Fundamenta Informaticae V. 20, Nos. 1–3 (Spring 1994), pp. 35–73 Submitted 11/93

Reasoned Assumptions and Rational Psychology

Jon Doyle DOYLE@LCS.MIT.EDU
Massachusetts Institute of Technology, Laboratory for Computer Science
Cambridge, Massachusetts 02139, U.S.A.

Abstract

Logical epistemology unduly sways theories of thinking that formulate problems of nonmonotonic reasoning as issues of nondeductive operations on logically phrased beliefs, because the fundamental concepts underlying such reasoning have little to do with logic or belief. These formulations make the resulting theories inappropriately special and hide the characteristic structures of nonmonotonic reasoning amid many unrelated structures. We present a more direct mathematical development of nonmonotonic reasoning free of extraneous logical and epistemological assumptions, and argue that the insights gained in this way exemplify the benefits obtained by approaching psychology as a subject for mathematical investigation through the discipline of *rational psychology*.

For Joseph A. Schatz, teacher and friend

1 Reasoning, logic, and psychology

Nonmonotonic reasoning, the study of making and revising assumptions in a reasoned or principled way, needs little introduction in artificial intelligence today thanks to years of extensive exposition, analysis, and application. In spite of an admirable history of progress, however, the subject stands in need of some rethinking and redirection as the strengths and limitations of the accepted theories become clearer. This paper seeks to further this rethinking and redirection by presenting the foundations of nonmonotonic reasoning through a mathematical and philosophical approach closer to the concepts and methods of modern physics and rational mechanics than to the standard formulations of artificial intelligence. I believe these concepts and methods, which seek to find the most appropriate means for describing and understanding psychological structure and behavior, will prove productive for rethinking other parts of artificial intelligence as well. This introduction thus attempts to set out some of the motivations for this rethinking and to motivate the methods underlying the formal treatment.

This paper celebrates the twentieth volume of *Fundamenta Informaticae*. The year of its writing (1993) also marks the twentieth anniversary of my involvement in the field of artificial intelligence; the fifteenth anniversary of the appearance of the original nonmonotonic logic (McDermott & Doyle, 1980); and the tenth and fifth anniversaries (respectively) of the appearance of my mathematical monograph (Doyle, 1983c) and my foundational monograph (Doyle, 1988), from which the present paper derives and upon which it improves, and looking back from these anniversaries has led me to include some personal interpretations of their history in this introduction.

1.1 Nonmonotonic reasoning

Though the reader will likely find the notion familiar, a few words about the term "nonmonotonic reasoning" should aid in understanding the discussion to follow.

Intuitively speaking, nonmonotonic reasoning refers to reasoning involving nonadditive changes in beliefs, preferences, intentions, and other mental attitudes. The intuitive notion, however, is meaningless on its own because reasoning is an activity, and activities are not inherently monotonic or nonmonotonic; any monotonicity and nonmonotonicity of reasoning must be relative to how we view the reasoning in terms of aspects of mental states. In the usual usage in theoretical artificial intelligence, one views mental states as consisting of sets of mental attitudes and reasoning as a process that fills out and changes these sets over time. One can thus identify two very different senses of nonmonotonicity of reasoning: *temporal* nonmonotonicity, in which mental attitudes may appear and vanish over time, and *logical* nonmonotonicity, in which filling out larger sets of attitudes may yield fewer conclusions than filling out smaller sets. Mathematically, temporal nonmonotonicity compares mental attitudes as time increases, while logical nonmonotonicity compares consequences as mental attitudes increase.

Temporal nonmonotonicity may occur routinely and unexceptionally, for example through direct temporal variation of mental attitudes by perceptual or cognitive systems that add and subtract attitudes to reflect changes or anticipated changes in the world (as might happen if some of the attitudes describe the contents of the retina). Logical nonmonotonicity may occur because the reasoner derives some attitudes as conclusions from others as long as the right circumstances obtain. In the canonical example, the reasoner infers that Tweety flies from the information that Tweety is a bird, but not from the information that Tweety is also a penguin, information that defeats or undercuts the usual conclusion. In general, however, the division between these two forms of nonmonotonicity is not sharp, as one may draw conclusions over time to convert logical nonmonotonicity into temporal nonmonotonicity, or replace mental simulations with atemporal reasoning or logics to convert temporal nonmonotonicity into logical nonmonotonicity (cf. (Makinson & Gärdenfors, 1991)).

Theorists and practitioners in artificial intelligence recognized the need for logically nonmonotonic reasoning early on, motivated by problems of reasoning about knowledge and actions, by the desire to make plausible commonsense inferences, and by the desire to speed problem-solving searches by making quick decisions about where to search that would yield information useful for guiding the search even if proven wrong. They suggesting ways of expressing nonmonotonic reasoning rules (e.g., (McCarthy & Hayes, 1969; Sandewall, 1972)) and implementing reasoning systems that performed versions of these (e.g., (Sussman, Winograd, & Charniak, 1971)), but rigorous and formal theories appeared later, for unlike ordinary logic, in which one takes contradictions to indicate flawed axioms, useful commonplace rules of nonmonotonic reasoning can provide conflicting conclusions in some cases, conflicts that call for adjudication, perhaps case by case, rather than for abandonment or revision of the conflicting rules. In another canonical example, the reasoner infers that Nixon is a pacifist because Nixon is a Quaker, but also infers Nixon is not a pacifist because Nixon is a (US) Republican, and has to decide which of these reasonable conclusions to accept while keeping the rules that led to them. The early proposals offered no precise ways of treating such conflicts, even when anticipated, as identifying coherent notions of nonmonotonic conclusions proved a perplexing task.

I formulated perhaps the first rigorous solution to this problem in 1976 as the two fundamental principles of my original reason maintenance system or RMS (Doyle, 1976, 1979) (renamed so from "truth maintenance system" or TMS in (Doyle, 1980)), which introduced the now-familiar notion of nonmonotonic justification. (Some may also consider McCarthy's (1977) probably contemporaneous early notion of circumscription a solution to this problem, or even credit the older logical theory of implicit definition (Doyle, 1985).) The RMS represents mental attitudes (or other representational or procedural items) by structures called *nodes* that the RMS labels as either *in* or *out* (of the current state). The RMS also records sets of *justifications* or *reasons* for each node, most of which express simple boolean combinations of the labelings of nodes we denote as "$A \mathbin{\backslash\!\backslash} B \Vdash c$" and read as "$A$ without B gives c", meaning that the node c should be *in* if each node in the set A is *in* and each node in the set B is *out*. The RMS then seeks to construct labelings for the nodes from these justifications, labelings that satisfy two principles: a "stability" principle of labeling each node *in* if and only if one of its reasons is *valid* in the labeling (i.e., expresses hypotheses "A without B" that match the labeling), and a "groundedness" principle demanding that labelings provide each node labeled *in* with a noncircular argument in terms of valid reasons. The structure for justifications given above makes both of these principles perfectly unambiguous. Indeed, these principles convert nonmonotonic reasoning tasks into problems for algorithmic analysis, and different versions of RMS explored different graph-theoretic techniques for analyzing systems of nodes and justifications.

1.2 Logical formalizations

The fundamental RMS principles led, in time, to a variety of formalizations of the stability and groundedness notions. The initial and most abidingly popular formalizations clothed these principles in logical garb: nonmonotonic logic (McDermott & Doyle, 1980) and the logic of defaults (Reiter, 1980), together with the circumscription rule of inference (McCarthy, 1977, 1980), which as a class of inference operators rather than a "logic" has a somewhat different character from nonmonotonic and default logics. Each of these theories formalizes nonmonotonic reasoning by encoding groundedness and the presence and absence of knowledge in terms of logical provability and unprovability, or in terms of logical consistency instead of provability. For example, the simplest transcription of the canonical example into nonmonotonic logic translates the inference as the implication $b \wedge \neg L \neg f \rightarrow f$, where L stands for the provability modality, b stands for Tweety's birdness, and f for Tweety's flying. A similar transcription of RMS justifications translates them into the form $a_1 \wedge \ldots \wedge a_m \wedge \neg L \neg b_1 \wedge \ldots \wedge \neg L \neg b_n \rightarrow c$.

Formalizations, as mathematical characterizations of ideas, may be either good or bad characterizations. The initial formalizations of nonmonotonic reasoning, along with their monotonously logical subsequent variants, improvements, and extensions, proved very good as ways of making the theoretical problems of logically nonmonotonic reasoning both interesting and accessible to a wider audience than artificial intelligence theoreticians. This accessibility and advertisement encouraged the involvement of many brilliant thinkers in the problems of artificial intelligence, and produced a large and vigorous literature that has significantly increased our understanding of nonmonotonic reasoning. It seems doubtful we would understand as much today had these formalizations not been developed and explored as they have.

At the same time, the logical formalizations proved very bad as conceptual characterizations of nonmonotonic reasoning because the fundamental concepts of nonmonotonic reasoning have little to do with the concepts of logic, which in these formalizations obscure and mislead one from attending to the concepts of interest. Understand that this does *not* mean the logical formalizations have no value, even as conceptual characterizations, for a bad formalization may serve well enough. But the logical formalisms deserve reconsideration because they highlight and enshrine at their core things essentially unrelated to nonmonotonic reasoning. We enumerate only three of the most important inappropriate aspects of the logical formalizations.

In the first place, the logical formalizations convert what in many systems is a fast and computationally trivial check for presence and absence of attitudes into a computationally difficult or impossible check for provability, unprovability, consistency or inconsistency. This inaptness seems especially galling in light of the initial problem-solving motivations for nonmonotonic assumptions, for which assumptions served to speed inference, not to slow it (cf. (de Kleer, Doyle, Steele, & Sussman, 1977; Ginsberg, 1991)).

In the second place, the logical formalizations impede development of realistic approaches to reasoning about inconsistent information. Standard logics make inconsistent theories trivial, and the use of logical provability or consistency to encode inferential permissions and guidance means that nonmonotonic theories must be consistent as well in order to be useful. But in practice, reasoners must deal with inconsistent information all the time. They may try to remove some inconsistencies when they deem the inconsistencies important enough to warrant the effort, but even so may take time to remove them, occasionally even a long time, and must keep operating reasonably throughout this process. A theory of reasoning basing the very definition of mental state on logical consistency requirements makes representation of such processes impossible.

In the third place, the logical formalizations encourage the view that nonmonotonic rules and defaults express beliefs or factual information, even though reasoning, and nonmonotonic reasoning in particular, may involve desires, intentions, and other mental attitudes besides belief (cf. (de Kleer et al., 1977; Doyle, 1980; McCarthy & Hayes, 1969; McDermott, 1978)). Now no one disputes that some nonmonotonic rules carry or presume some beliefs, but taking this special case as the general one has proven very misleading. For example, some theories attempting to provide guidance about choosing from among the possible interpretations of complex sets of nonmonotonic rules interpret these rules as qualitative statements about highly likely conditions. While high conditional probabilities of the conclusions may motivate adoption of some defaults and thus deserve attention as a special case, this interpretation does not even make sense for rules reasoning with other mental attitudes, and might indicate the wrong conclusions for all we know now. But such theories have been proposed as theories of nonmonotonic reasoning in general, not as theories of special cases. Nonmonotonic theories based on logical consistency provide another example with the same character. These theories convert reasoning about reasoning generally into reasoning about the consistency of beliefs. Now while one may motivate a number of reasonable principles for guiding reasoning about beliefs in terms of avoiding inconsistency, these principles simply do not apply to reasoning about desires and some other attitudes. The conclusions thus motivated hardly seem appropriate for general use without much further argument.

It is tempting to point the finger of blame for these mistakes elsewhere, even back to McCarthy and Hayes' (1969) original brief suggestion, but I bear much responsibility for these misleading formalizations due the initial nonmonotonic logic I developed with Drew McDermott in 1978 (McDermott & Doyle, 1980). I well understood the distinction between the underlying RMS ideas and the logical encoding, and was quite pleased when Reiter's (1980) default logic later remedied a number of the unsatisfying characteristics of our original nonmonotonic logic through an approach closer to the RMS. Nevertheless, I was for years as encouraged as anyone in thinking that some logic would eventually provide the right formulation for these ideas—at least until I undertook in 1981 to develop logics capturing the RMS approach even more precisely. I soon found a great variety of possible schemes of reasoning, which suggested that the fundamental ideas were best isolated and pursued independently rather than attempting to formulate each variant as a special quirky logic. Mathematical classification seemed more appropriate than philosophical or metaphysical proposal, especially as many of the dozens (if not hundreds) of the schemes imaginable seemed well suited to some special purpose, rather than some logic dominating the rest.

1.3 Rational psychology

Eventually it became clear that in spite of the great progress made through exploiting logical tools and theories in artificial intelligence, one cannot expect *a priori* that the most appropriate theory of psychology should have much to do with logic or should necessarily make extensive use of logical concepts. Psychology, as a subject for investigation, includes the study of many aspects of thought, feeling, and behavior, while standard logic mainly idealizes only part of one sort of mental activity, reasoning about facts or beliefs, and clearly does not address notions like intent, desire, and preference, nor even the relations between such attitudes and reasoning in terms of beliefs (to say nothing about its silence on love, fear, and other feelings). This is not to say one cannot cast parts of these larger theories in logical terms too; this enterprise goes back to Aristotle, and has been investigated diligently in modern philosophical logic to good effect. But even here, one finds no reason *a priori* to suppose that ordinary logic provides the most appropriate basis for these investigations. The beautiful theories of modern mathematical logic constitute a triumph of conceptual analysis; but while logic is great, it isn't psychology, and to use G. A. Miller's (1986) vivid metaphor, instead of carving up the subject of psychology at its joints and clearly revealing the structure of psychology's conceptual components, standard uses of logic may in fact dismember psychology, carving it up into ill-shapen lumps that reveal little of the attraction of the subject.

The rational mechanics of Newton and its modern renewal by Truesdell (1958, 1977) and his contemporaries provide a model for a more productive approach. Rational mechanics is a part of mathematics, the conceptual investigation of mechanics. "Rational" here indicates investigations based on reason alone, rather than on experiment, engineering, or computation, the rational analysis of the concepts and theories whose applicability and feasibility are studied in experimental, engineering, and computational projects. We call the corresponding part of mathematics devoted to the study of psychology *rational psychology* (Doyle, 1983d) (a term used by James (1892) and earlier writers to refer to philosophical psychology). Rational psychology is not the study of rational agents, but instead the mathematical approach to the problems of agents and their actions, whether one thinks these agents and actions rational or irrational. It aims to understand psychological ideas through mathematical classification of all possible minds or psychological systems, to describe and study mental organizations and phenomena by the most fit mathematical concepts, seeking the best (most appropriate, illuminating, edifying, powerful, ...) ways of formalizing psychology. It excludes the problem of identifying important psychological phenomena except as a byproduct of organizing ideas about psychologies into a coherent mathematical whole.

The mathematical formulation of nonmonotonic reasoning presented below resulted from this shift in focus from logic to conceptual analysis, with many of the central concepts and results of the new formulation first circulated in my monograph (Doyle, 1983c), a dense exposition short on explanation that revealed the mathematical structure of the subject but doubtless exceeded the tolerance of most potential readers. That monograph introduced the notion of *simple reason*, which resembles a RMS justification or a propositional default in the logic of defaults, but which appears as a conclusion as well as as a rule, and which so also resembles one of Minsky's "K-lines" (Minsky, 1980). Simple reasons formed the focus of the original exposition, which deliberately presented the core of the development *twice*—once for simple reasons alone (no mention of any logical structure), and then again for states closed with respect to a compact abstract deducibility relation—to drive home the irrelevance of even minimal logical notions to the central mathematical structures of reasoned assumptions. That presentation also deviated from logic in embracing a wide conception of semantics and meaning as the theory of pure designation, without requiring the compositionality of meanings usually demanded by logical theories. It also hewed close to the RMS conception in making no assumptions that reasoned entities represented beliefs, or that presence or absence amounted to consistency or inconsistency, or that contrary beliefs should not be held indefinitely, or that any mental entities were inherently contradictory with any others. At the same time, it followed logic in aiming to describe reasoning agents regardless of computability, in seeking to describe the mind of God as well as the mind of Man. The present exposition also draws on my later monograph (Doyle, 1988), which provided better development of the motivation, better forms for some of the concepts, and better notation.

This paper omits treatment of many important topics addressed in (Doyle, 1983c) and (Doyle, 1988), in their numerous sequelae, and to some extent in (Doyle, 1980). We have omitted most discussion of the motivations for nonmonotonic reasoning in particular and the basic structures of mental architectures in general; of the nature of meaning and the structure of semantical theories; of "psycho-logics" derived from the state spaces of agents; of uniformly defeasible reasons; of the theory of denials and contradictions; of logical encodings of reasoning; of psychological attitudes; of conservatism and other topics concerning the evolution of mental states; of probabilistic constructions over trajectories and their relation to reasoning and the strength of mental attitudes; of the failings of practical systems like the RMS; of social, economic, and political structures within minds; of most connections with ideas from economics; and of reflection, deliberation, and action. The mathematical formulations of these topics presented in the earlier works sometimes need emendation, but generally fit well into the framework elaborated here.

This paper seeks to present the fundamental concepts and results about nonmonotonic reasoning in a setting free of unnecessary logical ornament. Section 2 introduces the notion of framings as different ways of viewing mental states, while Section 3 presents the underlying tools for describing the constitution or special structure of mental states, including the notion of satisfaction system used to express half of the RMS stability principle. Section 4 introduces simple reasons in semantic terms, as elements of mental states bearing certain constitutive meanings rather than as exhibiting specific structures, and develops canonical descriptions for them. Section 5 introduces constitutions for reasoning that express the other half of the RMS stability principle as well as the RMS groundedness principle, while Section 6 analyzes some of the structure of reasoned states.

References

Abelson, H. (1978). Towards a theory of local and global in computation. *Theoretical Computer Science*, *6*, 41–67.

de Kleer, J., Doyle, J., Steele, G. L. Jr., & Sussman, G. J. (1977). Amord: Explicit control of reasoning. In *Proceedings of the ACM Symposium on Artificial Intelligence and Programming Languages*, pp. 116–125.

Doyle, J. (1976). The use of dependency relationships in the control of reasoning. Tech. rep. Working Paper 133, MIT AI Laboratory.

Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, *12*(2), 231–272.

Doyle, J. (1980). A model for deliberation, action, and introspection. Ai-tr 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139.

Doyle, J. (1983a). Admissible state semantics for representational systems. *IEEE Computer*, *16*(10), 119–123.

Doyle, J. (1983b). A society of mind: Multiple perspectives, reasoned assumptions, and virtual copies. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 309–314.

Doyle, J. (1983c). Some theories of reasoned assumptions: An essay in rational psychology. Tech. rep. 83-125, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Doyle, J. (1983d). What is rational psychology? toward a modern mental philosophy. *AI Magazine*, *4*(3), 50–53.

Doyle, J. (1985). Circumscription and implicit definability. *Journal of Automated Reasoning*, *1*, 391–405.

Doyle, J. (1988). Artificial intelligence and rational self-government. Tech. rep. CS-88-124, Carnegie-Mellon University Computer Science Department.

Doyle, J. (1989). Constructive belief and rational representation. *Computational Intelligence*, *5*(1), 1–11.

Doyle, J. (1992). Reason maintenance and belief revision: Foundations vs. coherence theories. In Gärdenfors, P. (Ed.), *Belief Revision*, pp. 29–51. Cambridge University Press, Cambridge.

Doyle, J., & Wellman, M. P. (1990). Rational distributed reason maintenance for planning and replanning of large-scale activities. In Sycara, K. (Ed.), *Proceedings of the DARPA Workshop on Planning and Scheduling*, pp. 28–36 San Mateo, CA. Morgan Kaufmann.

Doyle, J., & Wellman, M. P. (1991). Impediments to universal preference-based default theories. *Artificial Intelligence*, *49*(1-3), 97–128.

Fahlman, S. E. (1979). *NETL: A System for Representing and Using Real-World Knowledge*. The MIT Press, Cambridge, MA.

Gärdenfors, P. (1990). The dynamics of belief systems: Foundations vs. coherence theories. *Revue Internationale de Philosophie*, *172*, 24–46.

Gazzaniga, M. S. (1985). *The Social Brain: Discovering the Networks of the Mind*. Basic Books, New York.

Ginsberg, M. L. (1991). The computational value of nonmonotonic reasoning. In Allen, J., Fikes, R., & Sandewall, E. (Eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pp. 262–268 San Mateo, CA. Morgan Kaufmann.

Harman, G. (1986). *Change in View: Principles of Reasoning*. MIT Press, Cambridge, MA.

James, W. (1892). *Psychology*. Henry Holt & Co., New York.

Konolige, K. (1985). Belief and incompleteness. In Hobbs, J. R., & Moore, R. C. (Eds.), *Formal Theories of the Common-Sense World*, pp. 359–403. Ablex, Norwood.

Makinson, D., & Gärdenfors, P. (1991). Relations between the logic of theory change and the nonmonotonic logic. In Fuhrmann, A., & Morreau, M. (Eds.), *The Logic of Theory Change*, pp. 185–205. Springer-Verlag, Berlin.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B., & Michie, D. (Eds.), *Machine Intelligence 4*, pp. 463–502. Edinburgh University Press.

McCarthy, J. (1977). Epistemological problems of artificial intelligence. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pp. 1038–1044.

McCarthy, J. (1980). Circumscription — a form of non-monotonic reasoning. *Artificial Intelligence*, *13*(1), 27–38.

McDermott, D. (1978). Planning and acting. *Cognitive Science*, *2*, 71–109.

McDermott, D., & Doyle, J. (1980). Non-monotonic logic—I. *Artificial Intelligence*, *13*, 41–72.

Miller, G. A. (1986). Dismembering cognition. In Hulse, S. H., & Green, Jr., B. F. (Eds.), *One Hundred Years of Psychological Research in America*, pp. 277–298. Johns Hopkins University Press, Baltimore.

Minsky, M. (1965). Matter, mind, and models. In *Proceedings of the IFIP Congress*, pp. 45–49.

Minsky, M. (1980). K-lines: a theory of memory. *Cognitive Science*, *4*, 117–133.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*, 81–132.

Sandewall, E. (1972). An approach to the frame problem, and its implementation. In *Machine Intelligence 7*, pp. 195–204. University of Edinburgh Press.

Scott, D. S. (1982). Domains for denotational semantics. In Nielsen, M., & Schmidt, E. M. (Eds.), *Automata, Languages, and Programming: Ninth Colloquium*, Vol. 140 of *Lecture Notes in Computer Science*, pp. 577–613 Berlin. Springer-Verlag.

Sussman, G. J., Winograd, T., & Charniak, E. (1971). Micro-planner reference manual, (revised). Aim 203A, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139.

Touretzky, D. S. (1986). *The Mathematics of Inheritance Systems*. Morgan Kaufman, Los Altos, CA.

Truesdell, C. (1958). Recent advances in rational mechanics. *Science*, *127*, 729–739.

Truesdell, C. (1977). *A First Course in Rational Continuum Mechanics*, Vol. 1. Academic Press, New York.

Contents