AI Magazine Volume 16 Number 2 (1995) (© AAAI)

A Review of *Rules of Encounter: Designing Conventions for Automated Negotiation*

Piotr Gmytrasiewicz

Rules of Encounter: Designing Conventions for Automated Negotiation, Jeffrey S. Rosenschein and Gilad Zlotkin, The MIT Press, Cambridge, Massachusetts, 1994, 221 pp., $35.00, ISBN 0-262-18159-2.

The main contribution of the book *Rules of Encounter: Designing Conventions for Automated Negotiation,* by Jeffrey S. Rosenschein and Gilad Zlotkin, is the formulation of a principled framework within which to study interactions among artificial heterogeneous agents. This framework is based on the theory of games, which is aimed at decision problems faced by agents in situations in which the agent's welfare depends not only on its own actions but also on the actions of other agents. The examples are numerous: The personal digital assistants (PDAs) that might one day keep track of their users' itinerary will have to negotiate with PDAs of other people to adjust and synchronize their meeting schedules. Software agents looking for the right kinds of information on the Internet on behalf of their users might have to negotiate with other such agents over the access to resources. Computer agents that control a telecommunications network will have to interact with computers that control other networks and might find it beneficial to come to agreement with them.

The challenge that the book undertakes is to facilitate the design of rules of interaction and negotiation in such a way that the society of agents as a whole exhibits desirable properties, such as stability and efficiency, chosen by the designers of the systems. The added difficulty for this challenge is that in many of the domains in which it might be useful to rely on such agents, quite likely, they will be designed and manufactured by different companies. More importantly, they will represent interests and preferences that might not coincide and, in fact, might conflict. The issue is, therefore, how can a system of rules be designed that would ensure a stable and efficient interaction pattern and would prevent the selfish agents from manipulating the system by, say, lying to each other, cheating, and otherwise taking unfair advantage of others?

The authors envision that the companies that manufacture the agents will worry about these things and agree on the most desirable set of rules. Thus, the designers from IBM, Apple, and Sony might get together and agree on a protocol that determines what kind of deals the PDA-based agents can make when negotiating access to resources during scheduling. They have to consider a particular domain because as it turns out, the properties of a system using a protocol are dependent on the characteristics of the domain in which the interaction is taking place. The focus of the book, therefore, is on using the tools of game theory to analyze negotiation protocols in different domains and check whether they can be made to lead to interaction patterns that are stable, efficient, and nonmanipulable.

The domains are categorized into a hierarchy of classes. The simplest class consists of *task-oriented domains.* In these domains, agents are given a list of jobs to do; the jobs are nonconflicting and can be redistributed among the agents. The redistribution is what the agents will have to negotiate given that each agent is selfish and attempts to minimize its own effort, but they do not have to worry about conflicts over resources. On the higher level of the hierarchy is the class of *state-oriented domains,* which is a superset of the task-oriented domains. In state-oriented domains, each agent has to move the world from a given initial state to one of the set of given goal states for this agent. An example can be the block world in which one agent's goal is to have block A on block B, and another agent's goal is to have B on C. In this class of domains, the agents have to worry about the possible conflicts over the resources and try to find the state that satisfies the goals of all the agents. The most general class is the class of *worth-oriented domains,* which is the superset of state-oriented domains. In *worth-oriented domains,* each potential state has some degree of desirability, or worth (a real number), for each agent. Here, the agents can make compromises, accepting an outcome that has a lower desirability for them, provided, for example, that they do less work to bring it about.

The category of task-oriented domains is the most specific and the one with the strongest results. For example, consider the possibility that agents might attempt to benefit by deceiving each other during negotiation. It turns out that if two agents use a negotiation mechanism that maximizes the product of the utilities each of them gets, then certain kinds of deceptive behavior are unsafe because there is a positive probability that they will be discovered. For example, it is not beneficial for an agent to lie by claiming to have a phantom task, which it has to achieve in addition to its actual tasks, given that a sufficiently severe penalty is imposed once the deception is discovered. If one further restricts the domain to the so-called *subadditive task-oriented domain* (in which the cost of performing a sum of two sets of tasks cannot be larger than the sum of the costs of performing the sets of tasks separately), then there is a mechanism that makes it irrational for any agent to attempt to hide one of the original tasks. The importance of these results is clear: They present the protocol designers with ways to make the negotiation mechanisms used in the task-oriented domains immune to manipulation.

In the more general state-oriented domains, conflicts over the resources can occur. The agents, therefore, might have to be ready to expend more effort in a multiagent situation than in the situation where the other agents were not present. This difference is essentially the overhead of coordination. Furthermore, in these domains, even in a conflict situation, agents can benefit by helping each other out part of the way and then resolving the conflict, say, by flipping the coin to decide who will get its wish rather than decide to resolve the conflict at the beginning, with the winning agent doing all the work itself. This possibility can be exploited by the agents making a suitable deal, called a *semicooperative deal*. A mechanism that allows the agents to make these kinds of deals, the *unified negotiation protocol*, is then useful for resolving conflict situations as well as reaching cooperative agreements.

The matters get worse if the possibility of deception is considered. Unlike in the task-oriented domains, in the state-oriented domains, the agents might find it beneficial to lie about their set of tasks, and no mechanism is likely to make it irrational. This negative result seems to originate in part from an implicit assumption made by the authors that the agents are completely gullible. Thus, the agents are assumed to be smart enough to try to deceive others, but at the same time, they are assumed to believe everything they hear from others. The possibility that the agents could disbelieve what they hear is not considered. Although doing so would clearly make the analysis of such cases much more complicated, it might be one of the ways to make the agents less prone to deception while they operate in more general environments.

The most general class of domains that the authors consider is the *worth-oriented domain*, a generalization of the state-oriented domains. In these domains, agents can assign a measure of worth to each of the possible states of the world, representing the intuitive notion of partial goal satisfaction. Human users, for example, almost never have all-or-nothing goals, and a partial fulfillment is still better than none. The gradation of worth opens the field of deal making during negotiation to include trading off the worth of the final state as well as the agents' part in the work that brings about the final state. This gradation allows the agents to compromise in new ways and enables agreements that would be impossible otherwise, thus improving the effectiveness of the overall interaction. Here, agents can use the unified negotiation protocol that was defined previously for state-oriented domains.

… the role of th[is] book—and its main contribution—is in illuminating the issues that the protocol designers for multiagent systems might soon have to face. In providing this insight, the principled approach based on game theory used by the authors is particularly valuable.

A demanding reader might walk away somewhat disappointed with the generality of the results presented. For example, there does not seem to be a way to reliably discourage deception in general encounters. Furthermore, some of the assumptions the authors make are a little unclear. Apart from the gullibility assumption mentioned earlier, the authors concentrate on agents that agree on deals that maximize the product of the payoff values that each of them expects to get (called the *product-maximizing mechanism*). At the same time, the authors postulate that the agents be selfish utility maximizers. The two are not identical, as shown by the following example: An agent that maximizes its own payoff will prefer a deal that gives it a payoff of 1000 and gives 0 to another agent over a deal that gives a payoff of 1 to both. An agent operating under the product-maximizing mechanism, however, would be happy with the second deal. Now, it can be shown that even a selfish utility maximizer would adopt the product-maximizing mechanism if it wanted the interaction result to be fair and symmetric. The remaining question is, of course, what could convince a selfish agent to use these humanitarian values?

Apart from these minor inconsistencies, the book is clearly written. The authors, probably realizing that the formalism used in game theory is not the bread and butter of every AI reader, were careful to include plenty of convincing examples to get the point across and left some of the detailed proofs to the appendix. In summary, the role of the book—and its main contribution—is in illuminating the issues that the protocol designers for multiagent systems might soon have to face. In providing this insight, the principled approach based on game theory used by the authors is particularly valuable.

Piotr Gmytrasiewicz received his Ph.D. from the University of Michigan. He is now an assistant professor in the Computer Science and Engineering Department at the University of Texas at Arlington. His interests concentrate on applying decision-theoretic techniques to problems in multiagent interactions and AI.