Reasoning about Rational Agents

A Review

Shlomo Zilberstein

Reasoning about Rational Agents offers a thorough study of the logical foundations of the belief-desire-intention (BDI) model of rational agency. The origins of the BDI model lie in the theory of human practical reasoning developed by the philosopher Michael Bratman (1987) in the mid-1980s. The theory was further developed by the members of the Rational Agency Project at SRI International and has seen a number of successful implementations, most notably the PROCEDURAL REASONING SYSTEM (PRS) (Georgeff and Lansky 1987) and its many descendants.

BDI research encompasses three strands of work: (1) philosophy, (2) software architectures, and (3) mathematical logic. In philosophy, research has focused on understanding practical human reasoning. In software architectures, research has focused on particular knowledge representation schemes and inference engines suitable for real-world applications. In mathematical logic, research has focused on logic modalities to represent and reason about beliefs, desires, and intentions. The synergy between the three active strands has produced numerous successful systems and serves as a model for good AI research. Some critiques of the theory argue that the three strands are only loosely connected and that BDI architectures overemphasize the notion of following intentions and neglect the harder issues of how to form the "correct" intentions given one's desires and goals. Nevertheless, few other agent architectures have been used as widely as the BDI model by a broad community of researchers.

Among the three strands of BDI research, *Reasoning about Rational Agents* focuses almost exclusively on the logical foundations of the theory. The book develops a logic, called LOGIC OF RATIONAL AGENTS (LORA), and investigates its properties and ability to capture the relationships among beliefs, desires, and intentions in single-agent

■ *Reasoning about Rational Agents,* Michael Wooldridge, Cambridge, Massachusetts, The MIT Press, 2000, 226 pp., $35.00, ISBN 0-262-23213-8.

and multiagent setups. LORA combines four components: (1) a classical first-order logic component; (2) a BDI component to express the beliefs, desires, and intentions of agents within the system; (3) a temporal component to represent the dynamic aspects of systems; and (4) an action component to represent the effects of actions that agents perform. The book extends previously developed BDI logics, in particular the framework of Anand Rao and Michael Georgeff (1991). The key contribution of the book is its treatment of multiagent notions, such as teamwork, communication in the form of speech acts, and cooperative problem solving. The result is a comprehensive examination of the logical foundations of the BDI model from its basic notions to the most advanced ones that have been formalized to date.

Writing a book about such a rich modal logic presents the obvious challenge of balancing mathematical rigor and accessibility for the broad AI community. The author has navigated this terrain quite successfully. The book starts with a clear and gentle introduction to LORA that defines its four components and the underlying logical concepts and notations. The introductory sections are self-contained and easily accessible to students and researchers with only basic familiarity with discrete math and first-order logic. However, as successful as the author has been, this book is not easy to read, in part because the material is very much at the cutting edge of contemporary logic research and could serve as a source of ideas for researchers in pure logic.

The book starts with a brief introduction to rational agency, in which the author defines the notion of rationality used in the book and explains why a logical approach to formalizing it is the most appropriate one. An agent is said to be *rational* "if it chooses to perform actions that are in its own best interests, given the beliefs it has about the world" (p. 1). Logic, the book argues, is the most appropriate approach for investigating the problem of rational agency because it makes it possible to examine rigorously the expressive power of the theory; because it is transparent—properties, interrelationships, and inferences are open to examination; and because it uses the results of "the oldest, richest, most fundamental, and best-established branch of mathematics" (p. 12). The computational complexity of logical reasoning and the lack of complete axiomatization of LORA (which hinges on a major open problem in temporal logic) do not deter the author from seeing it as the preferred approach to reasoning about rational agents.

The remainder of the book is divided into three major parts. The first part (chapters 2 and 3) is background material. It offers an excellent introduction to the BDI model, from Bratman's original intention-based theory of practical reasoning to its application to the design of agents. A set of agent designs are detailed in pseudocode, highlighting some fundamental issues

such as commitment strategies, reconsidering intentions, and the general problem of interleaving action and deliberation. The introductory part also covers the background in logic that is needed to understand LORA and develops the notation used in the book.

The second part (chapters 4 and 5) defines LORA and examines its basic properties. It covers the syntax and semantics of LORA, presents some derived logical connectives used in the remainder of the book, and establishes some properties of the logic. It then investigates how LORA can be used to capture properties of rational agents, such as the possible interrelationships among beliefs, desires, and intentions. This part of the book is mostly about single-agent systems.

The third part (chapters 6, 7, and 8) shows how the framework can be used to capture properties of multiagent systems. It shows that LORA can be used to formalize collective mental states such as mutual beliefs, desires, intentions, and joint commitments. To address speech acts, the *inform* and *request* actions are defined in LORA. The inform action is used when an agent attempts to get another agent to believe something, and the request action is used when an agent gets another agent to intend something. A range of other speech acts are then defined using these primitives. LORA is also used to define a model of cooperative problem solving. The model has four stages: (1) an agent recognizes the potential for cooperation with respect to one of its actions, (2) the agent solicits assistance from a team of agents, (3) the team attempts to agree to a plan of joint action, and (4) the plan is executed by the group.

The book concludes with a chapter on the general role of logical theories in the development of agent systems, a comprehensive bibliography, and two appendixes. In the last chapter, the author adopts a software-engineering perspective to examine the utility of logical theories in specification, implementation, and verification of agent systems. Appendix A contains a summary of notation, and appendix B contains a self-contained introduction to modal and temporal logics.

The role of logic, and in particular

that of modal logic, in formalizing practical reasoning has been widely debated in AI (McCarthy 1997). Nevertheless, it continues to be an important, thriving research area that helps to crystallize some of the deepest questions in the field. I was puzzled, however, by the attempt in the book not only to emphasize the importance of modal logic but also to dismiss the merits of competing approaches. In particular, the arguments presented against the use of decision theory and game theory are perplexing. The main argument against decision theory is that it defines what optimal actions are, but it has nothing to say about how we might efficiently implement the action-selection function. This criticism applies to a large extent to LORA itself because the author admits that it "is not an executable logic, and there is currently no simple way of automating LORA" (p. 15).

The tension between BDI and competing approaches could have been explored in greater depth. The notion of "rationality" is rooted in economics and decision theory, and just based on the title of the book, I expected a more detailed examination of the decision-theoretic approach. The past decade has seen much progress in refining decision-theoretic approaches to rational agency by developing effective graphic representations of probabilistic models, new algorithms for belief revision, and useful models for probabilistic metareasoning that address the limited computational resources in the face of complexity. Some fundamental aspects of the BDI model appear to be inconsistent with decision theory. For example, suppose that an agent has the ability to reconsider its intentions and selected actions. Within the BDI framework, "if the agent chose to deliberate but did not change intentions, then the effort expended on deliberation was wasted" (p. 38). The conclusion in this case is that the reconsider function is not optimal. This is not the case in decision theory, which could justify deliberation in the

Simulating Organizations

Computational Models of Institutions and Groups

Edited by Michael J. Prietula, Kathleen M. Carley, and Les Gasser

6 x 9, 350 pp., $45.00, ISBN 0-262-66108-X
(Prices higher outside the U.S. and subject to change without notice.)

To order, call 800-356-0343 (US and Canada) or (617) 625-8569.
Distributed by The MIT Press, 5 Cambridge Center, Cambridge, MA 02142

face of uncertainty, even if it leads to no change of strategy. Imagine, for example, a company intending to purchase a software package to improve customer service. A study of the effectiveness of the product is published that could alter the company's decision. Under certain circumstances, it can be shown that examining the study and its implications is rational, even if it is time consuming and requires a delay in purchasing the new software. Furthermore, this decision could be rational (and optimal) even if it leads to no change in intention. Identifying and explaining such contradictions between BDI and decision theory would have been helpful for readers.

The attempt to justify a logicist approach and dismiss competing approaches to rational agency goes against a growing recognition in the field of AI that there is a real need for a multitude of reasoning techniques, even within a single-agent architecture. The complexity and richness of "practical" reasoning is unlikely to be completely captured within a single formalism, which is why the late Herbert Simon (1995) recommended we "secure and maintain a tolerance throughout our discipline for a plurality of approaches." The key question, which the book neglects to examine, is not which model is the best in some absolute sense, but what technique is most suitable for a particular class of environments. There has been plenty of evidence that the BDI model is useful for practical applications. However, the book leaves open an important question: What problem domains and characteristics of the environment make BDI the "correct" approach?

Overall, this is a well-written book that maintains a lot of enthusiasm for modal logic. It takes the reader on a thorough journey into the logical foundations of the BDI model of rational agency. It is an excellent resource for anyone who wants to understand, study, or use this model.

References

Bratman, M. E. 1987. *Intention, Plans, and Practical Reasoning.* Cambridge, Mass.: Harvard University Press.

Georgeff, M. P., and Lansky, A. L. 1987. Reactive Reasoning and Planning. In Proceedings of the Sixth National Conference on Artificial Intelligence, 677–682. Menlo Park, Calif.: American Association for Artificial Intelligence.

McCarthy, J. 1997. Modality, Si! Modal Logic, No! *Studia Logica* 59:29–32.

Rao, A. S., and Georgeff, M. P. 1991. Modeling Rational Agents within a BDI Architecture. In *Proceedings of Knowledge Representation and Reasoning,* eds. R. Fikes and E. Sandewall, 473–484. San Francisco, Calif.: Morgan Kaufmann.

Simon, H. A. 1995. Artificial Intelligence: An Empirical Science. *Artificial Intelligence* 77(1): 95–127.

Shlomo Zilberstein is an associate professor of computer science and director of the Resource-Bounded Reasoning Lab at the University of Massachusetts at Amherst. He received his B.A. in computer science summa cum laude from the Technion, Israel Institute of Technology, and his Ph.D. in computer science from the University of California at Berkeley. Zilberstein's research interests include decision theory, design of autonomous agents, heuristic search, monitoring and control of computation, planning and scheduling, reinforcement learning, and reasoning under uncertainty. His e-mail address is shlomo@cs.umass.edu.