1986. Topics in rational self-government, Pittsburgh: Carnegie-Mellon University, Department of Computer Science, in preparation.

Harman, G., 1986. *Change of View: Principles of Reasoning*, Cambridge: MIT Press.

Sussman, G. J., 1975. *A Computer Model of Skill Acquisition*, New York: American Elsevier.