A review of "Reasoning About Rational Agents" by Michael Wooldridge, MIT Press 2000

Gordon Beavers and Henry Hexmoor

Reasoning About Rational Agents is concerned with developing practical reasoning (as contrasted with theoretical reasoning) for rational computer agents within the "Belief, Desire, Intention" model. The BDI model for computer agents is based on the theory of rational action in humans put forward in 1988 by the philosopher M. Bratman in his book **Intention, Plans and Practical Reason**. BDI logics are multi-modal logics developed by Rao and Georgeff during the 1990s. Wooldridge's version of BDI logic, which extends the work of Rao and Georgeff, is called **LORA** for "Logic of Rational Agents". Wooldridge is concerned with the problem of writing a book that is both accessible and rigorous. The result is a book whose first three chapters can be easily read by anyone with a modest background in logic. However, Chapter 4 assumes a discontinuously higher expectation in the reader's knowledge. Chapters 6 through 9 presuppose some knowledge in the field of agency. After the introductory chapter one, Wooldrige divides the book into three parts: (part 1) chapter 2 provides background material on the BDI model while chapter 3 gives an introduction to **LORA**; (part 2) chapter 4 provides the formal syntax and semantics for **LORA** while chapter 5 covers some properties of rational agents; (part 3) investigates the use of **LORA** in multi-agent social systems.

At the end of each chapter is a very helpful section entitled *Notes and Further Reading* for the reader who wishes to take advantage of Wooldridge's vast knowledge in the field of agency. There is also a website (http://www.csc.liv.ac.uk/~mjw/pubs/rara/errata.html) where Wooldridge collects errata.

Chapter 1 introduces the reader to the subject of rational agents and gives an outline of the rest of the book. Wooldridge wishes to consider agents that are autonomous, proactive, reactive, and have some social ability in the sense of being able to negotiate and cooperate with other entities. From a software engineering perspective, agents are cast as a subset of Pneulian Reactive systems [Manna, et al, 1995]. In contrast to functional systems, the role of a Pneulian reactive system is to adaptively maintain an interaction with its environment. These systems are specified in terms of their on-going behavior. For example, a compiler is a functional system whereas a multiplayer game program with asynchronous concurrent processes might be a reactive system depending on its specification. If the game program

considers its interactions with the players in a non-functional manner, it is an agent.

Wooldridge's book is concerned solely with the belief-desire-intention (BDI) model of rational agency. The BDI model has been selected because (i) it is based on Bratman's theory of rational action in humans which is widely known, (ii) it has been implemented several times, e.g., Geogeff's PRS (although, Wooldridge's version has not been implemented), and (iii) the theory has been formalized in a family of BDI-logics. Intentions enable an agent to constrain the search space of possible actions to perform.

Wooldridge very briefly suggests some justification for his choice of a multi-modal branching-time logic over Decision Theory, Game Theory, and First-Order Logic.

Chapter 2 introduces the reader to the philosophical and to the software engineering components of the BDI model of agency. Practical reasoning is separated into deciding what to do (deliberation) and how to do it (means-end reasoning, planning), both of which can be computationally expensive. After a discussion of deliberation and means-end reasoning, intentions are characterized as “pro-attitudes” (inasmuch as they tend to lead to action) that drive means-ends reasoning, that persist, and that constrain future practical reasoning.

Wooldridge progressively refines an algorithm for implementing an agent with particular attention to the sequence *deliberate-plan-act*. Wooldridge's first revision uses perceptions to update beliefs and deliberation on beliefs to update intentions. Plans are then selected based on beliefs and intentions. Wooldridge then considers the deliberation process, which he decomposes into “option generation” and “filtering of the options generated”. Options (which Wooldridge calls “desires”) are determined by optimal beliefs and intentions and these options (or “desires”) are combined with beliefs and intentions to update intentions.

According to Wooldridge “(w)hen an option successfully passes through the filter function and is hence chosen by the agent as an intention, we say that the agent has made a commitment to that option” and commitment implies persistence. The agent needs a “commitment strategy” to determine when and how to drop intentions. Three common strategies are (i) maintain an intention until it is realized (blind or fanatical commitment), (ii) maintain an intention until either it is realized or it is not possible (single-minded commitment), and (iii) maintain an intention as long as it is still believed possible (open-minded commitment).

Wooldridge adds the ability to replan when a plan goes awry or when the agent has otherwise determined that its current intentions are inappropriate, so as to avoid "over commitment". A well-designed agent should exploit serendipity by reconsidering intentions during plan execution, but reconsideration is expensive. A cautious agent reconsiders intentions in each loop whereas a bold agent would never reconsider. Wooldridge outlines Kinny and Georgeff's experiments, which show that bold agents tend to outperform cautious agents in a stable environment, but the reverse is true in an unstable environment.

Wooldridge provides some justification for the use of BDI mentalistic terminology; in those situations where we do not know enough about a system to give a physical or design explanation of the system's behavior it yields greater understanding. Wooldridge compares the "intentional" to the "physical" and the "design" stances. The intentional stance is an abstraction tool that is appropriate when complexity or ignorance prevents a physical explanation.

Chapter 3 gives a very intuitive and accessible overview of **LORA**. **LORA** is first-order logic with modalities for the intentional notions of Belief, Desire and Intention, branching temporal structures and action expressions. Wooldridge mentions the problems associated with substitution into modal (opaque) contexts and with quantifying into temporal contexts.

Chapter 4 gives the syntax and semantics of **LORA**. This chapter is very formal and perhaps demands more of the reader than any other chapter in the book. Axioms for **LORA** are not given and consequently no proof theory is provided.

Chapter 5 explores the various ways in which beliefs, desires, and intentions can interact. Wooldridge issues several disclaimers. He first warns the unwary reader that **LORA** does little to capture the notions of belief, desire, and intention in humans. Although no axiomatization of BDI has been provided Wooldridge does consider BDI correspondence theory, that is, the relationships between properties of the accessibility relations in the models and the axioms of the logic. For example it is shown that if the desire accessibility relation is a structural subset of the belief accessibility relation then, if in some situation it is inevitable that agent i believes ϕ then it is inevitable that agent i desires that ϕ in that situation.

Wooldridge goes on to consider the plausibility of various interactions between beliefs, desires, and intentions. He contends, for example, that the

schema “if agent i believes that ϕ then i desires ϕ” (realism) is unreasonable. In his discussion of realism Wooldridge concludes that intention – desire consistency is reasonable unless reconsideration of intentions is required.

Chapter 6 considers collective mental states. Following an example demonstrating the need for mutuality Wooldridge provides the semantics of mutual belief including a fixed-point theorem. Wooldridge then shows that mutual belief behaves like a KD4 modal operator.

Wooldridge adopts a reductionist approach with respect to teamwork. Wooldridge’s development is influenced by the work of Cohen and Levesque [Cohen and Levesque, 1991].

Chapter 7 shows how certain types of communications between agents can be modeled in **LORA**. Wooldridge first reviews the speech act theory of J. L. Austin, the later contributions of J. R. Searle , the STRIPS formalism of Cohen and Perrault, and the agent communication languages KQML and FIPA. Building on this foundation Wooldridge defines the semantics of “inform”, “request” and other speech acts.

In chapter 8 Wooldridge takes up the question of how autonomous agents might achieve cooperative problem solving in **LORA**. He suggests a model with four stages: (i) recognition, (ii) team formation, (iii) plan formation, and (iv) team action. He begins with a formal definition of “ability” for both individuals and groups and then develops formal results in **LORA** for each of the four stages.

While most of the book before chapter 9 is directed towards an understanding of Wooldridge’s **LORA**, chapter 9 itself address the broader question of the role of logic in building rational agent software. Wooldridge considers the following roles for logic (i) as a specification language, (ii) as a programming language, and (iii) as a verification language. In each case he outlines how logic can be utilized in the role and the problems to be overcome. He also provides case studies for the use of logic as either a programming language or a verification language.

Two appendices conclude the book. The first is a short summary of the notation used in the book while the second is a twenty-page introduction to modal and temporal logics.

Reasoning about Rational Agents can be used as a primer text on the logical approach to agents and multiagency. However, the title might mislead the novice reader since there are other approaches to agents and multiagency that

can fit under this topic. Examples are decision theory and utility theory. Although Wooldrige explains the differences, we know those theories are rooted in economic viewpoints and have devoted followers in the agent community. To followers of that community, rational agents may mean agents that are based on those theories. Albeit, a comparable book to Wooldridge's does not exist in that area.

LORA represents a class of logics and one can readily tailor it to specific kinds of agents that can appear to have certain traits such as cautiousness, or boldness. It can also be used to define more complex notions such as teams or agents capable of emotions. This book gives us an accepted platform for building more logically oriented agents. An interesting companion book might be one on computational and implemented agents based on the logic of **LORA**.

References:

Michael Bratman, **Intentions, Plans, and Practical Reason**, Harvard University Press, Cambridge, 1987.

Philip Cohen. and H. Levesque, H. J. 1991. Teamwork. Nous 25(4), Special Issue on Cognitive Science and Artificial Intelligence, pp. 487-512.

Zohar Manna and the STeP group. STeP: The Stanford Temporal Prover (Educational Release), User's Manual. Technical report STAN-CS-TR-95-1562, Computer Science Department, Stanford University, November 1995.