Priority Awareness: Towards a Computational Model of Human Fairness for Multi-agent Systems

Steven de Jong, Karl Tuyls, Katja Verbeeck, and Nico Roos

MICC/IKAT, Maastricht University, The Netherlands
{steven.dejong,k.tuyls,k.verbeeck,roos}@micc.unimaas.nl

Abstract. Many multi-agent systems are intended to operate together with or as a service to humans. Typically, multi-agent systems are designed assuming perfectly rational, self-interested agents, according to the principles of classical game theory. However, research in the field of behavioral economics shows that humans are not purely self-interested; they strongly care about whether their rewards are *fair*. Therefore, multi-agent systems that fail to take fairness into account, may not be sufficiently aligned with human expectations and may not reach intended goals. Two important motivations for fairness have already been identified and modelled, being (i) inequity aversion and (ii) reciprocity. We identify a third motivation that has not yet been captured: priority awareness. We show how priorities may be modelled and discuss their relevance for multi-agent research.

1 Introduction

Modelling agents for a multi-agent system requires a thorough understanding of the type and form of interactions with the environment and other agents in the system, including any humans. Since many multi-agent systems are designed to interact with humans or to operate on behalf of them, for instance in bargaining [1,2], agents' behavior should often be aligned with human expectations; otherwise, agents may fail to reach their goals.

Usually, multi-agent systems are designed according to the principles of a standard game-theoretical model. More specifically, the agents are perfectly rational and optimize their individual payoff disregarding what this means for the utility of the entire population. Experiments in behavioral economics have taught us that humans often do *not* behave in such a self-interested manner [3,4,5]. Instead, they take into account the effects of their actions on others; i.e., they strive for *fair* solutions and expect others to do the same. Therefore, multi-agent systems using only standard game-theoretical principles risk being insufficiently aligned with human expectations. A prime example is the ultimatum game [4], in which purely rational players will not be able to obtain a satisfactory payoff. More generally speaking, the importance of fairness should be studied in any problem domain in which the allocation of limited resources plays an important role [6]. Examples from our own experience include decentralized resource distribution in large storage facilities [7], aircraft deicing [8], and representing humans in bargaining (e.g., [1,9]).

Thus, designers of many multi-agent systems should take the human conception of fairness into account. If the motivations behind human fairness are sufficiently understood and modelled, the same motivations can be transferred to multi-agent systems.

K. Tuyls et al. (Eds.): Adaptive Agents and MAS III, LNAI 4865, pp. 117–128, 2008.

More precisely, *descriptive* models of human fairness may be used as a basis for *prescriptive* models, used to control agents in multi-agent systems in a way that guarantees alignment with human expectations. This interesting track of research ties in with the descriptive agenda formulated by Shoham [10] and the objectives of evolutionary game theory [5,11].

In the remainder of this paper, we first briefly discuss related work in the area of fairness models. Then, we look at problems in which priorities play a role. We show that current descriptive models do not predict human behavior in such problems. Next, we provide our descriptive model, priority awareness, and perform experiments to show that the model performs a much better prediction of human behavior. We conclude with some directions for future work.

2 Related Work

Already in the 1950's people started looking at fairness, for instance in the Nash bargaining game [12]. Recently, research in behavioral economics and evolutionary game theory has examined human behavior in the ultimatum game and the public goods game [3,4,5,13,14]. In all cases, it was observed that standard game theoretical models predict a very selfish outcome in comparison to the fair outcomes reached by human players. In other cases, e.g. the Traveler's Dilemma [15], it was shown that humans can actually obtain a higher payoff by failing to find the rational solution, i.e., the Nash equilibrium. Using neuroscientific research, such as MRI scanning [16] and disrupting certain areas if the brain using magnetic stimulation [17], it has been assessed which brain areas are likely to be responsible for fair behavior. The current state of the art provides two important descriptive models of human fairness.

Inequity aversion. The first descriptive model for human fairness is *inequity aversion*. In [4], this is defined as follows: *"Inequity aversion means that people resist inequitable outcomes; i.e., they are willing to give up some material payoff to move in the direction of more equitable outcomes"*. To model inequity aversion, an extension of the classical game theoretic actor is introduced, named homo egualis [4,5]. Homo egualis agents are driven by the following utility function:

$$u_i = x_i - \frac{\alpha_i}{n-1} \sum_{x_j > x_i} (x_j - x_i) - \frac{\beta_i}{n-1} \sum_{x_i > x_j} (x_i - x_j) \tag{1}$$

Here, u_i is the utility of agent $i \in \{1, 2, \ldots, n\}$, which is based on its own reward, x_i, minus a term for other agents doing better (weighed by α_i), and minus a term for other agents doing worse (weighed by β_i). Agents using the homo egualis utility function care more about inequity if it is to their disadvantage than if it is to their advantage; i.e., $\alpha_i > \beta_i$. Research with human subjects provides strong evidence that this is a valid assumption [4]. The β parameter must be in the interval $[0, 1]$ to keep behavior realistic; with a higher value for β, agents would be willing to throw away money in order to reduce inequity.