Learning Fair Cooperation in Systems of Indirect Reciprocity

Jacobus Smit
University of Amsterdam
Amsterdam, Netherlands
jacobus.smit@uva.nl

Fernando P. Santos
University of Amsterdam
Amsterdam, Netherlands
f.p.santos@uva.nl

ABSTRACT

In a multi-agent setting, altruistic cooperation is costly yet socially desirable. As such, reinforcement learning agents can struggle to converge to efficient, cooperative policies. Indirect reciprocity (IR), whereby agents are given a chance to discriminate based on prior actions of others, is a mechanism that can stabilise cooperation. IR has been used to investigate the reputation rules that stabilise cooperation in homogeneous populations. However, in heterogeneous social systems, discrimination based on prior actions can supplement discrimination based on arbitrary (protected) characteristics, leading to unfair outcomes. In this work-in-progress paper, we propose a new model to investigate cooperation and fairness in systems of indirect reciprocity. As in previous IR literature, agents asynchronously play a donation game in which reputations and strategies co-evolve. Strategies and reputation assignment can however discriminate based on static group labels. We analyse a multi-agent system where strategies are adopted through reinforcement learning (Q-learning). We aim at identifying the settings where both cooperation and fairness emerge. In our preliminary analysis, we show that, in line with previous literature, imposing specific social norms (e.g., stern-judging or simple standing) allows cooperation to be learnt. Cooperation is not always fair, though: adding a group identity layer opens the door for inequality to emerge, as we highlight in some exemplifying scenarios. We plan to use this framework to comprehensively analyse intervention mechanisms that induce learning both high levels of cooperation and fairness.

KEYWORDS

Indirect Reciprocity, Cooperation, Evolutionary Game Theory, Reinforcement Learning, Fairness

1 INTRODUCTION

Cooperation is a fundamental research topic across disciplines [6, 20]. While cooperative populations tend to thrive, individuals are tempted to act selfishly, receiving the benefits of others' cooperation without exerting the effort themselves. The conundrum underlying this interaction is evident if we formally translate it into the so-called *donation game*, whereby a donor decides whether to pay a cost c to offer a benefit b to a recipient. Assuming that $b > c > 0$, this simple interaction illustrates the ubiquitous social dilemma of altruistic cooperation. Understanding how to engineer cooperation in these settings is a fundamental scientific challenge [18, 20] and a key frontier in artificial intelligence research [4, 17].

In artificial intelligence, particularly in the context of distributed and multiagent systems, research has focused on the design of autonomous systems where cooperation is stable [7]. Without explicitly designing to encourage cooperation, the individual costs of cooperating can cause free rider problems and cooperation to subsequently vanish over time. In such contexts, it is fundamental to understand how adaptive agents can learn, over time, to cooperate. The cooperation mechanisms observed in human societies [13, 20] can accordingly inspire formal methods to stabilise cooperation in groups of artificial agents.

Indirect reciprocity. One particularly effective mechanism to sustain cooperation among humans is indirect reciprocity (IR) [14, 16, 25]. Within such a framework, agents are assumed to strategically discriminate, and provide benefits, based on the social standing of others which encapsulates the social judgements of their previous actions. A central challenge in this domain is thereby understanding how reputations should be assigned for cooperation to be maximised. Previous work has shown that only a small set of social norms (i.e., rules followed to assign reputations) are able to stabilise cooperation in populations of homogeneous agents [15]. These norms have also been tested in the context of reinforcement learning agents [1].

In general, prior works also assumed that agents are only distinguished through their actions and reputations. In real settings, however, agents may be distinguishable by certain traits that give rise to specific group identities. This raises the question of how discrimination based on reputations might be affected by discrimination based on group identities. Tag-based cooperation [2, 28] and indirect reciprocity have been studied independently as mechanisms of cooperation, but the implications of combining the two are yet unexplored. The importance of such combinations has been raised in reports such as Efferson and Fehr [5]. This paper is related to similar efforts to combine cooperation mechanisms (such as direct and indirect reciprocity [26]) by combining reputation and group (or tag) based cooperation. The connection between reputation-based and group-based discrimination has been recently discussed in the context of social psychology [10, 21] and evolutionary biology [27]. Nevertheless, it remains under-explored how reputations and group identities might affect cooperation in groups of reinforcement learning agents.

1.1 Contribution

With this work-in-progress paper, we aim to show, through an exploratory analysis, how unfair discrimination can evolve in groups of reinforcement learning agents. By unfair cooperation we mean cooperation through discrimination based on an arbitrary group label. These results will be used as a basis to test new mechanisms that can sustain fair cooperation, which only discriminates based on prior actions by agents.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 9-10, 2023, London, UK, https://alaworkshop2023.github.io. 2023.

We develop a new model where agents play a *donation game* in two different roles (*donor* and *recipient*), and receive payoffs according to the rules introduced above (cooperation cost c by a donor, and cooperation benefit b when received by a recipient; no costs nor benefits are distributed when a donor decides to defect). We start by noting that applying reinforcement learning in this context faces a fundamental challenge: agents only have costs when playing in the role of donor, and benefits of acting cooperatively are only accrued at a later stage, when playing as a recipient. As we show, this requires re-formulating the way by which rewards are used to update Q-values (in the context of Q-Learning).

Furthermore, we assume that agents can discriminate both based on reputation and an agent's arbitrary group "label". This label is randomly assigned to each agent in the beginning of our simulations, such that a predetermined proportion of agents in the population have each label. Contrary to reputations, which are dynamic, group labels are static to each agent regardless of their actions. Such a label may be informative about their propensity or ability to cooperate through heterogeneity in the likelihood of committing errors, or be a "red herring" and simply a distraction to stifle cooperation or cause unfairness.

1.1.1 Preliminary findings. In the preliminary results we present, we are able to show that, as in prior works on evolutionary game theory [8, 15, 19, 22, 24, 25, 29], in the context of reinforcement learning there are also so called "leading" rules to assign reputations (i.e., social norms) that allow agents to learn cooperation. By introducing group identities, however, we show that agents can learn more quickly to cooperate with members of a certain group, even if the norm governing the population is group-agnostic, when one group is a majority. On the other hand, we show that introducing norms that assign reputations based on group identities (e.g., determining that a good reputation is only deserved when cooperation happens with a recipient belonging to the same group as the donor) leads to unfair cooperation. These results will provide the basis to investigate the following:

(1) Group-dependent norms to allow cooperation to evolve at equal rates regardless group identity,
(2) Mechanisms to reinstate universal cooperation even in the presence of unfair group-dependent norms.

1.1.2 Structure. The paper is structured as follows: first, in the remainder of this section, we discuss previous work related to cooperation and indirect reciprocity in humans and multi-agent systems. In section 2 we introduce our model of indirect reciprocity with heterogeneous agents and group labels, and perception and execution errors. In section 3, we translate this model into a multi-agent system where strategies are learned with a Q-learning approach. In section 4 we show that the policies learned in the multi-agent model are aligned with those predicted by previous theoretical models, and that the multi-agent model provides additional insights into the dynamics of learning. Finally, in section 5 we discuss the implications of our preliminary results so far and how we will develop them in future.

1.2 Related Work

1.2.1 The cooperation dilemma. Explaining and inspiring cooperation among humans is a fundamental research topic across disciplines [6, 18, 20]. In the field of AI, there is a growing interest in expanding the ability for AI to interact with and contribute more directly to society. In a recent commentary [4], the authors argue that AI requires "social understanding" to achieve success in tasks that require complex interactions such as navigating pavements, financial markets, and online communication. Many tasks that AI engage with also require cooperation with humans or other AI and so recent works have explored mechanisms to help enable cooperation. The proposed methods include introducing inequality-averse agents who pro-socially punish defectors [9], intrinsic motivations [11], an introspective self-play mechanism [1], or non-adaptive agents playing a fixed pro-social strategy [1, 23].

1.2.2 Indirect reciprocity and multi-agent reinforcement learning. Another recent paper which explores how indirect reciprocity (IR) can be incorporated into Q-learning is Anastassacos et al. [1]. Instead of a norm being predetermined, the goal of the paper is to examine how agents can establish an effective reputation mechanism by themselves. To do so, they must collectively learn and come to a consensus about both the social norm and the interpretation of agents' reputations. To aid in this, the researchers propose seeding the population with fixed agents and introspective self-play, where agents evaluate their own strategy against themselves. They find that a combination of both mechanisms can sustain cooperation.

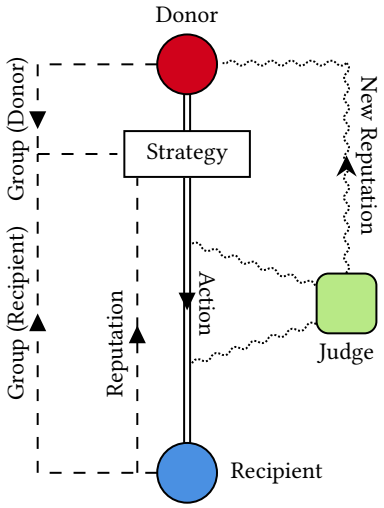
While our model also has Q-learning agents, one key difference is the game being played. In the model of Anastassacos et al. [1], players take on the role of donor and recipient *simultaneously*, making learning more consistent and simpler than in our model. In section 3.2 we discuss how we adapt Q-learning to the setting of delayed rewards and how its impact can be seen in our observations.

Moreover, while the techniques used are similar, the focus of our work is very different. While in Anastassacos et al. [1] the goal is to internalise the reputation mechanism and examine its effects on learning cooperation, we take the norm to be an external constant, introduce another variable agents can discriminate with respect to, and see how inequality can emerge in spite of cooperation.

1.2.3 Indirect reciprocity and group structured populations. Some prior works studied IR in populations where agents explicitly belong to groups, through the lens of evolutionary game theory. In this domain, Kessinger et al. [12] assume that different groups might use different social norms and focus on the effect of different information broadcasting mechanisms, whereby information about individuals can spread only between members of the same groups or publicly (as in traditional models). Contrarily to the setting we explore here, [12] assumes that strategies only discriminate based on reputations and not group identity. The authors find that in such systems cooperation ultimately depends on the rate of in/out-group interactions and cooperation can collapse if information remains within the same groups.

In a more recent work, also considering the interplay between group-structured populations and reciprocity, Stewart and Raihani [27] study how stereotypes might be formed through group reciprocity: the authors find that stereotyping can lead to *negative*

(a) Diagram exemplifying an interaction between agents.



(b) Walkthrough of the interaction in 1a.

Player	Donor	Recipient
Group	Red	Blue
`Rel`	—	Out-group (`O`)
`Rep`	—	Good (`G`)
Strategy	1100_2 (aka `Disc`)	—
`Act`	$^1\mathtt{O}, \mathtt{G}^{\mathtt{O}} = {}^1\mathtt{0}, 1^{\mathtt{O}} = 10_2 = 2$	
	$\Rightarrow$ $1100_2 \gg 2¼ = 1 = \mathtt{C}$	
Norm	11000011_2 (aka *stern-judging*)	
Judgement	$^1\mathtt{O}, \mathtt{G}, \mathtt{C}^{\mathtt{O}} = {}^1\mathtt{0}, 1, 1^{\mathtt{O}} = 110_2 = 6$,	
	$\Rightarrow$ $11000011_2 \gg 6¼ = 1 = \mathtt{G}$	

Figure 1: An illustration (1a) and tabular walkthrough (1b) (using notation from Table 1 to denote group relations (`Rel`), reputations (`Rep`), and actions (`Act`)) of an interaction between two agents which is observed by a judge. The donor's Strategy determines the Action taken based on information about the Groups of the Donor and Recipient, in this case their relation to each other, and the Reputation of the donor. The Judge then observes the Action and the information used to determine it in order to assign the Donor a New Reputation based on the society's social norm. In this example, the donor plays `C`, and the judge assigns the new reputation is `G` based on the so-called *stern-judging* norm.

judgement bias in which individuals become pessimistic about the willingness of out-group members to cooperate.

Although these works study dynamics of IR under reinforcement learning [1] and dynamics of reciprocity associated with group identities [27] the combination of indirect reciprocity, group identity and reinforcement learning remains under-explored. In this paper we propose a model that contributes to fill this gap.

Table 1: Boolean encoding, names and abbreviations for information used by players and judges.

Boolean type	`abbrev.`	"False" value		"True" value	
Group relation	`Rel`	Out-group	`O`	In-group	`I`
Reputation	`Rep`	Bad	`B`	Good	`G`
Action	`Act`	Defect	`D`	Cooperate	`C`
Group		Blue		Red	

2 MODEL

We consider a well-mixed population where agents interact pairwise by playing a donation game. A donation game is characterised by parameters b and c where $b > c > 0$, and has one player taking on the role of donor, while the other is the recipient. The donor has the opportunity to "cooperate": paying a cost c to confer a benefit b to the recipient, who itself has no action to take.

Given that cooperation (referred to `C`) is costly to the donor, defection (referred to as `D`) is the dominant strategy in the one-shot form of this game. In order to encourage players *not* to play their dominant strategy, donors' strategies are allowed to discriminate based on the current reputation of the potential recipient. This reputation is binary, and is determined by an external judge who examines the interactions between agents and determines in each case whether the action taken by the donor should confer the donor a good (`G`) or a bad (`B`) reputation.

Novel to this paper, we allow donors' strategies to discriminate based on the group identifier of their game partner, particularly, whether their potential recipient is in their in-group (`I`) or out-group (`O`). In this work we have two groups which we refer to as "red" or "blue", and these identifiers are unable to be changed once assigned. Moreover, the proportion of red agents is a parameter p_{Red} of our model with $p_{\text{Blue}} = 1 \quad p_{\text{Red}}$. For consistency we assume that the red group is the majority in the population i.e. $p_{\text{Red}} > 0.5$ but note that this decision is arbitrary.

The role of the judge is to update the reputations of donors after each interaction. The judge determines the goodness of actions by the *social norm* which governs the society (see Table 2 for examples of norms). A social norm is a logical function N of the group relation of the two individuals (`O` or `I`), the recipient's reputation (`B` or `G`), and the donor's action (`D` or `C`), and returns the donor's new reputation (`B` or `G`). By encoding these inputs as Booleans, as detailed in Table 1, we can write:

$$N : \mathtt{Rel} \quad \mathtt{Rep} \quad \mathtt{Act} \; ! \; \mathtt{Rep}, \tag{1}$$

$$N : \mathtt{f0, 1g}^3 \; ! \; \mathtt{f0, 1g} \tag{2}$$

Noting that the domain and range of the function has a finite number of elements, we can enumerate all possible social norms and assign one a unique integer value. Take the input $^1\mathtt{I, B, D}^{\mathtt{O}} = {}^1\mathtt{1, 1, 0}^{\mathtt{O}}$, we can concatenate[1] the digits to give $011_2 = 3$, and then store the output of the norm at this input in the 3rd digit of a (0-indexed) binary number N_2 i.e. $N\,^1\mathtt{1, 1, 0}^{\mathtt{O}} = N_2 \gg 001_2¼$. Enumerating all possible inputs from 000_2 to 111_2 gives rise to an 8-bit integer for every possible norm. Figure 1 exemplifies a pairwise interaction.

[1] Concatenation reverses the order of the digits as the first digit of a number is on the right, but the first dimension of an array is indicated by the left-most index.

2.1 Perception and execution errors

So far, agents have been infallible in their ability to assess information and execute their intended action, but, similarly to the exploration rate ϵ in Q-learning, Mistakes are both realistic and required in an analytic IR model to ensure that the space of reputations is properly explored[2].

An agent subject to an execution error $\epsilon \in [0, 1]$ will play the opposite strategy than they intended with probability ϵ. Expanding on the assessment errors in [3], an agent with perception error $\delta = (\delta_{\text{Rel}}, \delta_{\text{Rep}}) \in [0,1]^2$ will, independently of other bits, perceive the opposite information in a certain bit i with probability δ_i.

2.2 Evolutionary stability

Given the model introduced above, a natural question one can pose is: Given a specific norm, which strategies are more likely to end up being played by agents? One way this can be answered is with evolutionary game theoretical (EGT) tools. A key concern of EGT is that of the evolutionary stability of strategies: as agents play the donation game, they gain and lose utility based on their strategy and its impact on their reputation. Assume that strategies S_{Red} and S_{Blue} have proliferated the entire incumbent population of red and blue agents respectively, and that the population is governed by social norm N. By calculating the expected utility of a player of both groups playing their respective strategy (S_{Red} or S_{Blue}), we can determine whether a random strategy mutation in either group could outperform the incumbents of that group. One can derive the expected utilities of incumbents and mutants and determine whether the norm-strategy-strategy triple $(N, S_{\text{Red}}, S_{\text{Blue}})$ is an *evolutionarily stable state* (ESS).

Stronger than the traditional Nash equilibrium, a strategy S_I is evolutionarily stable on the condition that if any alternative strategy S_M arises in a group that S_I has proliferated and that the proportion of agents playing this alternative is sufficiently small, then this alternative strategy will perform worse than the incumbent strategy S_I and die out. We say that a triple is an ESS if both of its strategies are evolutionarily stable.

While EGT and an ESS analysis is informative in terms of which strategies are more or less stable under each norm, we still need to understand 1) how prevalent each equilibrium point is and 2) how likely a population of learning agents is to converge to a certain ESS. For both purposes, we can use reinforcement learning.

3 REINFORCEMENT LEARNING MODEL

Rather than learning through imitation found in typical EGT models, it is interesting to examine whether a finite population of stochastic agents whose strategies are learned over time through interactions would converge to cooperative sets of strategies (eventually the same states forming ESS strategies).

3.1 Obstacles to learning

Beyond the difficulty inherent to agents simultaneously learning to coordinate, previous IR work has shown that, even in finite population models with imitation-based learning, larger exploration rates causes a breakdown in cooperation [24]. This has implications for any reinforcement learning (RL) approach, where exploration is initially very common as it is necessary for learning the quality of an action in a given state.

Another potential source of difficulty is the delayed nature of rewards in indirect reciprocity, as the rewards of acting as a donor are only later (and indirectly) obtained as a recipient. The problem of learning with sparse and delayed rewards is a well-known challenge in reinforcement learning. Delayed rewards in RL models with some stochastic aspect can cause uncertainty in the attribution of utility to actions e.g. how to know that some actions lead to a goal if it is not clear how the goal is reached? Even in deterministic systems, delayed rewards tends to lead to slower learning and therefore more exploration making cooperative equilibria in IR models less stable.

In the case of our model, actions are only taken by donors, and they can only understand the effects of their actions in subsequent interactions where they are the recipient. This means that, without sufficient exploration permitted, an initially cooperative agent could play as a donor multiple times in a row and reinforce the idea that cooperating is a costly endeavour which provides less utility than defecting. In this case the agent would converge towards a policy of unconditional defection (known as AllD), which is always stable and has a larger basin of attraction the smaller the level of exploration [24]. However, the fact that reputations are *memoryless* in the sense that they are overwritten by actions and we don't allow for strategies that take an agent's previous reputation into account (as in [25]) the current state (information) in our model has no implications on past or future states. This means utility will, in absence of perception errors, be attributed to the correct action.

3.2 Q-learning for IR Agents

With this in mind, we carefully apply a Q-learning algorithm similar to [1] . Action selection is standard: at each encounter, a donor's policy π chooses the best action given the information about the recipient with probability $1 - \epsilon$, and chooses either D or C uniformly at random $\mathcal{U}(\{C, D\})$ otherwise i.e.

$$\pi(s) = \begin{cases} \text{argmax}_{a \in \{C,D\}} Q(s, a), & \text{with probability } 1 - \epsilon \\ \mathcal{U}(\{C, D\}), & \text{with probability } \epsilon \end{cases} \tag{3}$$

Following a typical tabular Q-Learning approach, each agent maintains a 2×4 Q-table where $Q(a, s)$ is the Q-value associated with action $a \in \{0, 1\}$ subject to information $s \in \{0, 1, 2, 3\} = \{ij_2 : i \in \text{Rel}, j \in \text{Rep}\}$. After an interaction, whether or not a donation occurs, the Q-table of *both* players is updated by the equation

$$Q(a, s) \leftarrow Q(a, s) + \gamma(\mu - Q(a, s)) \tag{4}$$

where μ is the (possibly negative or zero) utility incurred in the interaction and $\gamma \in (0, 1]$ is the learning rate. A donor who cooperates will receive $\mu = -c$, discouraging cooperation in the short term. For a recipient, we set a to be the last action they took as a donor. Importantly, Q-values decay for both players even when the interaction causes no utility to be gained or lost because the donor chose to defect ($\mu = 0$ for both players.

For our tests, we had a model of 50 learning agents with $\gamma = 0.01$, $\epsilon = 0.1$, and initialised the Q-table (somewhat arbitrarily) such that

$$Q(a, s) \overset{i.i.d.}{\sim} \mathcal{N}\left(\frac{b_{\text{Red}} - b_{\text{Blue}}}{2}, 1\right), \tag{5}$$

[2]Execution errors must be strictly non-zero to satisfy the existence and uniqueness conditions of solutions to the reputation dynamics equations in section 2.2.

where b_{Red} and b_{Blue} are the benefit conferred to someone in the majority and minority groups respectively. We then ran the model for 5,000 episodes, each consisting of $50^2 = 2500$ interactions each such that, on average, every agent would interact with every other agent once every episode. In each interaction, the donor is chosen deterministically (for performance reasons), but the donor chooses another agent to be the recipient entirely at random. The proportion of red agents was 0.9, and we set $b = 8, c = 1$.

4 OBSERVATIONS

4.1 Cooperation

The norm governing the population has a deciding effect on the level of cooperation in a society. We find this to clearly be the case as we see in Figure 2, where three of the four established norms lead to cooperation, but *image scoring* (IS) does not. Further to this, the figure also reveals that the rate at which the policy is learned is also affected by the governing norm. We see that while cooperation is learned quickly under SJ, it is far slower under *shunning* (where the only "good" action is cooperation with a good player).

4.2 Fairness

Our model permits unfair norms that discriminate based on label. In Figure 3 we can see two examples of these norms: 208 and 224. These norms are each one bit different to *shunning* (which is 192). Norm 208 flips the bit for cooperating with bad outsiders from bad to good, meaning it is *out-group biased*, whereas Norm 224 does the same for cooperation with bad insiders, being *in-group biased*.

In the case of Norm 208, we can see that the majority group quickly learn to cooperate only in cases where it will provide them a good reputation. Yet, despite the norm, the minority group is not able to learn to cooperate in any situation and quickly becomes overrun by unconditional defectors.

However, the emergent effect we see for Norm 224 is rather unexpected. The figure shows us that the probability of majority-majority cooperation if the recipient is bad is close to 50%. This is due to the fact that half of the majority population have a policy close to Disc, and the other half only defect against bad outsiders. Looking closer at the Q-values in Figure 4, we can see that they are roughly evenly scattered to the left and right of the line $y = x$, explaining the previous probability of cooperation. This is a stable state that wouldn't have been discovered through EGT alone as the method of discovery in this case is a simple enumeration where each population is assumed to *all* play the same strategy.

Due to the majority-minority interaction rate being lower than majority-majority rate, we see that out-group interactions are learned at a slower rate. This slower effective learning rate means that there is an asymmetric learning rate for majority-minority interactions: minority agents (after a much shorter amount of time) know exactly what to do when interacting with a majority agent, but majority agents haven't had the opportunity to learn a definitive policy yet.

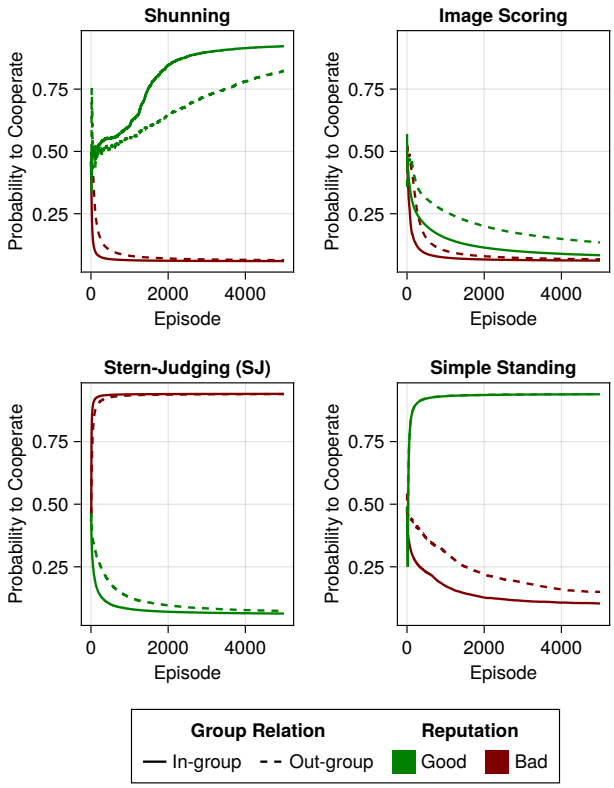


Figure 2: Time-series representing the probability that an agent belonging to the majority group cooperates with an agent of the same group (full-lines), an agent of the out-group (dashed-lines), an agent with a good reputation (green) or an agent with a bad reputation (red). Here we represent four well-known (fair) norms, that do not explicitly discriminate based on group identity. Each norm is able to sustain cooperation to varying degrees. In the long run, *stern-judging* (SJ), *simple-standing* (SS) and *shunning* (SH) lead to high levels of cooperation, although SJ and SS induce agents to learn cooperation faster. *Image scoring* (IS) fails to steer agents into cooperation. These observations match previous well-known results in the context of evolutionary game theory, although here agents adapt through individual reinforcement learning. Although here we only represent fair norms, we can already observe that agents might learn to cooperate faster with members of the in-group (e.g., compare the green dashed and full lines in SH)

5 CONCLUSION AND DISCUSSION

Here we propose a new model combining indirect reciprocity [14, 15, 25] with group-identity, in a setting where agents adapt through multi-agent reinforcement learning. We observe that cooperation emerges under typical norms if recipients are able to use their rewards to update the Q-values of actions played as donors. Furthermore, we observe that unfair norms – assigning reputations in ways that discriminate based on both actions and group-identity – trigger biased cooperation: agents learn to cooperate only with in-group members. This work connects multi-agent reinforcement learning, cooperation and ongoing discussions related with fairness in AI systems. In the field of algorithmic fairness there is a

Name	Encoding	(O,B,D)	(I,B,D)	(O,G,D)	(I,G,D)	(O,B,C)	(I,B,C)	(O,G,C)	(I,G,C)
All bad	0	0	0	0	0	0	0	0	0
Shunning	192	0	0	0	0	0	0	1	1
Stern-judging	195	1	1	0	0	0	0	1	1
Simple standing	207	1	1	0	0	1	1	1	1
Image score	240	0	0	0	0	1	1	1	1
Norm-208	208	0	0	0	0	1	0	1	1
Norm-224	224	0	0	0	0	0	1	1	1

Table 2: Norms are used by observers (e.g., a Judge) to assign reputations to individuals based on their actions and opponents' characteristics. We represent norms by their binary encoding, concatenating their outputs in all possible contexts (see Table 1 for notation). The named norms are "fair": each consecutive pair of entries (covering the same input except group relation) is the same. The unnamed norms (208 and 224) show either in or out-group bias, the effects of which can be seen in Figure 3.

growing interest in understanding discrimination from the perspective of benefits precluded based on protected characteristics. In this context, different fairness metrics have been formalised to capture how agents are treated differently based on possibly arbitrary group identifiers. We have shown that introducing such identifiers may complicate the maintenance of cooperation and the rate at which cooperation develops in the context of indirect reciprocity

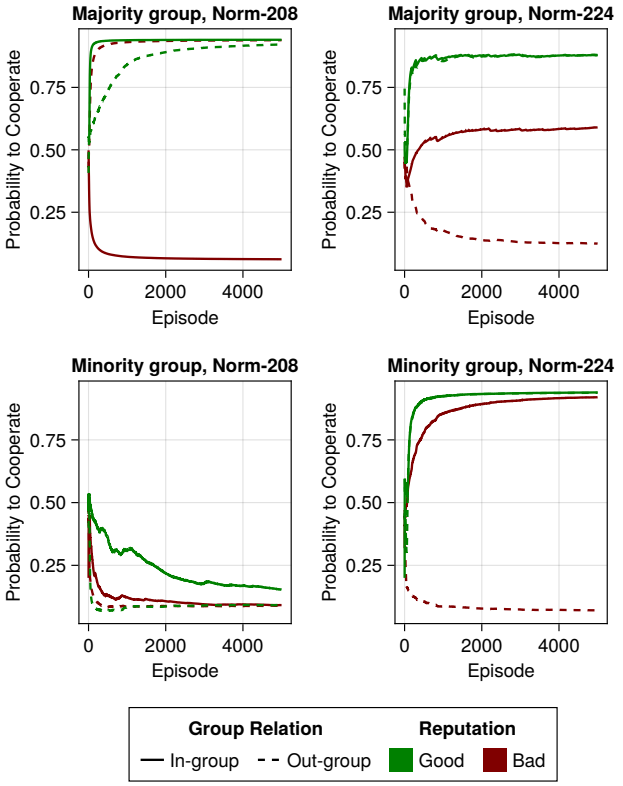


Figure 3: Unfair norms cause unfair outcomes. With Norm-224 (see Table 2) agents learn to prefer in-group (over out-group) cooperation (i.e., dashed and full lines do not match).

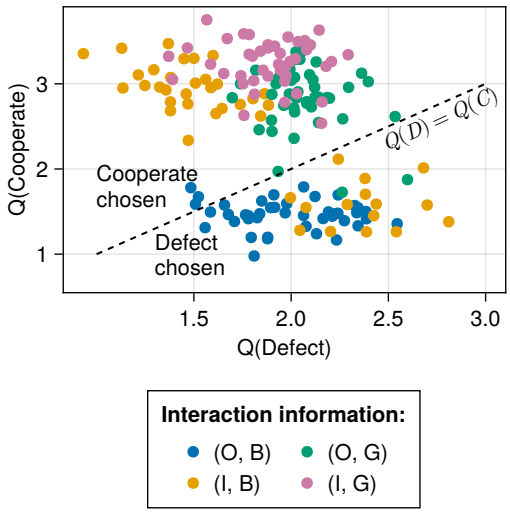


Figure 4: Final Q-values of majority players under Norm 224.While most information has its Q-values cleanly to one side of the line $y = x$, the (in-group, bad) markers are spread evenly on either side of the line indicating that a roughly equal proportion will cooperate or defect with bad insiders.

and multi-agent reinforcement learning, even when reputations are assigned *without* considering the identifier.

Despite this, there are many other angles by which to view this model which raises more questions. In future work, it would be interesting to test which other variables in our model impact learning fair cooperation and which do not have a large effect. This would provide a set of guidelines for more complex reinforcement learning models to follow when trying to encourage fair cooperation.

Finally, by using the previous extension or otherwise, we aim to study intervention mechanisms to counteract the unfairness brought upon by unfair norms or other inequalities such as cost, benefit, or group size.

REFERENCES

[1] Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. https://doi.org/10.48550/arXiv.2102.07523 arXiv:2102.07523 [cs].

[2] Tibor Antal, Hisashi Ohtsuki, John Wakeley, Peter D Taylor, and Martin A Nowak. 2009. Evolution of cooperation by phenotypic similarity. *Proceedings of the National Academy of Sciences* 106, 21 (2009), 8597–8600.

[3] Hannelore Brandt and Karl Sigmund. 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology* 231, 4 (Dec. 2004), 475–486. https://doi.org/10.1016/j.jtbi.2004.06.032

[4] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593, 7857 (May 2021). https://doi.org/10.1038/d41586-021-01170-0

[5] Charles Efferson and Ernst Fehr. 2018. Simple moral code supports cooperation. *Nature* 555, 7695 (March 2018), 169–170. https://doi.org/10.1038/d41586-018-02621-x Bandiera_abtest: a Cg_type: News And Views Number: 7695 Publisher: Nature Publishing Group Subject_term: Human behaviour.

[6] Ernst Fehr and Urs Fischbacher. 2003. The nature of human altruism. *Nature* 425, 6960 (Oct. 2003), 785–791. https://doi.org/10.1038/nature02043

[7] Michael R. Genesereth, Matthew L. Ginsberg, and Jeffrey S. Rosenschein. 1986. Cooperation without communication. In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence (AAAI'86)*. AAAI Press, Philadelphia, Pennsylvania, 51–57. https://doi.org/10.1016/B978-0-934613-63-7.50026-7

[8] Christian Hilbe, Laura Schmid, Josef Tkadlec, Krishnendu Chatterjee, and Martin A. Nowak. 2018. Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences* 115, 48 (Nov. 2018), 12241–12246. https://doi.org/10.1073/pnas.1810565115

[9] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html

[10] Hirotaka Imada, Angelo Romano, and Nobuhiro Mifune. 2022. Dynamic indirect reciprocity; When is indirect reciprocity bounded by group membership? *Evolution and Human Behavior* (Dec. 2022). https://doi.org/10.1016/j.evolhumbehav.2022.12.001

[11] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*. PMLR, 3040–3049.

[12] Taylor A. Kessinger, Corina E. Tarnita, and Joshua B. Plotkin. 2022. Evolution of social norms for moral judgment. https://doi.org/10.48550/arXiv.2204.10811 arXiv:2204.10811 [q-bio].

[13] Martin A. Nowak, Akira Sasaki, Christine Taylor, and Drew Fudenberg. 2004. Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428, 6983 (April 2004), 646–650. https://doi.org/10.1038/nature02414 Number: 6983 Publisher: Nature Publishing Group.

[14] Martin A. Nowak and Karl Sigmund. 2005. Evolution of indirect reciprocity. *Nature* 437, 7063 (Oct. 2005), 1291–1298. https://doi.org/10.1038/nature04131 Number: 7063 Publisher: Nature Publishing Group.

[15] Hisashi Ohtsuki and Yoh Iwasa. 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* 231, 1 (Nov. 2004), 107–120. https://doi.org/10.1016/j.jtbi.2004.06.005

[16] Isamu Okada. 2020. A review of theoretical studies on indirect reciprocity. *Games* 11, 3 (2020), 27.

[17] Ana Paiva, Fernando Santos, and Francisco Santos. 2018. Engineering Pro-Sociality With Autonomous Agents. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.12215

[18] Elizabeth Pennisi. 2005. How did cooperative behavior evolve? *Science* 309, 5731 (2005), 93–93.

[19] Cedric Perret, Marcus Krellner, and The Anh Han. 2021. The evolution of moral rules in a model of indirect reciprocity with private assessment. *Scientific Reports* 11, 1 (Dec. 2021), 23581. https://doi.org/10.1038/s41598-021-02677-2 Number: 1 Publisher: Nature Publishing Group.

[20] David G. Rand and Martin A. Nowak. 2013. Human cooperation. *Trends in Cognitive Sciences* 17, 8 (Aug. 2013), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

[21] Angelo Romano, Daniel Balliet, and Junhui Wu. 2017. Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership? *Journal of Experimental Social Psychology* 71 (July 2017), 59–67. https://doi.org/10.1016/j.jesp.2017.02.008

[22] Fernando Santos, Jorge Pacheco, and Francisco Santos. 2018. Social Norms of Cooperation With Costly Reputation Building. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.11582 Number: 1.

[23] Fernando P Santos, Jorge M Pacheco, Ana Paiva, and Francisco C Santos. 2019. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6146–6153.

[24] Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. 2016. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Scientific Reports* 6, 1 (Nov. 2016), 37517. https://doi.org/10.1038/srep37517 Number: 1 Publisher: Nature Publishing Group.

[25] Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. 2021. The complexity of human cooperation under indirect reciprocity. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376, 1838 (Nov. 2021), 20200291. https://doi.org/10.1098/rstb.2020.0291 Publisher: Royal Society.

[26] Laura Schmid, Krishnendu Chatterjee, Christian Hilbe, and Martin A. Nowak. 2021. A unified framework of direct and indirect reciprocity. *Nature Human Behaviour* 5, 10 (Oct. 2021), 1292–1302. https://doi.org/10.1038/s41562-021-01114-8 Number: 10 Publisher: Nature Publishing Group.

[27] Alexander J. Stewart and Nichola Raihani. 2023. Group reciprocity and the evolution of stereotyping. *Proceedings of the Royal Society B: Biological Sciences* 290, 1991 (Jan. 2023), 20221834. https://doi.org/10.1098/rspb.2022.1834 Publisher: Royal Society.

[28] Arne Traulsen and Martin A. Nowak. 2007. Chromodynamics of Cooperation in Finite Populations. *PLOS ONE* 2, 3 (March 2007), e270. https://doi.org/10.1371/journal.pone.0000270 Publisher: Public Library of Science.

[29] Jason Xu, Julian García, and Toby Handfield. 2019. Cooperation with Bottom-up Reputation Dynamics. (2019).