Artificial Agents Facilitate Human Cooperation Through Indirect Reciprocity

Alexandre S. Pires[a,*] and Fernando P. Santos[a]

[a]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
ORCID (Alexandre S. Pires): https://orcid.org/0000-0002-0648-5702, ORCID (Fernando P. Santos):
https://orcid.org/0000-0002-2310-6444

Abstract. Indirect reciprocity (IR) is a key mechanism to explain cooperation in human populations. With IR, individuals acquire reputations which can be used by others when deciding to cooperate or defect: the costs of cooperation can therefore be outweighed by the long-term benefits of keeping a good reputation. Although IR has been studied assuming populations fully composed of humans, social interactions nowadays involve the ever-increasing presence of artificial agents (AAs) such as social bots, conversational agents or even collaborative robots. It remains unclear how IR dynamics will be affected once artificial agents co-exist with humans. Here we develop a theoretical model to investigate the potential effect of AAs, deployed with a fixed strategy, in the evolving cooperation levels observed in a population. We study settings where AAs are subject to the same reputation update rules as the remaining adaptive agents and settings where AAs have a fixed reputation. We show that introducing a small fraction of AAs with a discriminating strategy (i.e., cooperate only with *good* agents) increases the cooperation rate in the whole population. Moreover, the positive effect of AAs is exacerbated when these are unconditionally assessed as *good*. We also demonstrate the vulnerability of cooperation towards purely defecting AAs, and the inefficacy of non-discriminating cooperators in promoting cooperation. Our theoretical work contributes to identify the settings where artificial agents, even with simple hard-coded strategies, can help humans to solve a social dilemma of cooperation.

1 Introduction

Altruistic cooperation requires that individuals spend a cost (c) to provide a benefit (b) to others. When $b > c$, cooperation implies a social dilemma: cooperation is socially desirable yet, given the cost involved, refusing to cooperate is the dominant strategy. Explaining cooperation is a fundamental challenge across disciplines [15] and previous research has identified mechanisms to stabilize it [25]. Among these, indirect reciprocity (**IR**) stands as a primary mechanism to enable cooperation between unrelated individuals [27]. **IR** requires that interactions are observed and individuals assigned reputations [3], which spread through the population (e.g., via gossiping [12]). Simply put, under **IR** cooperating today might contribute to build a reputation that leads others to reciprocate tomorrow [48].

Research in **IR** spans many disciplines. This mechanism is intrinsically related to the evolution of morality, culture and was pointed as a crucial component of a cohesive social structure [3]. Importantly,

however, the viability of **IR** as a mechanism to sustain cooperation in hybrid populations – composed of humans and artificial agents (AAs) – remains unknown.

Humans are now increasingly co-existing with AI systems, particularly with socially interactive agents [23]. Examples of these include collaborative robots for navigation [37, 54] or education [44]. It is pressing to understand the impacts that AI systems will have on our collective behavior [2] and our ability to trust and cooperate with AI [7, 36]. Previous works have suggested the role of communication, embodiment [22] and the perception of facing an artificial agent as important aspects of cooperation [20] in hybrid populations. In the context of **IR**, however, one must identify the differences in how humans and artificial agents are assessed, as well as the role of AAs that discriminate in pre-defined ways to opponents' reputations.

In this article, we aim to provide a step in addressing two central questions related to **IR** in hybrid populations: **1) What is the impact of artificial agents introduced in a human population interacting under IR? 2) Which social norms promote cooperation in a system composed of humans and artificial agents?**

To answer these questions, we develop a theoretical model based on evolutionary game theory (EGT) [48] where a finite population composed of adaptive agents (representing humans) and AAs repeatedly play a *donation game*. In this game, an agent playing as donor can cooperate (C), that is, donate, or defect (D) with a receiver. As introduced above, cooperate means paying a cost c to concede a benefit b, where $b > c > 0$. In our model, detailed in Section 3, agents have reputations that are dynamically updated based on a social norm, i.e., rules that maps the action of the donor and the reputation of the receiver to a new reputation for the donor. While adaptive agents revise their strategy over time, AAs employ a pre-defined strategy, thus omitting implementation details and focusing solely on their resulting behavior [40, 17]. We also study the impact of biases against AAs, in the form of fixed reputations.

We show that introducing a small fraction of AAs whose actions are conditioned on reputations can trigger high levels of cooperation in settings where defection previous prevailed (i.e., low b/c ratio). This effect is further amplified if these AAs are perceived as *good*. Additionally, we observe that unconditional cooperative AAs are exploited, thereby unable to sustain cooperation, while anti-social AAs undermine all cooperative behavior in the population. These results contribute to identify the settings where AAs can help humans to solve a social dilemma of cooperation. [1]

---

* Corresponding Author. Email: a.m.dasilvapires@uva.nl

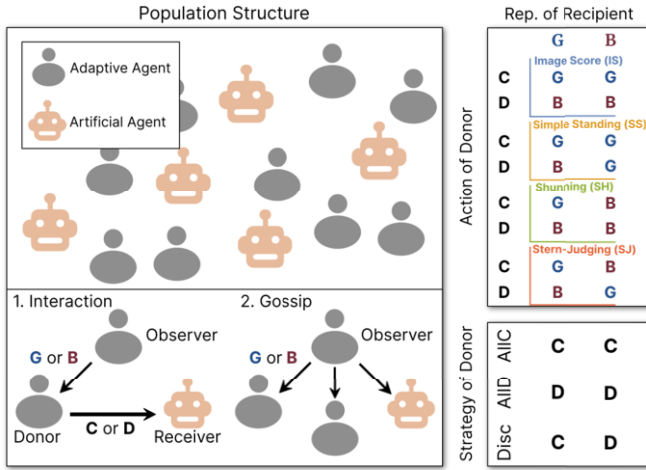[1] Code and supplementary material is available on [1].

Figure 1. We consider a hybrid population consisting of adaptive agents (representing humans), in gray, and artificial agents, in orange. We assume interactions involving three agents: a Donor, Receiver and Observer(s). The Donor chooses whether to cooperate or defect (C or D) with the Receiver. The interaction is witnessed by a third party, the Observer, who determines the new reputation (Good, **G**, or Bad, **B**) of the donor following a social norm. The reputation is shared by the Observer and assumed to spread to all other agents, becoming public knowledge. Top right: Reputations are assigned following a social norms that is fixed in a population and followed by all agents. Social norms determine the next reputation of a donor given the reputation of the recipient and the action of the donor. Bottom right: The three strategies considered: *AllC*, which always cooperates, *AllD*, that always defects, and *Disc*, that cooperates with **G** and defects against **B**.

2 Related Work

Indirect Reciprocity: A wide body of literature has studied the connection between **IR** and cooperation [26, 27]. Firstly, **IR** requires a social norm to discern between good and bad behaviors. Previous work has explored which social norms, and under which interaction contexts, can best sustain cooperation. More specifically, seminal work by Ohtsuki et al. has identified the leading norms able to stabilize cooperation [30, 31]; subsequent works have considered the role of cognitive complexity and emotional displays in **IR** [42, 11], as well as the impact of private reputation systems [19]. Research on **IR** has historically been focused on scenarios with infinitely large populations [49, 5], both for the convenience of mathematical analysis, and to discourage direct reciprocity (cooperation emergent from repeated play against the same individual). For a review of indirect reciprocity models see [32].

Fixed-strategy artificial agents: Previous works have considered the role of agents with a fixed policy in cooperation dynamics. These agents were named "seeding" [4] or "fixed-strategy" agents [43, 16]. Whenever such agents cooperate unconditionally, they were also referred to as "pathological altruists" [34]. Besides AAs, human resilient cooperators are also shown to increase overall cooperation in experimental settings [24].

Cooperation in hybrid populations: Following the increased prevalence of AI in everyday life, research on cooperation has now been considering an alternative scenario, where humans coexist among AAs forming *hybrid populations* and enabling *hybrid intelligence* [2]. Studies on hybrid populations seek to gain insight into the potential implications of deploying socially engaged AAs in human cooperation. Crucially, it remains unclear what their long-term consequences will be on a societal level – if these systems, even when

potentially cooperative, will reinforce or undermine cooperation efforts [36, 47, 9, 8]. Theoretical and experimental work has already been conducted in understanding the small scale effect that AAs have on human decision-making, often through investigating direct interactions between humans and AI-systems [33, 51, 22, 7, 18] or by having an autonomous system in place during human interactions [46]. These works highlight a complex relationship between the two parties, where the results of a human-AI interaction often differ greatly from those of a human-human or an AI-AI interaction, despite still showing promising results. Modelling the long-term and large scale effects of AI systems is, however, a more challenging task. To this end, frameworks like EGT and multi-agent systems are often applied to provide insight into the dynamics of such social systems [48].

Hybrid populations have been studied over a myriad of scenarios, such as two-player one-shot games [17], collective risk dilemmas [51] and the multiplayer ultimatum game [40]. Under these settings, AAs almost consistently improve the cooperation levels of the population, even in small numbers. We aim to complement these works by studying such hybrid systems under the context of indirect reciprocity, adapting the methods stemming from evolutionary game theory [55] to describe how human cooperation is influenced by the presence of AAs. Our conclusions also provide insight to guide future human-AI experimental works.

3 Methods

3.1 Donation game

We consider a finite population of size Z where, following previous indirect reciprocity models, we assume that agents play Donation Games with each other [30, 29, 39]. In a Donation Game, one agent (the donor) decides to cooperate **Cooperate (C)** or **Defect (D)** with another agent (the recipient). If the donor cooperates, the recipient receives a benefit (b) and the donor pays a cost (c, where $b > c$). If the donor defects, it incurs no cost and no benefit is given to the recipient. This game captures the essence of social dilemmas of cooperation: cooperation results in a socially desirable outcome, as the benefit of cooperation outweighs its cost; cooperators must, however, pay a cost and thus defection is the only dominant strategy in this game, likely to be implemented by rational agents.

The agents' decision to cooperate or defect depends on their action rules and reputations. Each agent has a (public) binary reputation, **Good (G)** or **Bad (B)**, which can be updated after each donation game (see Section 3.3). Moreover, agents adopt strategies to decide either to cooperate or defect. A strategy is a tuple $p = (p_G, p_B)$, where p_G and p_B represent the probability of cooperating with an opponent with reputation **G** or **B**, respectively. We focus on the dynamics between the three key pure strategies typically studied in indirect reciprocity: *AllC* ($p = (1,1)$), *AllD* ($p = (0,0)$), and the discriminator strategy *Disc* ($p = (1,0)$). While individuals using *AllC* and *AllD* will always opt for strategy C and D, respectively, therefore ignoring reputations, a *Disc* will only cooperate with **G** individuals.

Although we consider pure strategies, we allow for the presence of so-called *execution errors*: with probability ϵ, an agent fails to correctly execute the desired action. The execution error can easily be introduced by directly modifying any strategy p to $p^\epsilon = (p_G^\epsilon, p_B^\epsilon)$ where $p_i^\epsilon = (1 - 2\epsilon) \cdot p_i + \epsilon$.

3.2 Artificial agents

In a hybrid population, humans co-exist with arbitrarily complex artificial agents [13]. Although such agents can widely vary in their form

and abilities, we assume that they are characterized by two minimal properties: 1) their strategies can be hard-coded by design [40, 51] and 2) they will be perceived differently than human players [18], thereby being assigned different reputations. As a result, we consider **fixed-strategy agents (FSAs)**, playing with a fixed, hard-coded strategy, and **fixed-reputation-and-strategy agents (FRSAs)**, which, besides having a fixed-strategy, are also always perceived as having a designated reputation. The latter assumption captures the effects of either *techno-optimism* [10], whenever the reputations of the FRSAs is fixed at **G**, or *techno-skepticism*, when it is instead fixed at **B**.

When considering FSAs, all artificial agents will have the same strategy. When, instead, FRSAs are introduced, all AAs will have both the same strategy and reputation. We denote by $\mathcal{A}$ the number of FSAs (or FRSAs) in the population of size Z, such that $\mathcal{A} \leq Z$ as they are a part of the population, $f_p \in \{AllC, AllD, Disc\}$ the fixed strategy of any existing FSA, and $f_r \in \{\mathbf{G}, \mathbf{B}\}$ the predetermined fixed reputation, when applicable.

3.3 Reputation dynamics

We study second-order social norms [42] in hybrid populations, where the next reputation of an agent (donor) depends on its action (C/D) and the reputation of the recipient (G/B). Following previous models [39, 31, 32], we encode social norms as a tuple with four entries $d = (d_{GC}, d_{GD}, d_{BC}, d_{BD})$, where each entry corresponds to the probability of assigning a good reputation given the action of the donor and the reputation of the receiver.

We assume that reputation dynamics incurs assignment errors, α (an observer wrongly attributes the reputation following an observation) and assessment errors, χ (an individual incorrectly perceives the reputation of another agent). The assignment error can be included in each entry of a social norm, d_{ij}, by changing its value to $d_{ij}^{\alpha} = (1 - 2\alpha) \cdot d_{ij} + \alpha$. Assessment errors are added by impacting the probability that an individual with strategy p cooperates with an individual of a given reputation, which is given by $C_G^p = (1-\chi)p_G + \chi p_B$ and $C_B^p = \chi p_G + (1-\chi)p_B$ when cooperating with a good and a bad individual, respectively. Coincidentally, the probability of defecting against a good and a bad individual is given by $\bar{C}_G^p = 1 - C_G^p$ and $\bar{C}_B^p = 1 - C_B^p$, respectively.

Although there are 16 second-order social norms, we focus on four key norms, previously identified as capable of sustaining cooperation [26, 50, 30, 31, 38]: **Image Score (IS)**, $d = (1, 0, 1, 0)$, where cooperating is always considered good, and defecting is always bad; **Simple Standing (SS)**, $d = (1, 0, 1, 1)$, where only defecting against a good individual is considered bad; **Shunning (SH)**, $d = (1, 0, 0, 0)$, where only cooperating with a good individual is considered good; and **Stern Judging (SJ)**, $d = (1, 0, 0, 1)$, where both cooperating with good agents and defecting against bad agents is considered good, and the remaining is considered bad [21, 35]. As a sanity check, we also include the trivial norm **All Good**, $d = (1, 1, 1, 1)$, which essentially prevents any correlation between reputations and actions, precluding indirect reciprocity. A visualization of the population structure and the underlying social interaction, together with a summary of the social norms and strategies under consideration, is presented in Figure 1.

To illustrate the impact of **IR**, in Figure 1 of the Supplementary Material [1], we show how different social norms result in distinct cooperation levels for varying values of b, the donation benefit, without any AAs. We observe that, in general, increasing b increases cooperation; notwithstanding, some norms (e.g., Shunning and All Good) fail to sustain cooperation even for high levels of b ($b = 8$).

Given the assumptions above, the probability of assigning a good (public) reputation to an individual using strategy p after an interaction can be defined as

$$G_G^p = (1 - \chi)(C_G^p d_{GC} + \bar{C}_G^p d_{GD}) + \chi(C_G^p d_{BC} + \bar{C}_G^p d_{BD}), \tag{1}$$

when interacting with a good individual, and when interacting with a bad individual as

$$G_B^p = \chi(C_B^p d_{GC} + \bar{C}_B^p d_{GD}) + (1 - \chi)(C_B^p d_{BC} + \bar{C}_B^p d_{BD}). \tag{2}$$

Adapting the process described in [39] to three strategies, we now define the transition probabilities of a Markov chain that represents the possible distributions of good individuals, given a strategy state defined by a tuple $n_{ijk} = (n_i, n_j, n_k)$, where n_i, n_j and n_k represent the number of individuals currently using strategy *AllC*, *AllD* and *Disc*, respectively, and $n_i + n_j + n_k = Z$. As such, for each strategy state n_{ijk}, there will be an associated Markov chain with a corresponding set of states $\mathcal{G}_{ijk} = \{g_{ijk} \mid 0 \leq g_t \leq n_t, \forall t \in \{i, j, k\}\}$ where $g_{ijk} = (g_i, g_j, g_k)$ is a state where there are g_i, g_j and g_k good individuals using strategy *AllC*, *AllD* and *Disc*, respectively. In the presence of FRSAs, one must also remove any states where the number of individuals of strategy f_p with reputation f_r is less than $\mathcal{A}$. We can represent this new set of reputation states as

$$\mathcal{G}_{ijk}^f = \{g_{ijk} \mid 0 \leq g_t \leq n_t \wedge \begin{cases} g_{f_p} \geq \mathcal{A} & , \text{if } f_r = \mathbf{G} \\ g_{f_p} \leq n_{f_p} - \mathcal{A} & , \text{if } f_r = \mathbf{B} \end{cases},$$
$$\forall t \in \{AllC, AllD, Disc\}\} \tag{3}$$

where we abuse notation by denoting g_{f_p} as the number of good individuals using strategy f_p, with an analog definition for n_{f_p}.

The transition probabilities of having one more (H_p^+) or one less (H_p^-) good individual using strategy p are given by

$$H_p^+(g_{ijk}) = \frac{n_p - g_p - Q_{p,f}^+}{Z} \left(\frac{|g_{ijk}|}{Z - 1} G_G^p + \frac{Z - |g_{ijk}| - 1}{Z - 1} G_B^p \right) \tag{4}$$

and

$$H_p^-(g_{ijk}) = \frac{g_p - Q_{p,f}^-}{Z} \left(\frac{|g_{ijk}| - 1}{Z - 1} \bar{G}_G^p + \frac{Z - |g_{ijk}|}{Z - 1} \bar{G}_B^p \right) \tag{5}$$

where $|g_{ijk}|$ is the total number of good individuals in the reputation state, $\bar{G}^p = (1 - G^p)$, and $Q_{p,f}^+ = \mathcal{A}$ if $f_r = \mathbf{G}$ and 0 otherwise, and $Q_{p,f}^- = \mathcal{A}$ if $f_r = \mathbf{B}$ and 0 otherwise, which represent the damping factors related to the inability of the FRSAs in changing reputation. Intuitively, these can be understood as reducing the number of agents available to change reputation in the population where FRSAs are present. If the FSAs do not have fixed reputations, then this damping factor should not be applied, and therefore $Q_{p,f}^{\pm} = 0$.

For a given strategy state, we can now fully define the transition matrix H for the reputation dynamics. Each entry $H_{a,b}$ represents the probability of transitioning from a state g_{ijk}^a to a state g_{ijk}^b. Thus, each entry in the matrix is defined as

$$H_{a,b} = \begin{cases} H_p^+(g_{ijk}^a) & \text{if } g_p^b = g_p^a + 1 \wedge g_{p'}^b = g_{p'}^a \wedge g_{p''}^b = g_{p''}^a \\ H_p^-(g_{ijk}^a) & \text{if } g_p^b = g_p^a - 1 \wedge g_{p'}^b = g_{p'}^a \wedge g_{p''}^b = g_{p''}^a \\ H^=(g_{ijk}^a) & \text{if } g_{ijk}^b = g_{ijk}^a \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where $p, p', p'' \in \{AllC, AllD, Disc\}$ and $p \neq p' \neq p''$.

The size of the reputation state space will be equal to $S_R = (n_i + 1 - T_f(AllC))(n_j + 1 - T_f(AllD))(n_k + 1 - T_f(Disc))$, where

$$T_f(p) = \begin{cases} 0 & \text{if } f_p \neq p \\ \mathcal{A} & \text{if } f_p = p \end{cases}. \qquad (7)$$

As it is possible to transition from any state to every other state in a finite number of time steps – that is, the Markov chain is irreducible – the stationary distribution σ exists, is unique, and can be found by searching for the eigenvector with the associated eigenvalue 1. Thus, it must verify the propriety $\sigma H = \sigma$. We clarify that, when AAs do not have fixed reputations, this stationary distribution is equal to that of the case when $\mathcal{A} = 0$ in each possible strategy state, independently of $\mathcal{A}$. This is not the case for FRSAs, as the reduction of available states causes the transition probabilities to differ.

3.4 Strategy dynamics

The adoption of strategies follows a birth-death process which models strategy imitation and mutation [28]. Individuals tend to imitate those who perform better than them, which we model using the pairwise comparison rule (also known as the Fermi update rule) [53]. As such, the probability that an individual using strategy p adopts strategy p' is given by $P_{p \to p'}(n_{ijk}) = 1/(1 + e^{-\beta \Delta F_{p,p'}})$, where $\Delta F_{p,p'}(n_{ijk}) = \bar{F}_{p'}(n_{ijk}) - \bar{F}_p(n_{ijk})$ is the difference between the average fitness of strategy p' and strategy p, and β is the strength of selection, where a higher value ($\beta \to \infty$) results in a more deterministic evolutionary process, and a lower value ($\beta \to 0$) approximates the random selection process.

We now determine the average fitness of a strategy p in the context of a donation game, with benefits b and costs c as introduced in Section 3.1. The expected payoff of a strategy p in the strategy state n_{ijk} and reputation state g_{ijk} is given as $F_p(n_{ijk}, g_{ijk}) = bR_p(n_{ijk}, g_{ijk}) - cD_p(n_{ijk}, g_{ijk})$, where $R_p(n_{ijk}, g_{ijk})$ represents the probability that an individual of strategy p will receive a donation, and is given by

$$R_p(n_{ijk}, g_{ijk}) = \frac{g_p}{n_p} \left(\frac{n_p - 1}{Z - 1} C_G^p + \frac{n_{p'}}{Z - 1} C_G^{p'} + \frac{n_{p''}}{Z - 1} C_G^{p''} \right)$$
$$+ \frac{n_p - g_p}{n_p} \left(\frac{n_p - 1}{Z - 1} C_B^p + \frac{n_{p'}}{Z - 1} C_B^{p'} + \frac{n_{p''}}{Z - 1} C_B^{p''} \right), \qquad (8)$$

where p' and p'' are the other available strategies, and $D_p(n_{ijk}, g_{ijk})$ is the probability that an individual using strategy p will donate, and is calculated as

$$D_p(n_{ijk}, g_{ijk}) = \frac{g_p}{n_p} \left(\frac{|g_{ijk}| - 1}{Z - 1} C_G^p + \frac{Z - |g_{ijk}|}{Z - 1} C_B^p \right)$$
$$+ \frac{n_p - g_p}{n_p} \left(\frac{|g_{ijk}|}{Z - 1} C_G^p + \frac{Z - |g_{ijk}| - 1}{Z - 1} C_B^p \right). \qquad (9)$$

Finally, having access to the stationary distribution σ over reputation states, the average fitness is determined by

$$\bar{F}_p(n_{ijk}) = \sum_{g_{ijk} \in \mathcal{G}_{ijk}^f(n_{ijk})} \sigma_{g_{ijk}} F_p(n_{ijk}, g_{ijk}), \qquad (10)$$

where $\sigma_{g_{ijk}}$ is the value of the stationary distribution in the reputation state g_{ijk}.

Similar to the process described in [38, 41], we now define another Markov chain which describes the strategy adoption dynamics. Such a Markov chain will be constructed following the process defined above for the reputation dynamics, only transitioning between strategy states, n_{ijk}, instead of reputation states. Given the presence of FSAs, the state space of the strategy Markov chain is given by

$$\mathcal{M}^f = \{n_{ijk} \mid n_i + n_j + n_k = Z \wedge n_{f_p} \geq \mathcal{A}\}. \qquad (11)$$

As such, it is necessary to first define the transition probabilities between each state. Thus, we define the probability that an agent using strategy p changes to strategy p' under state n_{ijk} to be

$$M_{p \to p'}(n_{ijk}) = O_f^p(n_{ijk}) \left(\bar{\gamma} \frac{n_p}{Z} \frac{n_{p'}}{Z - 1} P_{p \to p'}(n_{ijk}) + \gamma \frac{n_p}{2Z} \right), \qquad (12)$$

where γ is the chance of mutation and $\bar{\gamma} = (1 - \gamma)$, and $O_f^p(n_{ijk})$ represents the factor accounting for the presence of FSAs of strategy p, which are unable to change strategy, and is given by

$$O_f^p(n_{ijk}) = \begin{cases} 0 & \text{if } f_p = p \wedge n_p < \mathcal{A} \\ \frac{n_p - \mathcal{A}}{n_p} & \text{if } f_p = p \wedge n_p \geq \mathcal{A} \\ 1 & \text{otherwise.} \end{cases} \qquad (13)$$

Given that one strategy is always constrained to be greater or equal to $\mathcal{A}$, the size of the strategy state space will be $S_S = \binom{Z - \mathcal{A} + 2}{2}$.

Finally, we define the transition matrix M for the strategy adoption dynamics. Similar to Equation (6), each entry $M_{a,b}$ will correspond to the probability of transitioning from state n_{ijk}^a to state n_{ijk}^b, and is defined as

$$M_{a,b} = \begin{cases} M_{p \to p'}(n_{ijk}^a) & \text{if } n_p^b = n_p^a - 1 \wedge n_{p'}^b = n_{p'}^a + 1 \\ & \wedge n_{p''}^b = n_{p''}^a \\ 1 - \sum M_{p \to p'}(n_{ijk}^a) & \text{if } n_{ijk}^b = n_{ijk}^a \\ 0 & \text{otherwise} \end{cases}, \qquad (14)$$

where $p, p', p'' \in \{AllC, AllD, Disc\}$ and $p \neq p' \neq p''$.

We can now define the gradient of selection, that is, the vector that points towards the most likely evolutionary path, in a given strategy state, as $\vec{v}(n_{ijk}^a) = (M_{AllC}^+ - M_{AllC}^-, M_{AllD}^+ - M_{AllD}^-, M_{Disc}^+ - M_{Disc}^-)$ where $M_p^+ = M_{p' \to p} + M_{p'' \to p}$ and $M_p^- = M_{p \to p'} + M_{p \to p''}$ represent the probability that an agent adopts or abandons strategy p, respectively.

3.5 Cooperation Index

Having the Markov chain that describes the strategy adoption dynamics, we now aim to calculate the cooperation index [39], which measures the fraction of actions that lead to cooperation. For that end, we first define the average cooperation of a strategy p over a strategy state n_{ijk} to be

$$\mathcal{C}_p(n_{ijk}) = \sum_{g_{ijk} \in \mathcal{G}_{ijk}^f(n_{ijk})} \sigma_{g_{ijk}} D_p(n_{ijk}, g_{ijk}). \qquad (15)$$

Denoting θ as the stationary distribution of M, over strategy states, (which is found similarly to σ, see Section 3.3), we define the cooperation index as

$$I = \sum_{n_{ijk} \in \mathcal{M}^f} \theta_{n_{ijk}} \frac{1}{Z} \sum_{p \in \{AllC, AllD, Disc\}} \mathcal{C}_p(n_{ijk}) n_p(n_{ijk}). \qquad (16)$$

3.6 Approximations through Reputation Down-sampling

While the previous sections described how to achieve the exact co-operation index for a given population size Z, the number of strategy and reputation states quickly expands as Z increases. As such, computational resources might not suffice. To this end, we expand on the method described before by employing an approximation of the reputation dynamics, drastically reducing the necessary number of reputation states while preserving the original dynamics.

Our method relies on determining the weighted average reputation state g^*_{ijk} in any given strategy state n_{ijk}. We note that, by increasing the population size, the average fraction of **G** individuals in any strategy remains the same, save for rounding errors, as the transition probabilities depend solely on the fractions of good and bad individuals, and never solely on the population size. Thus, one can safely approximate g^*_{ijk} for a population of size Z by calculating it for a lower population size Z', and scaling it accordingly to fit Z. Under this approximation, the number of reputation states will scale at a constant factor of $O(Z'^3)$. The average reputation state is computed from the stationary distribution as

$$g^*_{ijk} = \sum_{g_{ijk} \in \mathcal{G}^f_{ijk}(n_{ijk})} \sigma_{g_{ijk}} g_{ijk}. \tag{17}$$

From there, the calculations that rely on σ are modified to instead use g^*_{ijk}. These are the average fitness, which becomes $\bar{F}_p(n_{ijk}) = F_p(n_{ijk}, g^*_{ijk})$, and the average cooperation of a strategy p, becoming $\mathcal{C}_p(n_{ijk}) = D_p(n_{ijk}, g^*_{ijk})$. We note that using the average reputation is valid whenever the functions, in this case R_p and D_p, which produce both $\bar{F}_p$ and $\mathcal{C}_p$, are linear in respect to changes in g_{ijk} in any dimension. As this is the case, this approximation is applicable without loss of generality.
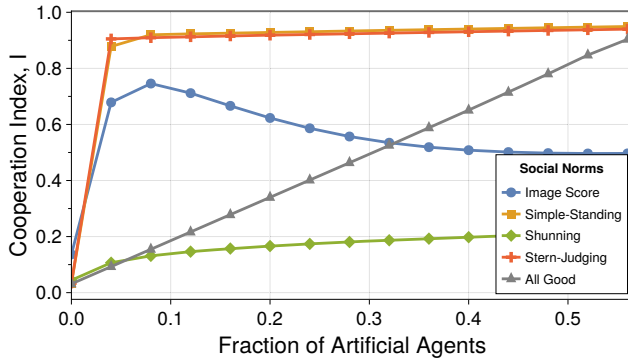
4 Results

Figure 2. Cooperation levels for different fractions $\mathcal{A}/Z$ of fixed-strategy Artificial Agents with a Discriminator strategy (FSA-Disc) present in the population, for each social norm. We observe that even low fractions of FSA-Disc improve cooperation under the **SS** and **SJ** norms; the effect of FSA-Disc is limited with **IS** and **SH**. For **IS**, after an initial cooperation peak, further introducing FSA-Disc slightly reduces cooperation for **IS**. $Z = 100, Z' = 25, b = 2, c = 1, \epsilon = \alpha = \chi = 0.01, \gamma = 0.01$.

In this section, we present the results on the cooperation index, when artificial agents (AAs) are introduced in the population.

Figure 2 presents the cooperation index for the previously mentioned social norms, as a function of the fraction of introduced *Disc*

Fixed-Strategy Agents (FSAs), in a scenario of low cooperation benefits ($b/c = 2$). As well known from previous work (and for completeness illustrated in Figure 1 of the Supplementary Material [1]), under low values of b/c ($b/c < 2.5$) cooperation cannot be achieved. Figure 2 reveals that, however, even under $b/c = 2$ cooperation can be facilitated by introducing agents with a Disc strategy. We observe that introducing a small fraction of FSAs with a Disc strategy increases cooperation levels for **SS**, **SJ** and **IS**.

To better understand the positive effects on cooperation of introducing FSAs, we analyze the most prevalent states of the strategy Markov chain before and after the introduction of *Disc* FSAs. In Figures 3 and 4, with and without FSAs, respectively, we observe the resulting gradient of selection, average reputation and stationary distribution at each strategy state, under **IS**, **SS** and **SJ** – this corresponds to the transition presented in Figure 2. We observe that, while reputations necessarily stay equal in each state throughout both scenarios, the gradient of selection is remarkably different, resulting in entirely distinct cooperation indexes. The cause of such an abrupt transition is further illustrated when studying the gradient of selection among the *AllD-Disc* edge (Figure 5, Supplementary Material [1]), where we observe that a small fraction of *Disc* AAs is enough to make selection promote discriminators, thus resulting in the rapid transition towards cooperation when reputations are good.

In Figure 5, we study the consequence of having a biased reputation towards AAs. Whenever **G** FRSAs are introduced, the prior effects under **IS**, **SS** and **SJ** are again present, as under these norms the reputations of *Disc* agents is often good. Most importantly, **G** FRSAs stand as the only type of agent that can sharply increase cooperation across all the social norms tested, including **SH**, even when comprising a large fraction of the population. As for **B** FRSAs, despite promoting cooperation when present in low fractions, a greater incidence of these agents leads to a drop in cooperation, suggesting a saturation effect, as these AAs will often be uncooperative towards each other and hinder overall reputations.

To understand the effects of discriminator FSAs, in Figure 2 of the Supplementary Material [1], we expand to a more challenging scenario, where $b/c = 1.1$. We observe a similar influence of discriminating FSAs in cooperation. However, for FSAs and **G** FRSAs, a greater fraction (around 0.15% as opposed to 0.05%) is necessary to trigger the transition to almost system-wide cooperation under **SS** and **SJ**. Crucially, **G** FRSAs remain effective at lower fractions of the population across all norms. Furthermore, **B** FRSAs have a lower overall capacity to improve cooperation, with low peaks of cooperation happening at high fractions of FSAs. Additionally, we investigate the scenario where $b/c = 5$ (Figure 3, Supplementary Material [1]), as it presents moderately elevated cooperation without the presence of AAs under **SS** and **SJ**. In this setting, while the influence of dynamic-reputation FSAs and **G** FRSAs follows prior experiments, we observe drops of cooperation with the addition of **B** FRSAs under **SJ** and a rise-and-fall behavior in both **SS** and **IS**, similar to Figure 2 for **IS**. Finally, we note that **AG** shows a predictable pattern across all experiments, as adding discriminator FSAs when every agent is considered good leads to their consistent cooperation: since these agents cannot effectively outperform *AllD* until in high enough numbers, cooperation rises proportionally to the fraction of *Disc* added, showing little capacity to increase the cooperation of adaptive agents.

Given the well-mixed setting we assume here, there are three components at play when considering the impact of AAs on cooperation: 1) the change in the number of role models that can be imitated, 2) the resulting payoffs after interacting with a different distribution of strategies, and 3) the higher incidence of a given reputation as a
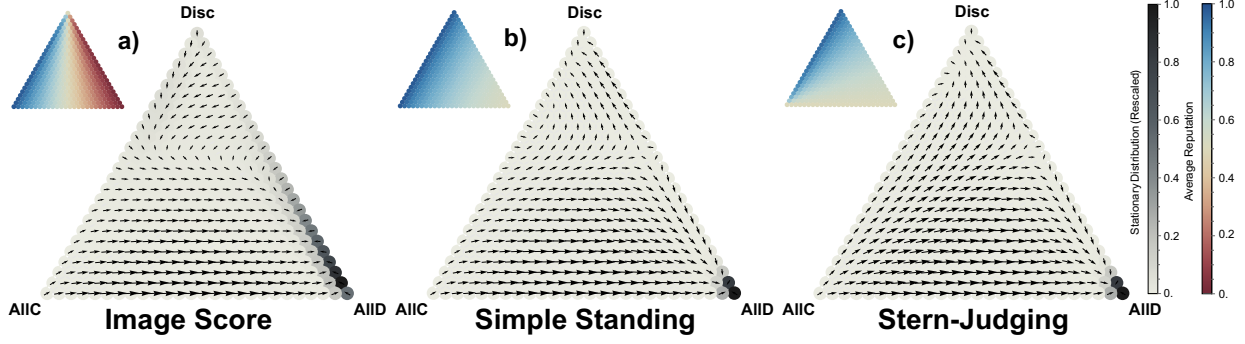
Figure 3. Evolutionary dynamics under **a) IS**, **b) SS**, and **c) SJ** without the presence of FSAs. Each circle corresponds to a possible combination of strategies. Arrows indicate the direction and magnitude (through the size) of the gradient of selection, $\vec{v}$ (Eq. 16). The large simplex indicate the stationary distribution over each strategy-combination state, where darker colors represents the states in which the system spends more time in. The smaller simplex illustrates the average reputation of an individual in each state, where blue indicates reputations are good, and red that reputations are bad. Given the low b considered (see Figure 1 of the Supplementary Material [1]), we observe, across all norms, a high prevalence of strategy *AllD*, leading to low cooperation rates. $Z = 100, Z' = 25, \mathcal{A} = 0, b = 2, c = 1, \epsilon = \alpha = \chi = 0.01, \gamma = 0.01$.
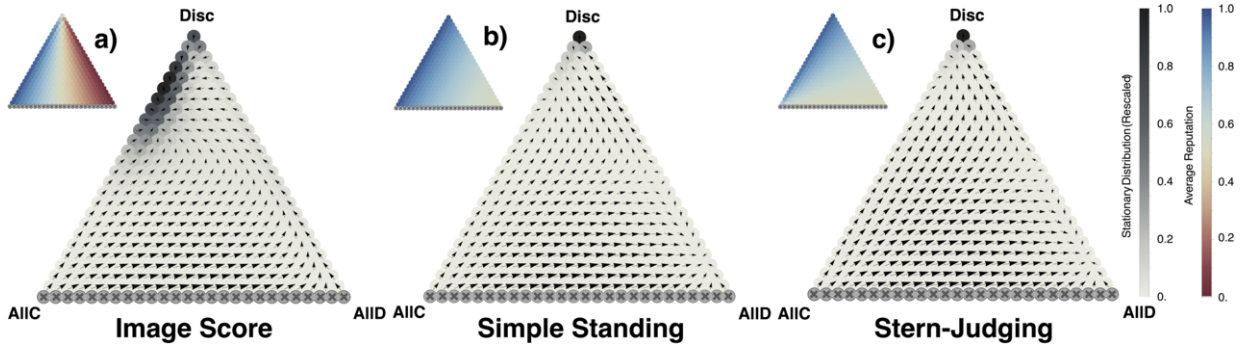

Figure 4. Evolutionary dynamics under **a) IS**, **b) SS**, and **c) SJ** in the presence of a small fraction ($\mathcal{A}/Z = 0.04$) of artificial agents. We use the same notation and parameters as Figure 3, except $\mathcal{A} = 4$. States with gray crosses indicate unreachable states, due to the presence of FSAs. By introducing Disc-FSA agents, the gradient of selection now favors *Disc* and, in **IS**, *AllC* individuals, leading to a considerably higher cooperation across all norms.

consequence of the social norm and, when applicable, the fixed reputation. In the case of discriminator FSAs, these two components will cause either a raise in good or bad discriminators, which then lead to a prevalence of either cooperation or defection, an effect that becomes more prominent as the fraction of FSAs increases. As such, the increased cooperation stemming from introducing *Disc* FSAs is a natural consequence of a higher prevalence of cooperative discriminators, which are sustained in the presence of defectors. This higher fraction of *Disc* agents, considerably greater than the fraction of *Disc* FSAs introduced, is, in turn, what allows social norms such as **SH** to foster cooperation under **G** FRSAs, as these will be necessarily labeled good, allowing other *Discs* to become good too.

To clarify which of these effects is more relevant, additional experiments were conducted (Figures 6 and 7, Supplementary Material [1]) where adaptive agents were not allowed to imitate AAs. This was done by altering Equation (12) to prevent adaptive agents from imitating artificial agents. Our results show that when AAs have no effect on the number of role models available for imitation, their overall effect is weaker. However, the changes in cooperation follow the same patterns as when imitation is possible, suggesting that the change in dynamics introduced by AAs is due to both the changes in payoffs after increasingly interacting with *Disc* agents and the increase in the number of *Disc* role models.

Additionally, we conduct experiments with dynamic-reputation

FSAs using the other available strategies. In the scenario with *AllD* FSAs (Figure 8, Supplementary Material [1]), whatever cooperation exists before the introduction of artificial agents is quickly shattered by their presence. Even in scenarios where the donation benefit is high, under no social norm can cooperation be sustained for small fractions of FSAs. The introduction of *AllC* FSAs (Figure 9, Supplementary Material [1]), although seemingly the most adequate choice to promote cooperative behavior, results in direct exploitation of the introduced agents. In fact, when accounting solely for the cooperation from the adaptive agents – that is, discarding the increased cooperation stemming from adding the FSAs – we observe that cooperation remains largely unchanged in the adaptive population.

We also analyze the robustness of FSAs to various types of errors [14, 39]. These are: execution errors, corresponding to the inability to cooperate correctly; assignment errors, where the resulting reputation is incorrectly assigned; and assessment errors, corresponding to a failure to recognize the reputation of the other agent. Figures 10 to 13 of the Supplementary Material [1] show the resulting cooperation index under all social norms for a varying rate of each type of error, on scenarios with and without dynamic-reputation FSAs and **G** FRSAs. For **SJ**, we observe that, as opposed to the environment without artificial agents, the presence of FSAs is negatively influenced by high error rates. However, a considerably high assignment and assessment error rates are necessary for cooperation to lower, at
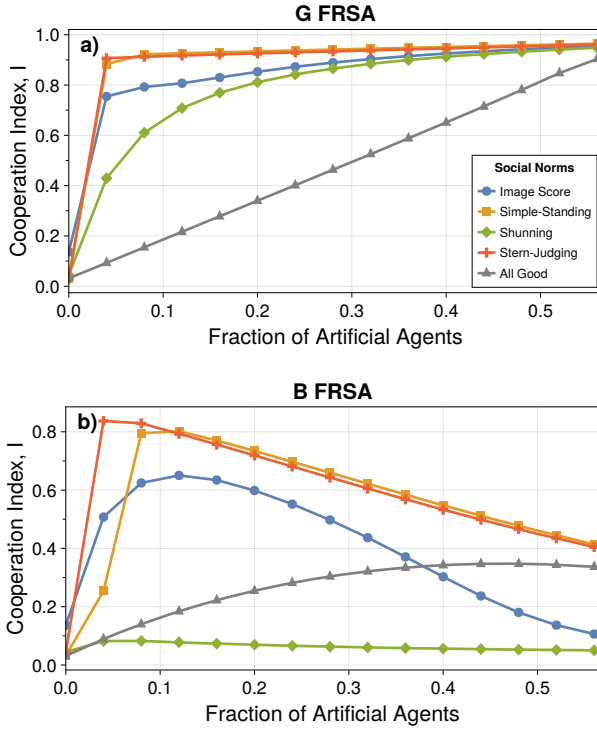
Figure 5. Cooperation levels for different fractions $\mathcal{A}/Z$ of Discriminator Fixed-reputation-and-strategy artificial agents (FRSAs) in the population, for each social norm. **a)** FRSAs are introduced with a **G** reputation. **b)** FRSAs have a **B** reputation. Compared to FSAs, **G** FRSAs are more effective in improving cooperation under **IS** and **SH**. In the case of bad FRSAs, although they initially promote cooperation across **IS**, **SS** and **SJ**, as more agents are introduced (higher $\mathcal{A}/Z$), cooperation lowers. The same parameters as those used in Figure 2 are applied.

which point it drops sharply. As for the remaining social norms: **SS** shares a very similar error profile to **SJ**; under **IS**, cooperation lowers across the three cases as the assignment and assessment errors increase; and, under **SH**, cooperation slightly increases without FSAs and with dynamic-reputation FSAs for certain ranges of errors, while with **G** FRSAs it consistently decreases as the errors increase. While the behavior of dynamic-reputation FSAs and **G** FRSAs aligns in **IS**, **SS** and **SJ** – both increase or decrease under the same error rates –, the presence of dynamic-reputation FSAs aligns instead with the scenario without FSAs under **SH**. We also note that **IS** and **SH** are more sensitive to noise than the remaining social norms considered, having noticeable effects in cooperation starting at error rates of 10^{-2}.

5   Discussion and Conclusion

In this work, we have investigated cooperation in adaptive populations under the presence of artificial agents in the context of indirect reciprocity. The study of **IR** is of particular interest, given the importance of this mechanism in explaining cooperation among unrelated individuals [27] and the possibility that artificial agents impact **IR** dynamics by acting as donors, receivers or observers. It is unclear if **IR** will work effectively when artificial systems permeate society.

To understand this, we developed a model to study the impact of artificial agents, implemented with a fixed strategy, integrated in a well-mixed population of adaptive agents, under four well-known social norms: **IS**, **SS**, **SH** and **SJ**. Our results indicate that the effects of such AAs depends primarily on the strategy they employ and on

the social norm at play, as well as if these agents are seen in a biased manner or if their reputation is defined by the social norm at play. We draw several conclusions: Firstly, the presence of dynamic-reputation *Disc* FSAs allows increased cooperation in previously uncooperative scenarios, under **IS**, **SS** and **SJ**. This result is of particular importance for **IS**, which is a first-order social norm with low cognitive complexity [42]. Furthermore, if *Disc* FRSAs are observed with a positive bias, always being assessed with a good reputation, the previously uncooperative **SH** enables high cooperation levels; cooperation levels under **IS** increase as well. Additionally, we highlight that negative biases towards *Disc* FRSAs result in two opposite forces: an increase of discriminators, which could increase cooperation, and a reduction of good individuals, which typically reduces cooperation. The effect of these agents is thus dependent on the social norm, but also on the benefit of cooperation versus that of defection – we do note that, in general, introducing a low fraction of these agents is still beneficial to cooperation. These findings align with the conclusions of other works on cooperation in hybrid populations outside indirect reciprocity [40, 4, 17, 45], where low fractions of (prosocial) seeding agents appear to considerably boost cooperation. We also remark the different robustness of FSAs in the presence of errors. These findings show that **G** FRSAs, while providing the greatest values of cooperation across all error rates, are also the most influenced by noise when compared to scenarios without FSAs, or with dynamic-reputation FSAs. Additionally, artificial agents are more robust to errors under **SS** and **SJ**, and more sensitive under **IS** and **SH**. We thus note that this type of errors should be of particular concern when considering physical autonomous systems [22, 6], where misinterpretations and operational errors are more common.

The effect of AAs which unconditionally defect is also of great importance, as it highlights a vulnerability of cooperative behavior to uncooperative agents. Our experiments demonstrated that cooperation does not evolve if a low fraction of agents are unconditional defectors. This poses the question of how to develop mechanisms that are resilient against these agents. Furthermore, we also highlight the inefficacy of *AllC* FSAs, which, due to the dominance of *AllD* in the adaptive population, lead to the exploitation of these agents, and prevent increasing the cooperation levels of the adaptive agents. While a purely theoretical model, these results provide a clear framework and baseline for future human-AI experiments, which can help steer AI development towards a focus on promoting pro-social behavior.

Finally, one must be cautious when extracting results from game theoretical models to inform real-world applications and policies [52]. We highlight the need for more thorough human-AI interaction studies [36] in order to bridge the gap between theoretical and experimental results. Despite suggesting that discriminating agents can promote cooperation, it is important to note the ethical concerns involved in having autonomous systems dictate what constitutes an acceptable action [47], as well as the fundamental difference in having systems that opt not to cooperate versus ones that actively defect. In addition, we note that the increased cooperation observed with artificial agents unconditionally deemed *good* should not be read as an argument for an acritical assessment of AI. Our results are constrained to the scope of donation games and indirect reciprocity, discarding eventual risks of over-trusting AI systems.

Acknowledgements

References

[1] *Supplementary Material and Code for "Artificial Agents Facilitate Human Cooperation through Indirect Reciprocity"*, Aug. 2024. Zenodo. Available at https://doi.org/10.5281/zenodo.13311651.

[2] Z. Akata, D. Balliet, M. De Rijke, F. Dignum, and et al. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, Aug. 2020.

[3] R. Alexander. *The biology of moral systems*. Routledge, 2017.

[4] N. Anastassacos, J. García, S. Hailes, and M. Musolesi. Cooperation and Reputation Dynamics with Reinforcement Learning, Feb. 2021. arXiv:2102.07523 [cs].

[5] H. Brandt and K. Sigmund. Indirect reciprocity, image scoring, and moral hazard. *PNAS*, 102(7):2666–2670, 2005.

[6] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th AAMAS*, pages 507–513, 2018.

[7] J. W. Crandall, M. Oudah, Tennom, F. Ishowo-Oloko, S. Abdallah, et al. Cooperating with machines. *Nature Communications*, 9(1):233, 2018.

[8] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. Open Problems in Cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020. Publisher: arXiv Version Number: 1.

[9] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. Cooperative AI: machines must learn to find common ground, 2021.

[10] J. Danaher. Techno-optimism: An analysis, an evaluation and a modest defence. *Philosophy & Technology*, 35(2):54, 2022.

[11] C. M. de Melo, K. Terada, and F. C. Santos. Emotion expressions shape human social norms and reputations. *iScience*, 24(3), 2021.

[12] T. D. Dores Cruz, I. Thielmann, S. Columbus, C. Molho, J. Wu, et al. Gossip and reputation in everyday life. *Philosophical Transactions of the Royal Society B*, 376(1838):20200301, 2021.

[13] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[14] M. A. Fishman. Indirect reciprocity among imperfect individuals. *Journal of Theoretical Biology*, 225(3):285–292, 2003.

[15] H. Gintis. Solving the puzzle of prosociality. *Rationality and Society*, 15(2):155–187, 2003.

[16] N. Griffiths and S. S. Anand. The impact of social placement of non-learning agents on convention emergence. In *AAMAS*, volume 12, pages 1367–1368, 2012.

[17] H. Guo, C. Shen, S. Hu, J. Xing, P. Tao, Y. Shi, and Z. Wang. Facilitating cooperation in human-agent hybrid populations through autonomous agents. *iScience*, 26(11), 2023.

[18] C. A. Hidalgo, D. Orghian, J. A. Canals, F. De Almeida, and N. Martin. *How Humans Judge Machines*. MIT Press, 2021.

[19] C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, and M. A. Nowak. Indirect reciprocity with private, noisy, and incomplete information. *PNAS*, 115(48):12241–12246, 2018.

[20] F. Ishowo-Oloko, J.-F. Bonnefon, Z. Soroye, J. Crandall, I. Rahwan, and T. Rahwan. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, 1(11): 517–521, 2019.

[21] M. Kandori. Social norms and community enforcement. *The Review of Economic Studies*, 59(1):63–80, 1992.

[22] I. Leite, C. Martinho, and A. Paiva. Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, 5:291–308, 2013.

[23] B. Lugrin, C. Pelachaud, and D. Traum, editors. *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, volume 37. ACM, 1 edition, 2021.

[24] A. Mao, L. Dworkin, S. Suri, and D. J. Watts. Resilient cooperators stabilize long-run cooperation in the finitely repeated prisoner's dilemma. *Nature Communications*, 8(1):13800, 2017.

[25] M. A. Nowak. Five rules for the evolution of cooperation. *Science*, 314 (5805):1560–1563, 2006.

[26] M. A. Nowak and K. Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577, 1998.

[27] M. A. Nowak and K. Sigmund. Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298, 2005.

[28] M. A. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428 (6983):646–650, 2004.

[29] H. Ohtsuki and Y. Iwasa. How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of theoretical biology*, 231(1): 107–120, 2004.

[30] H. Ohtsuki and Y. Iwasa. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of theoretical biology*, 239(4):435–444, 2006.

[31] H. Ohtsuki and Y. Iwasa. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *Journal of theoretical biology*, 244(3):518–531, 2007.

[32] I. Okada. A Review of Theoretical Studies on Indirect Reciprocity. *Games*, 11(3):27, July 2020.

[33] R. Oliveira, P. Arriaga, F. P. Santos, S. Mascarenhas, and A. Paiva. Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior*, 114:106547, Jan. 2021.

[34] J. M. Pacheco and F. C. Santos. The messianic effect of pathological altruism. In B. Oakley, A. Knafo, G. Madhavan, and D. S. Wilson, editors, *Pathological Altruism*, page 300. Oxford University Press, 2011.

[35] J. M. Pacheco, F. C. Santos, and F. A. C. Chalub. Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Computational Biology*, 2(12):e178, 2006.

[36] A. Paiva, F. Santos, and F. Santos. Engineering pro-sociality with autonomous agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[37] S. Rosenthal, J. Biswas, and M. M. Veloso. An effective personal mobile robot agent through symbiotic human-robot interaction. In *AAMAS*, volume 10, pages 915–922, 2010.

[38] F. P. Santos, J. M. Pacheco, and F. C. Santos. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Scientific Reports*, 6(1):37517, 2016.

[39] F. P. Santos, F. C. Santos, and J. M. Pacheco. Social norms of cooperation in small-scale societies. *PLoS Computational Biology*, 12: e1004709, 2016.

[40] F. P. Santos, J. M. Pacheco, A. Paiva, and F. C. Santos. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6146–6153, 2019.

[41] F. P. Santos, S. Mascarenhas, F. C. Santos, F. Correia, S. Gomes, and A. Paiva. Picky losers and carefree winners prevail in collective risk dilemmas with partner selection. *Autonomous Agents and Multi-Agent Systems*, 34(2):40, Oct. 2020.

[42] F. P. Santos, J. M. Pacheco, and F. C. Santos. The complexity of human cooperation under indirect reciprocity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1838):20200291, Nov. 2021.

[43] S. Sen and S. Airiau. Emergence of norms through social learning. In *IJCAI*, volume 1507, page 1512, 2007.

[44] S. Serholt, W. Barendregt, A. Vasalou, P. Alves-Oliveira, A. Jones, S. Petisca, and A. Paiva. The case of classroom robots: teachers' deliberations on the ethical tensions. *AI & Society*, 32:613–631, 2017.

[45] G. Sharma, H. Guo, C. Shen, and J. Tanimoto. Small bots, big impact: solving the conundrum of cooperation in optional prisoner's dilemma game through simple strategies. *Journal of The Royal Society Interface*, 20(204):20230301, 2023.

[46] H. Shirado, S. Kasahara, and N. A. Christakis. Emergence and collapse of reciprocity in semiautomatic driving coordination experiments with humans. *PNAS*, 120(51):e2307804120, Dec. 2023.

[47] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

[48] K. Sigmund. *The calculus of selfishness*. Princeton University Press, 2010.

[49] K. Sigmund. Moral assessment in indirect reciprocity. *Journal of Theoretical Biology*, 299:25–30, 2012.

[50] N. Takahashi and R. Mashima. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. *Journal of Theoretical Biology*, 243(3):418–436, 2006.

[51] I. Terrucha, E. Fernández Domingos, F. C. Santos, P. Simoens, and T. Lenaerts. The art of compensation: How hybrid teams solve collective-risk dilemmas. *Plos One*, 19(2):e0297213, 2024.

[52] A. Traulsen and N. E. Glynatsi. The future of theoretical evolutionary game theory. *Philosophical Transactions of the Royal Society B*, 378 (1876):20210508, 2023.

[53] A. Traulsen, M. A. Nowak, and J. M. Pacheco. Stochastic dynamics of invasion and fixation. *Physical Review E*, 74(1):011909, 2006.

[54] M. Veloso, J. Biswas, B. Coltin, and S. Rosenthal. Cobots: robust symbiotic autonomous mobile service robots. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 4423–4429. AAAI Press, 2015.

[55] J. W. Weibull. *Evolutionary game theory*. MIT Press, 1997.