Fairness in Multi-Agent Systems

Steven de Jong
MICC, Maastricht University, The Netherlands
steven.dejong@micc.unimaas.nl

1. THE NEED FOR FAIRNESS

In our increasingly interconnected world, research in the area of multi-agent systems is becoming more and more important [1, 2]. Multi-agent systems are generally accepted as valuable tools for designing and building distributed dynamical systems, by using several interacting agents, possibly including humans. In practice, multi-agent systems are often performing tasks in co-operation with, or instead of humans. Examples include software agents participating in online auctions or bargaining [3, 4], electronic institutions [5], developing schedules for air traffic [6] and decentralized resource distribution in large storage facilities [7, 8].

Although multi-agent systems have many potential advantages, designing them raises many difficulties. One of the key problems lies in controlling the behavior of individual agents in such a way that the system as a whole reaches a certain goal. This problem becomes even more prominent in multi-agent systems that interact with humans. Usually, multi-agent systems are designed assuming perfectly rational, self-interested agents, according to the principles of classical game theory. However, recently, this strong assumption has been relaxed in various ways, for instance by including well-known concepts such as bounded rationality [9] and social welfare [10, 11]. Research in the field of behavioral economics shows us that humans are not purely rational and self-interested; their decisions are often based on considerations about others. For instance, humans strongly care about receiving a fair share [12, 13, 14]. The concept of a fair share relates closely to a set of problems called *social dilemmas*. In such dilemmas, agents have to choose between being selfish (i.e., being individually rational and caring only for their own benefit) or being social (i.e., being driven by fairness considerations and also taking into account the benefit of others). The dilemma lies in the fact that being perfectly individually rational (and thus selfish) may lead to a lower benefit than being fair (and thus social). There are two distinct reasons why this may happen. First, other agents in the system may get frustrated by selfish actions and may decide to reject the offender. Second, cooperation may simply be necessary in the problem at hand to obtain a satisfactory payoff. In other words, the failure of selfish behavior may be caused either by other agents in the system or by the problem at hand.

We will provide an example of each of these two situations here, both of which are stylized games from behavioral economics research. First, in the Ultimatum Game [15], an agent proposes how to divide a reward with a second agent. If the second agent accepts this division, the first gets his demanded payoff and the second gets the rest. If however the second agent rejects, neither gets anything. The individually rational, selfish solution to this game is for the first agent to leave the smallest positive payoff to the other agent. After all, the other agent can then choose between receiving this payoff by agreeing, or receiving nothing by rejecting. However, human players consistently offer more, and low offers are almost always rejected. Thus, playing selfishly leads to a very low payoff for the first agent. Second, in the Public Goods Game [16], potentially many agents are given a sum of money and are asked to invest in a common pool. After everyone invested (or not), this pool is multiplied by a factor (usually three) and subsequently divided over all the players. Thus, if every agent invests, everyone will receive more money than they invested. Playing selfishly, agents will not invest, since they will be able to keep their money and additionally receive a share from the common pool. However, if all agents play selfishly, the common pool is empty, which means that nobody earns any money.

More generally speaking, being aware of concepts such as fairness may lead to better results in any problem domain in which the allocation of limited resources plays an important role [17], as in many of the aforementioned examples. Therefore, concepts such as fairness, discovered in such diverse fields as behavioral economics, economical psychology and evolutionary game theory, must be well-understood by developers of multi-agent systems. Fairness has been extensively studied, resulting in so-called *descriptive* models of human fairness, explaining why and how humans reach fair solutions instead of individually rational ones. These models may be used as a basis for *prescriptive* or computational models, used to control agents in multi-agent systems in such a way that alignment with human expectations is achieved.

2. RESEARCH SUMMARY

Since our goal is to obtain fairness in multi-agent systems by looking at human fairness, we first study literature on human fairness extensively. In our research on this topic, we identify two main existing descriptive models of human fairness in existing research, which can be summarized with the terms *inequity aversion* and *reciprocal fairness*. The first model focuses on one-shot interactions between humans, such as the Ultimatum Game, and addresses the observation that humans tend to dislike large differences in payoffs, with an emphasis on disadvantageous differences. The second model focuses on repeated interactions, such as the Public Goods Game, and shows that humans tend to be reciprocal, i.e., they are nice to others that are nice to them, and willing to punish others that are somehow offensive, even if this punishment is costly. In order to know who is nice and who is nasty, the reputation of others must

Cite as: Extended Thesis Abstract for Doctoral Mentoring Program, S. de Jong, Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008), Padgham, Parkes, Müller and Parsons (eds.), May, 12-16., 2008, Estoril, Portugal, pp. 1742-1743.

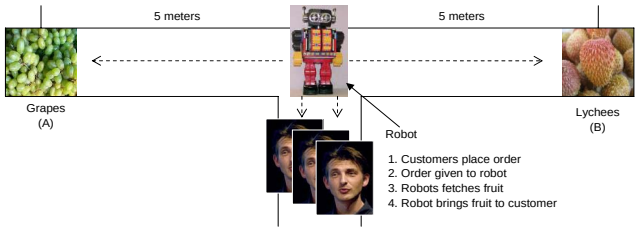

Figure 1: A small experiment, demonstrating human priority awareness in a vegetable shop. Respondents are asked to place the robot on the line between (A) and (B), in such a way that all customers of the shop are satisfied.

somehow become known.

We also perform our own experiments with humans, to gain understanding in the two existing models, to verify whether the models accurately predict human strategies, and if not, to identify possible flaws. We find that indeed, the current two existing models are missing an important element (which has been present in classical game theory for a long time), viz. *bargaining power*, *immediate reputation* or (as we call it throughout this thesis) *priority*. In other words, humans do not necessarily need repeated interactions to be able to classify an other person as being nice or nasty; additional (explicit or implicit) information they may have about this other person immediately influences their strategies. To demonstrate this, we initially developed a survey (see Figure 1), in which 50 faculty members and students participated. Here, the probability that customers order the item located at (A) influences the position people choose for the robot. A similar phenomenon was observed in Ultimatum Games in which the players were told that their opponents were not equally wealthy [18]. To address this phenomenon, we develop *priority awareness*, our own descriptive model of human fairness [18, 19].

After having acquired sufficient knowledge on descriptive models of human fairness, we turn to prescriptive, computational models. First, we argue that the notion of 'fairness' is overly vague and ambiguous and should therefore be clearly defined in a computational context. Then, we focus on obtaining computational fairness in adaptive multi-agent systems, more precisely, using multi-agent reinforcement learning in single-state problems (mostly related to bargaining). In order to learn fair strategies, agents have been equiped with a fairness utility function. We develop suitable utility functions that enable our agents to learn computational fairness, in accordance with each of the three descriptive models. Using experiments and analysis, we determine whether these utility functions are in accordance with what we currently know about human fairness. In other words, we determine whether our agents reach bargains that are similar to human bargains.

3. CONTRIBUTIONS OF THIS THESIS

Here, we outline the most important contributions of this thesis. First, this thesis presents an overview of the current state of the art in descriptive modeling of human fairness [19]. Second, we show that an important element is missing in current descriptive models of human fairness (i.e., humans do not need multiple interactions for reputation or priorities to emerge); to address this element, we introduce our own descriptive model, priority awareness [8, 18]. Third, we provide an operationalization of computational fairness for multi-agent systems. Fourth, we use (principles behind) known descriptive models of human fairness to obtain fairness in multi-agent systems, using multi-agent reinforcement learning (e.g., [20]).

4. REFERENCES

[1] N.R. Jennings, K. Sycara, and M. Wooldridge. A Roadmap of Agent Research and Development. *Autonomous agents and Multi-Agent Systems*, 1:275 – 306, 1998.

[2] Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

[3] Jayant Kalagnanam and David C. Parkes. Auctions, Bidding and Exchange Design. In David Simchi-Levi, S. David Wu, and Max Shen, editors, *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era*, Int. Series in Operations Research and Management Science, chapter 5. Kluwer, 2004.

[4] Chris Preist and Maarten van Tol. Adaptive agents in a persistent shout double auction. In *ICE '98: Proceedings of the first international conference on Information and computation economies*, pages 11–18. ACM Press, 1998. ISBN 1-58113-076-7.

[5] Juan A. Rodriguez-Aguilar. *On the design and construction of agent-mediated electronic institutions*. PhD thesis, Monografies de l'Institut d'Investigació en Intelligència Artificial, 2003.

[6] Xiaoyu Mao, Adriaan ter Mors, Nico Roos, and Cees Witteveen. Agent-Based Scheduling for Aircraft Deicing. In Pierre-Yves Schobbens, Wim Vanhoof, and Gabriel Schwanen, editors, *Proceedings of the 18th Belgium - Netherlands Conference on Artificial Intelligence*, pages 229–236. BNVKI, October 2006. ISBN 1568-7805.

[7] Danny Weyns, Nelis Boucke, Tom Holvoet, and Wannes Schols. Gradient Field-Based Task Assignment in an AGV Transportation System. In *Proceedings of EUMAS*, pages 447–458, 2005.

[8] Steven de Jong, Karl Tuyls, and Ida Sprinkhuizen-Kuyper. Robust and Scalable Coordination of Potential-Field Driven Agents. In *Proceedings of IAWTIC/CIMCA 2006, Sydney*, 2006.

[9] Herbert Simon. *Models of Man*. John Wiley, 1957.

[10] L. M. Hogg and N. R. Jennings. Socially Rational Agents. In *Proc. AAAI Fall symposium on Socially Intelligent Agents*, pages 61–63, 1997.

[11] Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. A Short Introduction to Computational Social Choice. In *Proceedings of the 33rd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM-2007)*, volume 4362 of *LNCS*, pages 51–69. Springer-Verlag, 2007.

[12] Samuel Bowles, Robert Boyd, Ernst Fehr, and Herbert Gintis. Homo reciprocans: A Research Initiative on the Origins, Dimensions, and Policy Implications of Reciprocal Fairness. *Advances in Complex Systems*, 4:1–30, 1997.

[13] E. Fehr and K. Schmidt. A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics*, 114:817–868, 1999.

[14] H. Gintis. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction*. Princeton University Press, 2001.

[15] Werner Gueth, Rolf Schmittberger, and Bernd Schwarze. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization*, 3 (4):367–388, 1982.

[16] Astrid Dannenberg, Thomas Riechmann, Bodo Sturm, and Carsten Vogt. Inequity Aversion and Individual Behavior in Public Good Games: An Experimental Investigation. *SSRN eLibrary*, 2007.

[17] Yann Chevaleyre, Paul E. Dunne, Ulle Endriss, Jérôme Lang, Michel Lemaître, Nicolas Maudet, Julian Padget, Steve Phelps, Juan A. Rodriguez-Aguilar, and Paulo Sousa. Issues in Multiagent Resource Allocation. *Informatica*, 30:3–31, 2006.

[18] Steven de Jong, Karl Tuyls, Katja Verbeeck, and Nico Roos. Priority awareness: towards a computational model of human fairness for multi-agent systems. *Adaptive Agents and Multi-Agent Systems III - Lecture Notes in Artificial Intelligence*, 4865, 2008.

[19] Steven de Jong, Karl Tuyls, and Katja Verbeeck. Fairness in multi-agent systems. *Knowledge Engineering Review*, (accepted).

[20] Steven de Jong, Karl Tuyls, and Katja Verbeeck. Artificial Agents Learning Human Fairness. In *Accepted at the international joint conference on Autonomous Agents and Multi-Agent Systems (AAMAS'08)*, 2008.