Cooperation and Fairness in Systems of Indirect Reciprocity

Martin Smit
University of Amsterdam
j.m.m.smit@uva.nl

Abstract

Across disciplines, cooperation is a fundamental research topic. While socially desirable to a population, it often bears a cost to the individual who, in their own self-interest, rationally chooses not to engage in costly cooperation. As such, much work has been done in understanding the biological mechanisms behind cooperation in human and animal populations. In my PhD project, I develop and apply these mechanisms both to artificial multi-agent systems and real social systems. I examine how factors such as agent heterogeneity and different learning algorithms affect not only the level of cooperation within a system, but also the level of fairness in the distribution of payoffs. In previous work, I showed how the effectiveness of the *social norm*-based mechanism of *indirect reciprocity* is affected when in-group biased cooperation is present. Beyond my future work on online platforms, I also plan to explore the effects of space, gossip, and partial and subjective observations to widen the potential scope of applications.

1 Introduction

Why do humans cooperate? This has been a core issue in a number of fields such as evolutionary biology, psychology and, more recently, AI. In many situations involving multiple agents, rational decisions lead to socially suboptimal outcomes due to some personal costs associated with cooperation and the risk that others may not also cooperate.

In a recent survey, Fatima *et al.* note that evolutionary selection derived from competition promotes selfish behaviour. They present an array of mechanisms that aim to curb the evolutionary pressure to be selfish which includes *indirect reciprocity* (IR), a mechanism based on reputations and social norms [Ohtsuki and Iwasa, 2004]. The goal of my thesis is to understand how reputations can be used to solve social dilemmas in ways that lead to fair outcomes.

The holistic, sometimes unconscious nature of human decision-making underlies the importance of considering fairness in social dilemmas. Firstly, similarly to how bias can creep into AI systems due to learning from biased human behaviour, a social norm that reflects a society's view of how to behave may also contain biases. These could be expressed as treating people differently based on characteristics such as age, gender, and ethnicity. Secondly, when reputation information and demographic information are both available when making decisions, biases in actions taken can affect outcomes for certain groups.

Both manifestations of bias are prevalent on many of the "gig economy" platforms that enable short-term economic interactions. On these platforms, users are assigned a reputation by other users, and some demographic information is available on one's profile. As such, those assessing your behaviour may have in-group biases, as in [Abrahao *et al.*, 2017] where AirBnB users were more likely to trust those more similar to them with an economic outcome. Furthermore, decisions can be affected by the presence of information such as names or profile pictures: Edelman *et al.* show that guests with a distinctly African American name have a 12% lower acceptance rate.

So far, my work has focused on incorporating the various types of bias into IR models. In much IR literature, a social norm is defined as a function $f : A \times C \rightarrow R$ for some some action space A, context space C and reputation space R. Typically, $A = R = \{0, 1\}$, and $C = \{0, 1\}^d$, where each dimension of C contains information such as current and past reputations or demographics.

For an example of a norm, take the pairwise *Donation game*, which gives one player (the *donor*) the opportunity to donate b utility to another player (the *recipient*) at a cost $c < b$ to themself. Clearly, in the one-shot game, a rational donor would never donate. But if the donor knows that their action will be judged, how it will be judged, and that everyone will know their reputation, how can we define a norm such that the benefit of donating outweights the temptation to defect?

It turns out that determining precisely which actions should be considered *pro*-social is a non-trivial endeavour. The cooperation level observed in a population depends on the distribution of reputations and the stability of cooperative strategies, which in turn depends on the social norms governing reputation updates. Previous works typically model the system using evolutionary game theory (EGT) and exhaustively search the space of social norms considered [Ohtsuki and Iwasa, 2004; Santos *et al.*, 2021]. Through such works, a number of "leading" norms were identified which successfully stabilise reputation-conditional cooperation, protecting against invasion both from defectors who benefit from the cooperation of others without contributing themselves *and* from cooperators who do not "enforce" the social norm through justified defection.

One issue that is rarely discussed in IR literature is that of the distribution of payoffs, resulting in different levels of *fairness*. This concept is particularly important if IR is used to model heterogeneous or group-structured systems, as norms may serve to perpetuate inherent inequalities if, for example, agents have varying abilities to perform cooperative actions.

To tackle this issue, I introduce an in-group/out-group distinction to the social norms and strategies of IR [Smit and Santos, 2024]. This combined mechanism is then applied to a multi-agent system of reinforcement learning agents, and the fairness, as well as the cooperation, of the resulting system is examined. While there is precedent in the IR community for examining IR in group-structured populations [Kawakatsu *et al.*, 2024], the approach is typically taken that either reputation information [Hilbe *et al.*, 2018] or the social norm itself [Chalub *et al.*, 2006] is not (or partially) shared between groups, and not that the norm *itself* distinguishes between interactions within a group and those between groups. Similarly, while IR applied to RL systems has been examined [Anastassacos *et al.*, 2021], the combination of group-structure and RL remains underexplored, particularly in regards to fairness.

2 Contributions

In my work so far, I assume that the population of agents is partitioned into two groups, that group membership is visible, and that there may be differences between groups such as likelihood to make mistakes when acting. The agents periodically engage in a pairwise donation game in a well-mixed manner. The fundamental question then becomes: "If the context space C contains a binary variable indicating whether the agents are in the same group, which social norms lead to the highest levels of cooperation *and* fairness?", where I define fairness as the demographic parity ratio i.e. the ratio between the payoffs of the best- and worst-off group.

I show that if agents' strategies can also discriminate based on the aforementioned binary group-relation variable, then a large number of social norms can stabilise cooperation among a majority group, but that this cooperation often does not extend fully to the minority group.

While previous work utilised only tools from evolutionary game theory, when applying IR to artificial agents, the method of adaption can be individual learning. As such, I additionally show that when agents adapt with tabular Q-learning instead of a birth-death process representing social learning, the expected level of cooperation in a group-structured population almost universally falls, and that social norms that previously led to fair outcomes can now produce in-group bounded cooperation depending on the initial conditions of the system.

3 Discussion and Future Work

While the model described above is purely theoretical, the idea of group-based discrimination combining with reputation-based discrimination can be seen in many real-life applications, as described in the previous AirBnB example. In a work in progress paper, I develop the current model of IR in group-structured populations to such online platforms to explore how to design better rating systems for all users, including how these systems can adjust for inherent biases in agents' decision making.

Additionally, I hope to expand the application domain of IR and fairness by studying more complex models that introduce additional variables such as spatial dimensions, explicitly model reputation spreading and coordination via gossip, and having agents strategically report their experiences. By introducing these additional parameters and mechanisms, I hope to capture a much broader set of behaviours and emergent phenomena and, in turn, examine how increasingly complex, real systems can be designed in a more equitable manner.

References

[Abrahao *et al.*, 2017] Bruno Abrahao, Paolo Parigi, Alok Gupta, and Karen S. Cook. Reputation offsets trust judgments based on social biases among Airbnb users. *PNAS*, 114(37):9848–9853, September 2017.

[Anastassacos *et al.*, 2021] Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. Cooperation and reputation dynamics with reinforcement learning. In *Proceedings of AAMAS 2021*, pages 115–123, 2021.

[Chalub *et al.*, 2006] F. A. C. C. Chalub, F. C. Santos, and J. M. Pacheco. The evolution of norms. *Journal of Theoretical Biology*, 241(2):233–240, July 2006.

[Edelman *et al.*, 2017] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22, April 2017.

[Fatima *et al.*, 2024] Shaheen Fatima, Nicholas R. Jennings, and Michael Wooldridge. Learning to Resolve Social Dilemmas: A Survey. *J. Artif. Intell. Res.*, 79, 2024.

[Hilbe *et al.*, 2018] Christian Hilbe, Laura Schmid, Josef Tkadlec, Krishnendu Chatterjee, and Martin A. Nowak. Indirect reciprocity with private, noisy, and incomplete information. *PNAS*, 115(48):12241–12246, 2018.

[Kawakatsu *et al.*, 2024] Mari Kawakatsu, Sebastián Michel-Mata, Taylor A. Kessinger, Corina E. Tarnita, and Joshua B. Plotkin. When do stereotypes undermine indirect reciprocity? *PLoS Comput. Biol.*, 20(3), 2024.

[Ohtsuki and Iwasa, 2004] Hisashi Ohtsuki and Yoh Iwasa. How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.*, 231(1):107–120, 2004.

[Santos *et al.*, 2021] Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. The complexity of human cooperation under indirect reciprocity. *Phil. Trans. R. Soc. B*, 376(1838), 2021.

[Smit and Santos, 2024] Martin Smit and Fernando P. Santos. Learning fair cooperation in mixed-motive games with indirect reciprocity. In *Proceedings of IJCAI 2024*, 2024.