Artificial Agents Mitigate the Punishment Dilemma of Indirect Reciprocity

Alexandre S. Pires
University of Amsterdam
Amsterdam, The Netherlands
a.m.dasilvapires@uva.nl

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

ABSTRACT

Altruistic cooperation is socially desirable yet costly, thereby challenging to promote in multiagent systems. Indirect reciprocity (IR), where the decision to cooperate or defect is based on reputations, serves as a key mechanism to elicit cooperation among selfish agents. However, IR faces challenges under private assessment, due to the so-called punishment dilemma: without mechanisms forcing reputation consensus, disagreements will emerge, resulting in apparently unjustified defections which are punished. Following the increasing prevalence of hybrid systems, where artificial agents (AAs) coexist with humans, we aim to understand the role of AAs in alleviating IR's punishment dilemma and improving cooperation. We develop an analytical evolutionary game-theoretical model to study cooperation under IR with private assessment. A fixed-strategy AA is embedded within an adaptive population, the latter simulating a population of humans adapting over time. We show that limited interactions with the AA are sufficient to impact the distribution of reputations in a population, allowing justified defection to be widely recognized and fostering cooperation. This work highlights the potential of using artificial agents, even with simple fixed strategies, to impact humans' moral assessments, generate reputation consensus and promote cooperation.

KEYWORDS

Cooperation; Indirect Reciprocity; Reputations; Hybrid Populations; Mixed-Motive Games; Evolutionary Game Theory; Agentic AI

ACM Reference Format:
Alexandre S. Pires and Fernando P. Santos. 2025. Artificial Agents Mitigate the Punishment Dilemma of Indirect Reciprocity. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

Cooperation requires that an individual spends a cost, c, to offer a benefit, b, to another individual. A social dilemma exists when $b > c > 0$, as cooperation provides a greater benefit, but defection is the rational choice [66]. Several mechanisms enable human prosocial behavior [35], and indirect reciprocity (**IR**) is fundamental for cooperation between unrelated individuals [37]. In particular, **IR** can promote cooperation through reputations [3] – even if individuals interact with others for the first time, they might have observed or heard about their prior interactions (e.g., via gossip [9]) and use this information to decide whom to help. A good reputation is thus important to keep, as it becomes key to receive cooperation.

IR has been studied under variable observability, reputation spreading mechanisms, and assessment rules [42]. A challenging scenario for maintaining cooperation occurs when reputations are private, that is, reputations are not publicly shared, which can prevent consensus on who deserves cooperation [23, 28]. As disagreements multiply, punishment against defectors might not be understood – this is known as the *punishment dilemma* [33, 72, 73].

While **IR** has been studied in the context of human populations, agentic AI systems [63] and Socially Interactive Agents [30] are now widely accessible and can impact dynamics of human pro-sociality and reciprocity [46]. Scenarios of hybrid populations [2, 8, 12, 48, 71], where humans coexist with artificial agents (AA), have gathered attention for their promise as tools to promote cooperation [18]. However, the effects of AAs on **IR**, particularly so in the context of private reputations, remain unclear. This context is relevant when interacting with cooperative robots [76], or chatbots [6], where interactions can have low observability. AAs also pose challenges under **IR**, as it has been shown that they are judged differently than humans. In particular, AAs are judged by their actions, and humans by their intentions [22]. This difference naturally influences how reputations are assigned to humans and AAs, affecting human-AI interactions, particularly so in prosocial behavior [47].

We pose the following research questions: **1) Can an artificial agent (AA) promote cooperation under indirect reciprocity with private reputations? 2) Can an AA increase agreement in human reputations, mitigating IR's punishment dilemma? 3) Do simplified human judgments against AAs affect their impact in promoting cooperation and agreement under IR?**

To address these questions, we introduce an analytical evolutionary game theoretical model [66] where a finite population of adaptive agents repeatedly play a *donation game* among each other and an AA. The adaptive population is used as a proxy for humans, who can adapt their strategies over time [66] [1]. Agents can cooperate, **C**, paying a cost c to offer the other agent a benefit b; or defect, **D**, giving no benefit at no own cost. In our model, agents attribute reputations depending on the type of agents involved, enabling distinct judgments between AAs and humans. A schematic view of our model is available in Figure 1. We show how a small fraction of interactions with the AA is enough to promote cooperation across most social norms, and observe how the AA is capable of increasing the distinguishability between defectors and cooperators in human-assigned reputations, mitigating the punishment dilemma.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[1] Although we refer to *adaptive agents* and *humans* interchangeably for simplicity, we clarify that no experiments with humans were conducted in this work.

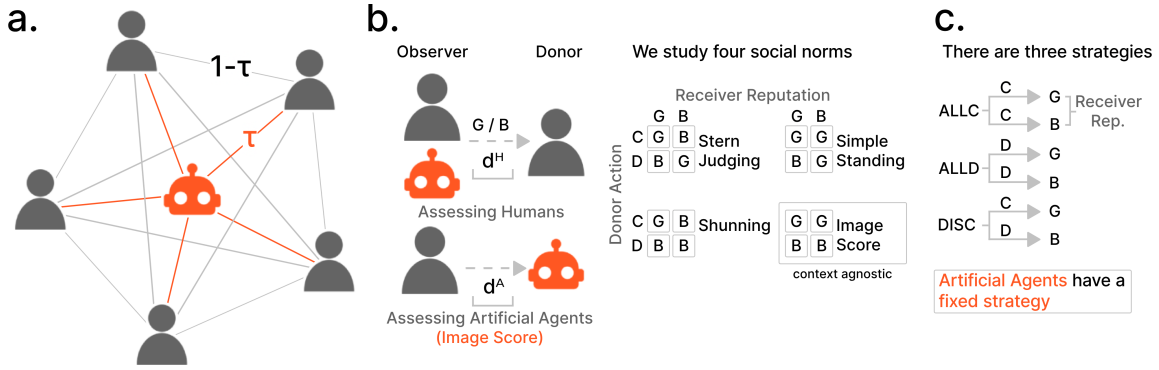

Figure 1: A schematic view of our model: a. A well-mixed population where with a probability τ humans interact with the AA, otherwise $(1-\tau)$ they interact with a randomly sampled human. b. Social norms define how reputations are assigned, based on the donor's action and the receiver's reputation. Norms applied to humans and AAs differ: AAs are judged solely by their action and not the receiver's reputation. c. We study dynamics between 3 strategies: unconditional cooperation (ALLC), unconditional defection (ALLD) and reputation-based discrimination (DISC), that cooperate only against those with good reputation.

2 RELATED WORK

Cooperation and consensus under indirect reciprocity: There has been much research into the role of **IR** in promoting cooperation across human societies [36, 42]. Emphasis has been placed on understanding the role of social norms, i.e., rules that dictate which reputation an individual is assigned following an interaction. In particular, previous works explored their emergence and evolution [37, 49, 78, 79], complexity [57], stability [39, 40], and relationship with culture and morality [16, 38]. The spreading of reputations throughout the population has also received considerable attention: while much work has assumed public reputations as a consequence of gossip, others have considered partial gossip or even fully private assessments [15, 23, 52]. As opposed to public assessment, private reputations pose challenges to cooperation through **IR**, as reputation disagreements lead to apparent *unjustified defections*, causing further disagreements and thus preventing consensus on who should receive cooperation [67, 72–74, 78]. Recent work has focused on reputation agreement [28] and proposed mechanisms, such as gossip and reputation aggregation, to promote it [25, 33].

Human judgment in hybrid populations: Recent work has shed light on how human judgment differs for actions performed by AAs or other humans [7, 19, 20, 31, 44]. Despite humans expressing the same ideal social norm for both types of agents before observation (e.g., "No one should kill."), after observation results differ: While humans are primarily judged by their intentions, the same does not hold true for AAs, as moral justifications are less readily available [32, 59]. Instead, humans employ a simplified judgment, where the action of the AA is the primary factor for the moral judgment of the agent [22]. These different judgment rules suggests that humans employ distinct social norms depending on the type of agent executing an action, and potentially receiving it.

Modelling hybrid populations: The study of artificial agents in hybrid scenarios has been conducted in many contexts. These AAs are characterized as designed agents that employ a selected fixed policy, which can be hard-coded by the agent designer[4]. These are often referred to as "seeding" [4] or "fixed-strategy" agents [17, 61]. Additionally, while the study of independent AAs in cooperation has been well documented [54], current technologies such as chatbots [6] or autonomous vehicles [10] present centralized architectures where either a single AA – or many fully-synchronized agents – exist. In the context of **IR**, these agents can help promote consensus, as there is no disagreement between their instances.

Cooperation in hybrid populations: The study of cooperation in scenarios of hybrid populations [2, 8, 12, 18, 62, 69] has gained large momentum, highlighting the benefits and limitations of AAs in various contexts. Despite this, research focused on **IR** has remained largely absent. [50] has shown that, when reputations are public, cooperation via can be highly promoted by introducing AAs in the population. However, it is unclear if these conclusions translate to private reputations, when AAs are judged differently than humans. In particular, it is important to understand both their influence on human cooperation and in generating consensus and distinguishability on reputations. While many mechanisms have been proposed to assist in synchronizing reputations [25, 27, 60], none has yet considered the role of artificial agents.

Social norms and normative systems: In our model, reputations are assigned based on assessment rules, also called social norms [37]. There is a vast literature on norms in multiagent systems [1, 11, 34, 58, 61]. According to a recent review [21], our norms can be considered *prescriptive* and *explicit*. Moreover, our norms are *essential* norms, used to to solve cooperation dilemmas, as opposed to *conventional*, which result from agents' coordination [77]. Furthermore, we apply norms in a top-down fashion, but their effectiveness is computed via a bottom-up process, where strategies evolve over time; as a result, our norms involve both a top-down "legalist" and bottom-up "interactionist" approach [21].

3 MODEL AND METHODS

We consider a finite and well-mixed population consisting of Z adaptive individuals, following prior work on **IR** [38, 39, 56]. These agents engage in repeated donation games, where an agent, designated as the donor, can either cooperate, **C**, paying a cost c to offer the other agent, the recipient, a benefit b, where $b > c > 0$,

or defect, **D**, where no donation is made, and thus no cost is paid. Other agents observe these interactions and hold a private view of every other agent. That is, any agent i can consider another agent j either **Good (G)** or **Bad (B)**. As these reputations are private, two individuals must not necessarily agree on the reputation of a focal agent. We detail how these reputations are assigned in Section 3.2. This is opposed to public reputation settings, where each agent has a publicly agreed upon reputation, as a consequence of gossip.

The action of each agent depends on their strategy, which itself uses the private reputation of the recipient. Formally, a strategy is a tuple $s = (s_G, s_B)$, where s_G and s_B are the probability of cooperating with an individual considered **G** and **B**, respectively. At any time, an agent will make use of one of the three following strategies: *ALLC* $(1, 1)$, where cooperation is always selected independently of the reputation of the recipient; *ALLD* $(0, 0)$, where the donor always defects; and *DISC* $(1, 0)$, where an individual will only donate to good individuals, and defect against bad individuals. Additionally, we include execution errors: with a probability e_e, an otherwise cooperative act will instead result in a defection [14].

3.1 Introducing an Artificial Agent

We consider a hybrid population, where humans coexist with artificial agents (AAs) [13]. In it, any of the three previous roles of the donation game (Donor, Recipient, and Observer) can be played by an AA. In our model, AAs also hold a private view of other agents, and act according to one of the three possible strategies. As opposed to adaptive agents, the strategy of an AA is hard-coded and thus constant in time [18, 54, 62]. Humans and AAs can be judged differently [22], which we implement through distinct social norms for assessing humans and the AA, as detailed in Section 3.2.

We assume that AAs are perfectly coordinated in their assessments, which can result from perfect communication between independent AAs, the existence of a common reputation database, or a single centralized AA [6]. This avoids disagreements in the private reputation views between AAs, allowing us to instead focus on the effect of coordinated AAs on potentially disagreeing humans. We define τ to be the probability that, for any interaction, a human will instead play with the AA. The fixed strategy of this AA is designated as $s_A \in S = \{ALLC, ALLD, DISC\}$.

3.2 Reputation Dynamics

Agents' private reputations are updated following a social norm. We use second-order social norms [33, 57], which consider the action of the donor and the reputation of the receiver, to assign a new reputation to the donor. These are encoded using a 4-bit tuple $d = (d_{G,C}, d_{G,D}, d_{B,C}, d_{B,D})$, representing the probability of assigning a good reputation in any of the four possible scenarios (e.g., $d_{B,C}$ represents the probability of assigning **G** to a donor using action **C** facing an individual seen as **B**). This allows for a total of 16 second-order social norms, of which we focus on four key norms known to sustain cooperation [39, 40, 68]: **Image Score (IS)** [36], $d = (1, 0, 1, 0)$, where cooperating is always good and defecting is always bad; **Simple Standing (SS)**, $d = (1, 0, 1, 1)$, where only defecting against a good individual is bad; **Shunning (SH)**, $d = (1, 0, 0, 0)$, where only cooperating with a good agent is good; and **Stern Judging (SJ)** [45], $d = (1, 0, 0, 1)$, where both cooperating with good agents and defecting against bad agents is good, and the remaining is bad. We allow for assessment errors, where with a probability e_a the reputation of an agent is incorrectly recalled. If $e_a = 0$ and every agent starts with the same initial assessment, reputations remain indefinitely synchronized, however, if $e_a > 0$, disagreements can appear and propagate. In case of error, we assume all AAs incorrectly recall the reputation of the agent.

Considering that the social norms we study differ only in $d_{B,C}$ and $d_{B,D}$, we can generalize a social norm as the probability of being effectively assigned a good reputation [26] in any of the scenarios above. After including errors, these probabilities are given by:

$$\begin{aligned} P^d_{G,C} &= (1 - e_e)(1 - e_a) + e_e e_a \\ P^d_{G,D} &= e_a \\ P^d_{B,C} &= d_{B,C}(P^d_{G,C} - e_a) + d_{B,D}(1 - P^d_{G,C} - e_a) + e_a \\ P^d_{B,D} &= d_{B,D}(1 - 2e_a) + e_a \end{aligned} \tag{1}$$

As mentioned, human and AAs are not necessarily assigned reputations following the same social norm [22, 29]. Two potentially distinct social norms are applied: d^H, for humans, and d^A, for AAs.

The reputation dynamics are linked to the strategy distribution in the population, and the success of each strategy depends on the distribution of reputations. We consider that the two dynamics happen at distinct timescales [23, 56]. More precisely, reputations are assumed to change at a much faster rate than strategies, and thus for any distribution of strategies it is possible to study the convergence of reputation dynamics. We define a strategy state as $n = (n_{ALLC}, n_{ALLD}, n_{DISC})$, where n_s represents the number of adaptive agents using strategy s, and $n_{ALLC} + n_{ALLD} + n_{DISC} = Z$.

In a well-mixed population, for any strategy state n, reputations are characterized by the probability that any agent, facing an individual of strategy s, will consider that individual good. Although reputations are not assigned based on strategies (as they are not in the norm), the resulting reputations depend on the actions used, and therefore each strategy will have a distinct probability of being considered good. Given the possible combinations of strategies and agent types, we define seven distinct reputation probabilities: $r^H_s, s \in S$, the probability that a human will consider another human using strategy s as good; r^A_s, the probability that the AA will consider a human using strategy s as good; and r^H_A, the probability that a human will perceive the AA as good. After enough pairwise interactions, where agents played both as donors and receivers, we can approximate these reputation probabilities by solving the following ordinary differential equations [49]:

$$\begin{cases} \dfrac{dr^X_s}{dt} &= g^X_s(t) - r^X_s(t), \quad s \in S, X \in \{H, A\} \\ \dfrac{dr^H_A}{dt} &= g^H_A(t) - r^H_A \end{cases}, \tag{2}$$

where $g^X_s(t)$ is the probability that an individual of type X will assign a good reputation to an individual using strategy s, at time step t (following the same scenarios of r^X_s). Adapting [25] to include AAs, each $g^X_s(t)$ will have a term related to human interactions and another for interactions with the AA. Each of these will consider the two scenarios where there is an interaction with an agent considered good and bad, and employ the relevant social norm. For the

probability of assigning a good reputation to the AA, $g_A^H(t)$, only human interactions are relevant, as there are no interactions between AAs. This probability will depend on the strategy of the AA. In the case of *DISC*, it is necessary to compute the probability that the observer and the donor agree in the reputation of the recipient. Considering these scenarios, these probabilities are given by:

$$
\begin{aligned}
g^H_{ALL\lambda}(t) &= \bar{\tau}\left[r^H(t)P^{d^H}_{G,\lambda} + \bar{r}^H(t)P^{d^H}_{B,\lambda}\right] + \\
&\quad \tau\left[r^H_A(t)P^{d^H}_{G,\lambda} + \bar{r}^H_A(t)P^{d^H}_{B,\lambda}\right] \\
g^H_{DISC}(t) &= \bar{\tau}\left[q^g_{H,H}P^{d^H}_{G,C} + q^d_{H,H}\tilde{P}^{d^H} + q^b_{H,H}P^{d^H}_{B,D}\right] + \\
&\quad \tau\left[q^g_{H,A}P^{d^H}_{G,C} + q^d_{H,A}\tilde{P}^{d^H} + q^b_{H,A}P^{d^H}_{B,D}\right] \\
g^A_{ALL\lambda}(t) &= \bar{\tau}\left[r^A(t)P^{d^H}_{G,\lambda} + \bar{r}^A(t)P^{d^H}_{B,\lambda}\right] + \tau P^{d^H}_{G,\lambda} \\
g^A_{DISC}(t) &= \bar{\tau}\left[q^g_{A,H}P^{d^H}_{G,C} + q^d_{A,\bar{H}}P^{d^H}_{B,C} + q^d_{A,\bar{H}}P^{d^H}_{G,D} + q^b_{A,H}P^{d^H}_{B,D}\right] + \\
&\quad \tau\left[g^H_A(t)P^{d^H}_{G,C} + \bar{g}^H_A(t)P^{d^H}_{G,D}\right] \\
g^H_A(t) &= \begin{cases} r^H(t)P^{d^A}_{G,\lambda} + \bar{r}^H(t)P^{d^A}_{B,\lambda} & \text{, if } s_A = ALL\lambda \\ q^g_{A,H}P^{d^A}_{G,C} + q^d_{\bar{A},H}P^{d^A}_{G,D} + & \\ \quad q^d_{A,\bar{H}}P^{d^A}_{B,C} + q^b_{H,A}P^{d^A}_{B,D} & \text{, if } s_A = DISC \end{cases}
\end{aligned}
\tag{3}
$$

where $\lambda \in \{C, D\}$; $\tilde{P}^d = P^d_{G,D} + P^d_{B,C}$; $\bar{\tau} = (1-\tau)$, the probability of a H-H interaction; $r^X(t) = \sum_{s\in S}(n_s/Z)\cdot r^X_s(t)$ is the average reputation of humans as perceived by agents of type $X \in \{H, A\}$; $\bar{g}(t) = 1 - g(t)$; and $\bar{r}(t) = 1 - r(t)$, which is the fraction of bad individuals in respect to $r(t)$. Furthermore, q^g_{XY} (q^b_{XY}) represents the average fraction of humans mutually considered good (bad) in the eyes of an individual of type X and another of type Y. Finally, q^d_{XY} represents the average fraction of humans over which there is a disagreement about the reputations in the perspective of type X and Y individuals. These include both scenarios where X considers one focal agent good and Y considers it bad, and vice versa. When instead we have $q^d_{\bar{X}Y}$, we have that only X considers a focal individual bad, and Y considers it good, and thus $q^d_{\bar{X}Y} + q^d_{X\bar{Y}} = q^d_{XY}$. The agreement and disagreement of private views over a focal individual can be calculated by [26, 52]:

$$
q^g_{XY} = \sum_{s\in S}\frac{n_s}{Z}r^X_s r^Y_s \quad q^b_{XY} = \sum_{s\in S}\frac{n_s}{Z}\bar{r}^X_s\bar{r}^Y_s \quad q^d_{XY} = \sum_{s\in S}\frac{n_s}{Z}\bar{r}^X_s r^Y_s \quad .
\tag{4}
$$

A thorough explanation for Equations 1 and 3 is presented in the supplementary material [51].

3.3 Strategy Adoption Dynamics

We model the adoption of strategies via a birth-death process, where two mechanisms exist: mutations (a probability γ of adopting another available strategy) and social learning. The latter is modelled using the *pairwise comparison rule* [70], otherwise known as the *Fermi update rule*, where an individual will imitate the strategy of another with a probability that increases with the difference in fitness of the two strategies. The probability that an individual using strategy s imitates another using strategy s' is given by $P_{s\to s'}(n) = (1 + e^{-\beta\Delta F_{s,s'}})^{-1}$, where $\Delta F_{s,s'}(n) = \bar{F}_{s'}(n) - \bar{F}_s(n)$ is the difference between the average fitness of strategy s' and strategy s, and β is the strength of selection. A higher strength of selection ($\beta \to \infty$) leads to a deterministic evolutionary process, while a lower value ($\beta \to 0$) converges to a random selection process.

In the donation game, the average fitness of a strategy is determined by two components: b, the benefit a recipient obtains when it is cooperated with; and c, the cost incurred by a donor when it cooperates. The fitness of a strategy will naturally depend on the distribution of strategies and reputations of the population. As such, we first determine the average fitness of a strategy under a strategy state n, given by $F_s(n) = bR_s(n) - cD_s(n)$, where $R_s(n)$ is the probability that an individual of strategy s is cooperated with, which for human individuals is given by

$$
R_s(n) = (1-e_e)\left[\bar{\tau}\left(\frac{n_{ALLC}}{Z} + \frac{n_{DISC}}{Z}r^H_s\right) + \tau C(s)\right] \quad , \tag{5}
$$

where $C(s)$ is the probability that the AA will opt to cooperate (before errors) with an individual of strategy s. As such, $C(s) = 1$ if $s_A = ALLC$, $C(s) = 0$ if $s_A = ALLD$, and $C(s) = r^A_s$ if $s_A = DISC$. For the AA, the probability of receiving a donation is given by

$$
R_A(n) = (1-e_e)\left(\frac{n_{ALLC}}{Z} + \frac{n_{DISC}}{Z}r^H_A\right) \quad . \tag{6}
$$

Likewise, $D_s(n)$ is the probability that an individual using strategy s will donate, and is calculated for humans as

$$
D_s(n) = \begin{cases} 1 - e_e & \text{, if } s = ALLC \\ 0 & \text{, if } s = ALLD \\ \bar{\tau}D^H_{DISC}(n) + \tau D^A_{DISC}(n) & \text{, if } s = DISC \end{cases} , \tag{7}
$$

where $D^H_{DISC}(n) = (1-e_e)r^H$ and $D^A_{DISC}(n) = (1-e_e)r^H_A$ are the probabilities of a *DISC* human cooperating with a human and an AA, respectively. The probability that an AA will donate, $D^A(n)$, is equal to humans, except when $s_A = DISC$, where $D^A(n) = (1-e_e)r^A$.

Using the reputations and fitness at each strategy state, we can analyze the evolution of strategy adoption by employing a Markov chain [53, 55] with the state space given by all the possible strategy states $\mathcal{M} = \{n \mid n_i + n_j + n_k = Z\}$, for a total of $S = \binom{Z+2}{2}$ states. The transition probability between two states that differ only by the strategy of one agent is equal to the probability that an individual using strategy s changes to strategy s' when in state n, via either mutation or imitation, and is given by

$$
M_{s\to s'}(n) = (1-\gamma)\frac{n_s}{Z}\frac{n_{s'}}{Z-1}P_{s\to s'}(n) + \gamma\frac{n_s}{2Z}, \tag{8}
$$

where γ is the aforementioned mutation probability. This formulation, which we follow in the Results section, does not consider that a human can imitate an AA, removing its direct influence on imitations and focusing instead on its impact on reputations and payoffs. However, it is also possible to permit this by replacing $\frac{n_{s'}}{Z-1}P_{s\to s'}(n)$ with $\left(\bar{\tau}\frac{n_{s'}}{Z-1}P_{s\to s'}(n) + T(s')\right)$, where $T(s') = 0$ if $s \neq s_A$ or $T(s) = \tau P_{s\to s_A}(n)$ if $s' = s_A$, which represents the probability of imitating the AA if it is using strategy s'. The transition matrix M of the Markov chain, where each entry $M_{a,b}$ is equal to

the probability of transitioning from state n^a to state n^b, is given by

$$M_{a,b} = \begin{cases} M_{s\to s'}(n^a) & \text{if } n^b_s = n^a_s - 1 \wedge n^b_{s'} = n^a_{s'} + 1 \\ & \wedge\, n^b_{s''} = n^a_{s''} \\ 1 - \sum M_{s\to s'}(n^a) & \text{if } n^b = n^a \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where $s, s', s'' \in S$ and $s \neq s' \neq s''$. Finally, as M is irreducible, its stationary distribution σ is unique and equal to the eigenvector associated with eigenvalue 1 [75], satisfying $\sigma M = \sigma$. We denote as σ_n the value of the stationary distribution at state n.

Additionally, we can determine the gradient of selection, the vector that points towards the most probable evolutionary trajectory in a given strategy state, through $\vec{v}(n) = (M^+_{ALLC} - M^-_{ALLC}, M^+_{ALLD} - M^-_{ALLD}, M^+_{DISC} - M^-_{DISC})$, where $M^+_s = M_{s'\to s} + M_{s''\to s}$ and $M^-_s = M_{s\to s'} + M_{s\to s''}$ are the probabilities that an individual adopts or replaces strategy s, respectively.

3.4 Cooperation and Disagreement Indexes

We measure cooperation through a cooperation index [56], which estimates the fraction of donations in the population. However, as our population contains not just distinct strategies, but distinct types of individuals, it is relevant to distinguish between the different directions of donations. To that end, we define three types of cooperation index: $I^{H,H}$, which accounts for human-human cooperation; $I^{H,A}$, which considers cooperation from humans towards the AA; and I^A, which conversely measures the cooperation of the AA towards humans. These can be calculated as follows:

$$I^{H,X} = \sum_{n\in\mathcal{M}} \sigma_n \frac{1}{Z}\left(D_{ALLC}(n)\cdot n_{ALLC} + D^X_{DISC}(n)\cdot n_{DISC}\right)$$
$$I^A = \sum_{n\in\mathcal{M}} \sigma_n D^A(n) \quad (10)$$

where $X \in \{H, A\}$. It is also possible to quantify the average disagreement among reputation assignments in the human population:

$$q^d = \sum_{n\in\mathcal{M}} \sigma_n q^d_{H,H}(n), \quad (11)$$

where $q^d_{H,H}(n)$ is the disagreement between human reputations at the strategy state n, as expressed in Equation (4).

4 RESULTS

We study the impact of an artificial agent (AA) by measuring the prevalence of cooperation (via the cooperation index, see Methods Section 3.4) as a function of the fraction of interactions between humans and the AA (τ). A focus is given to human-human cooperation, $I^{H,H}$, as it is our primary concern: AAs cooperation is a byproduct of the fixed-strategy considered, and human to AA cooperation is here not assumed to increase the social benefits of cooperation, as we are mainly interested in adaptive agents' welfare. Furthermore, we also explore how disagreements between humans reputation assignments, q^d, change depending on the social norms employed and the presence of the AA. As previously mentioned, in an exclusively human population under private reputations, the perception of unjustified defections (stemming from low agreement on reputations) leads to low cooperation [23, 41, 43, 72, 73]. This is visible in Figure 2, where cooperation is close to zero across all norms tested, even when the donation benefit, b, is considerably larger than c, the cost of donating. An exception is the **IS** norm, which by nature does not consider past reputations, and thus, when errors are not present, works similarly to when reputations are public [72, 74]. Although it achieves the highest cooperation of all norms tested, it still requires a high b/c for cooperation to be as common as defection. Our objective is to study if the presence of an AA can assist in promoting cooperation by increasing coordination in reputations, mitigating the punishment dilemma present when reputations are private.

Firstly, in Section 4.1, we explore scenarios where humans judge AAs based solely on their actions, [20, 22] by fixating d^A = **IS**. In Section 4.2, we then study scenarios where humans and AAs are judged equally, clarifying whether the simplified judgment towards AAs is detrimental to their capacity to promote cooperation.

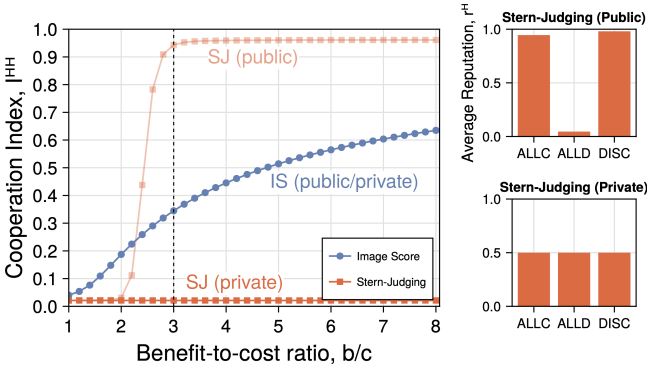

Figure 2: Left: Human cooperation under public and private reputations, $I^{H,H}$, without artificial agents, for two exemplifying social norms and benefit-to-cost ratios, b/c. Right: The average reputation assigned to each strategy under public (top) and private (bottom) reputations ($b/c = 3$). While cooperation increases with b/c under public reputation, when reputations are private cooperation remains close to 0 for every norm except IS, which still leads to low (< 0.65) cooperation. This is attributed to lack of distinguishability between strategies, leading to *DISCs* being unable to punish defectors. $Z = 100, e_e = e_a = 0.01, \gamma = 0.01, \beta = 1$.

4.1 Cooperation under Simplified Human Judgment

In Figure 3, we present the $I^{H,H}$ as a function of b/c, for different fractions of interactions with the AA employing a *DISC* strategy, when fixating d^A = **IS** and varying d^H. When d^H = **IS**, we observe a rise in cooperation from its already modest cooperation rate as τ increases. Similarly, cooperation under **SS** is greatly boosted by the AA, with a low value of τ being enough to reach very high cooperation rates. At a higher τ, **SH** and **SJ** feature phase transitions in their cooperation rates, reaching almost universal cooperation.

These results provide an initial look at how an AA can potentially promote cooperation even under previously uncooperative norms.

As for the other strategies that AAs can employ: using *ALLC*, cooperation under **IS** and **SS** is actually reduced as τ increases. The other norms continue with no cooperation in most cases. Similarly, an *ALLD* AA deters cooperation in the majority of cases. We present and analyze these results in the supplementary material [51].

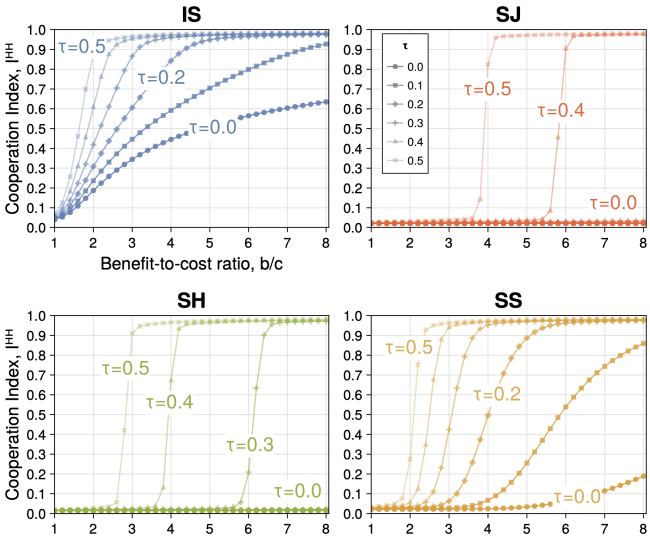

Figure 3: Human cooperation, $I^{H,H}$, at different frequencies of interaction with the *DISC* AA and benefit-to-cost ratios, b/c, when d^A = IS, and varying d^H. A low τ is enough to promote cooperation under IS and SS; But a greater frequency is necessary to promote cooperation under SJ and SH, at which point it features a phase transition to system-wide cooperation. We follow the parameters of Figure 2.

By observing the strategy stationary distribution (see Section 3.3), shown in Figure 4 for **SS** and $b/c = 3$, we can clarify the effect of the *DISC* AA. While at $\tau = 0$ the entire population adopts *ALLD* across all norms, when the transitions to cooperation occur via a higher τ, the composition shifts to a majority of *ALLC* and a fraction of *DISC*. This results from the AA compensating the payoff of *ALLC* and *DISC* over that of *ALLD*, causing a transition. Furthermore, the proportion of *DISC* adopters is related to the reputation of humans: In **IS** and **SS**, whose *ALLC-DISC* simplex edge always contains good individuals (except in the presence of errors, discussed in the supplementary material [51]), *DISC* is more common in the presence of the AA, as its behavior will be undistinguishable from *ALLC*. Inversely, when reputations are overall lower, such as under **SJ** and **SH**, we observe a greater proportion of *ALLC* (see supplementary material [51] for all remaining simplexes).

Although cooperation can drastically increase by introducing an AA, disagreement in reputations does not follow the same pattern across social norms. Figure 5 presents disagreement as a function of b/c, similarly to Figure 3. As expected, at $\tau = 0$, disagreement is very low in **IS**, due to not using reputations, and **SH**, as reputations are low overall. Disagreement is maximized in **SJ**, due to its symmetric nature, and in **SS**, for its symmetry in the *ALLD* equilibrium. However, despite cooperation in **SJ** and **SH** behaving similarly after the addition of the AA, disagreement decreases in **SJ** while increasing in **SH**. In **SS**, cooperation also increases as disagreement decreases. This suggests a norm dependent relation between cooperation and disagreement: Increased disagreement in **SH** is expected to result in better cooperation, as reputations are often low, so more agreement means some agents will instead have good reputations; In **SJ**, due to its symmetry, disagreement stays constant if everyone applies the same norm, but interactions with the AA while the population is positioned at *ALLC* result in higher agreement. Finally, in **SS**, lower disagreements stem from synchronized negative judgments towards defectors, resulting in better performance of *DISC*s.

By studying the evolution of average reputations as τ increases, presented in Figure 6, the impact of the AA in human cooperation becomes clearer. Across all norms, as τ increases, both humans and the AA assign better reputations to *ALLC* and/or *DISC* relative to *ALLD*. In particular: **IS** shows only an increase in *DISC* reputations, as *ALLC* and *ALLD* are universally agreed to be good and bad, respectively; starting from fully neutral reputations at $\tau = 0$, **SJ** shows an increase in the human-assigned reputation of *ALLD* and *DISC* and a decrease of *ALLC* before the phase transition, which then switch to an increase of *ALLC* and *DISC* after the phase transition. However, the AA-assigned reputations instead consistently increase that of *ALLC* while reducing *DISC* and *ALLD* before the phase transition. **SH** increases primarily the reputation of *ALLC* and *DISC* after the phase transition for humans, but across the full τ range for the AA; and **SS** reduces that of *ALLD*, with a slight increase to *DISC*. The difference between human and AA reputations results not just from the simplified human judgment towards AAs, but also from the AA assigning reputations following interactions as opposed to observations, resulting in an easier distinction between *ALLD* and the remaining strategies. This allows the AA to better target punishments towards *ALLD* and benefit cooperative strategies. Furthermore, it justifies the phase transitions observed in **SH** and **SJ**, as such occurs when the payoff of *ALLC* surpasses that of *ALLD*. As human reputations become distinct for *ALLD* and cooperative strategies, the punishment dilemma is mitigated.

4.2 Cooperation under Homogeneous Norms

We now study hybrid populations when humans and AAs are judged equally, that is, $d^H = d^A$, providing a comparison point to understand the impact of simplified human judgment against AAs. The equivalent plots for this scenario are presented in the supplementary material [51]. The results under this setup present key differences to the prior section. In particular, **SS** shows greater overall cooperation, with a higher reputation assigned to the AA. **SJ** requires a lower b/c to feature phase transitions, yet shows neutral human-assigned reputations, indicating the inability of the AA in mitigating the punishment dilemma. Inversely **SH** sees the transitions happen at a higher b/c, as *ALLC* and *DISC* reputations are much lower. As **IS** presents equal results, it is omitted from discussion. Regarding disagreements, as expected, **SJ** shows maximal disagreement independent of τ, as the AA no longer shapes reputations. On the other hand, **SH** shows no decay in disagreement after the phase transition. This suggests that the simplified human

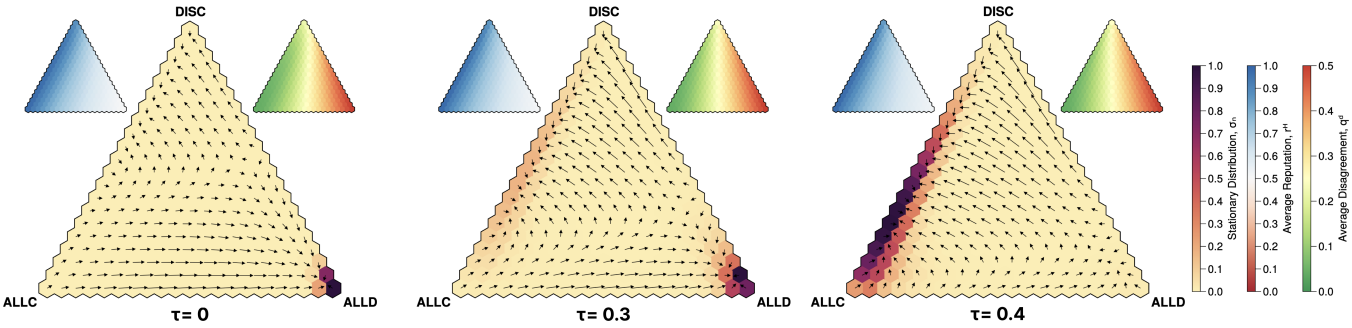

Figure 4: The stationary distribution (rescaled such that the maximum value is 1), gradient of selection, average reputation and average disagreement at each strategy state for different frequencies τ of interaction with a *DISC* AA, when d^A = IS and d^H = SS. The population transitions from an *ALLD* state, with high reputation disagreement, to states where *ALLC* and *DISC* co-exit and every agent accrues the same (good) reputation, resulting in higher cooperation. We follow the same parameters as Figure 2.

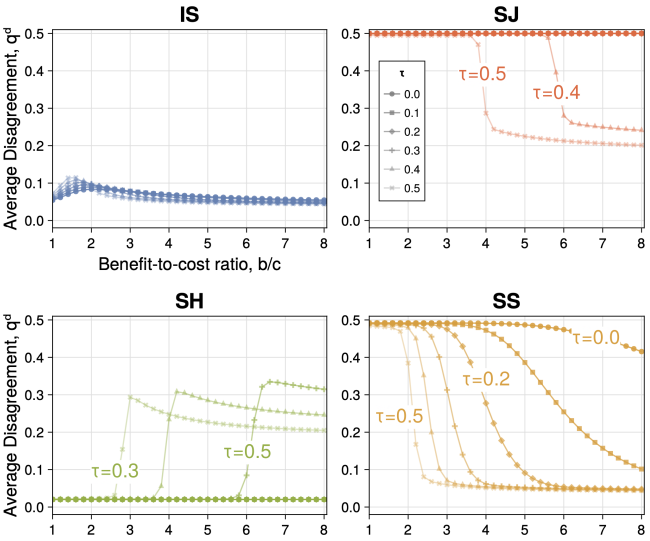

Figure 5: Human disagreement, q^d, at different frequencies of interaction with the *DISC* AA, for different benefit-to-cost ratios, b/c, when d^A = IS, and varying d^H. We see that a low prevalence of the AA decreases disagreement under SS and SJ, but increases it under SH. IS remains largely unaffected by the AA. We follow the parameters of Figure 2.

judgment against AAs can actually have both advantages and disadvantages regarding cooperation and agreement. The reason for this difference is that judging the AA using a distinct norm then influences how humans judge those who interact with it, influencing how the AA then acts, which then cycles back to judging the AA differently. As such, a stricter judgment of the AA results in lower reputations for cooperative strategies, and thus lower cooperation. Since judging the AA with **IS** is less strict than **SH**, we see an increase in cooperation under the simplified human judgment. On the other hand, **SS** is less strict than **IS**, so the simplified judgment results in lower cooperation. We illustrate this in the supplementary material [51] by interpolating between norms for d^A, showing how cooperation increases with less strict judgment of the AA.

5 DISCUSSION AND CONCLUSION

Of the many cooperation mechanisms present in society, indirect reciprocity (**IR**), and reputations, are key to ensure stable cooperation among unrelated individuals [37]. At the same time, AI tools such as chatbots, recommendation, and reputation systems have an increased role in shaping social dynamics [30]. As artificial agents (AAs) permeate physical society through robots, they can fundamentally change the behavior of humans. Although there have been efforts to assess the short-term impact of AAs in human cooperation [47, 64], their potential through **IR** has remained unexplored. Laboratory experiments can help bridge this gap, yet scaling empirical works faces important logistical and technical challenges, and testing future scenarios through longitudinal studies is especially arduous. To this end, we presented a theoretical model to study the impact of an AA, akin to those found in online platforms [6], in a human population interacting through **IR**. In it, agents use context-dependent social norms, contingent on the nature (human or AA) of the observed agents, allowing us to consider that humans judge AAs solely on their outcome, and not their intention [22].

We show that cooperation can be improved, in the donation game and under private reputations, by introducing an AA employing a fixed strategy. While prior work showed that this is the case outside of **IR** [12, 54, 69], we also show how cooperation can still emerge under the stricter assumption that humans will disregard its intentions and not imitate the strategy of the AA. The latter, typically not followed in work on hybrid populations, allows us to better generalize our results with a focus on the impact of AAs in reputations and payoffs.

In our initial experiment, where the AA is judged solely for its action [22], we see a clear benefit by introducing the AA across all norms, albeit with different degrees of success. Norms such as Image-Score (**IS**) and Simple Standing (**SS**) have a greater increase in cooperation than norms such as Stern Judging (**SJ**) and Shunning (**SH**), although all norms can achieve almost universal cooperation with enough frequency of interactions with the AA. Furthermore, the disagreement in assigned human reputations is highly affected by the AA, with disagreement decreasing under **SS** and increasing with **SH**, when interactions with the AA become more frequent. This suggests a deeper relation between cooperation

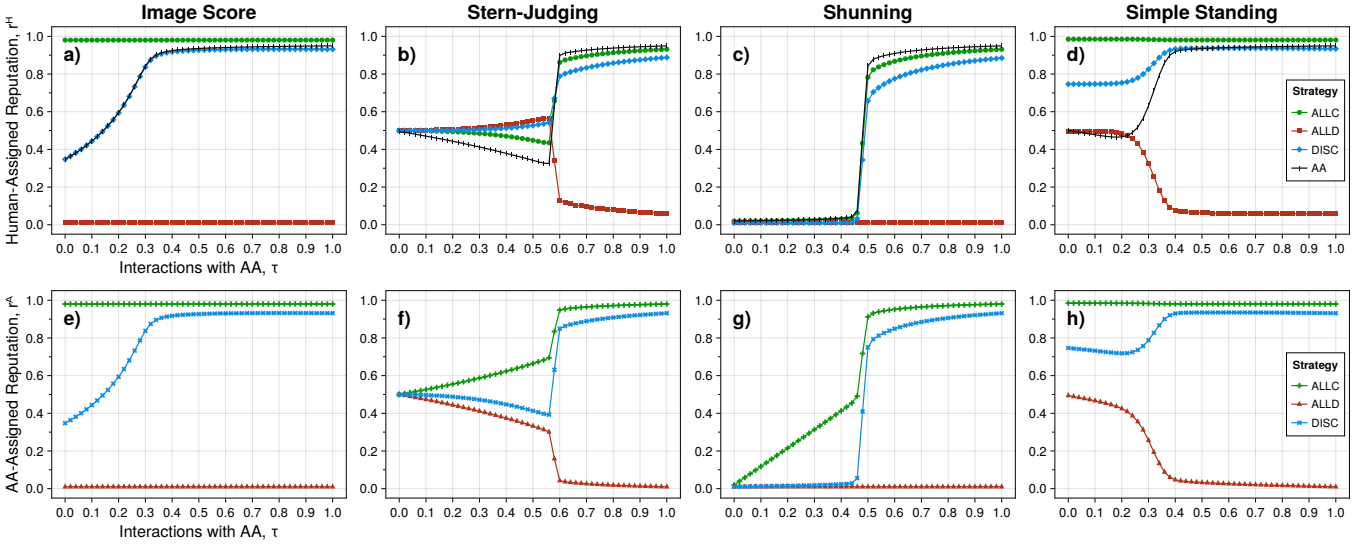

Figure 6: Average reputations, $r_s^H, s \in S$ and r_A^H (top) and $r_s^A, s \in S$ (bottom), at different frequencies of interaction with a *DISC* AA, when d^A = IS, and varying d^H. Introducing an AA increases the average human and AA-assigned reputation of *DISC* and/or *ALLC* relative to *ALLD*, resulting in more targeted punishments. We follow the same parameters as Figure 2, with $b/c = 3$.

and disagreement. Across all norms, the reputations of cooperators (DISC / ALLC) and defectors (ALLD) becomes distinct from each other, assisting both human *DISCs* and the AA in avoiding cooperating with defectors and thus promoting cooperative strategies: the punishment dilemma of private indirect reciprocity is mitigated. Finally, by studying a simpler scenario where AAs are judged like humans, we clarify the impact of the simplified judgment against AAs: although cooperation is higher in **SJ** and **SS**, it is also lower under **SH**. Furthermore, **SJ** presents no reduction in disagreement and thus AAs no longer mitigate the punishment dilemma. This suggests that simplified AA judgment might prove more beneficial in triggering cooperation than more complex judgment rules.

We draw five major conclusions: **1)** For an AA to promote human cooperation through **IR**, it must reward cooperators, and possibly punish defectors, which requires it to be able to readily identify cooperative actors. We conclude that, for this requirement, even simple social norms such as Image-Scoring can achieve positive results. **2)** An AA can promote human cooperation by increasing the gap between human reputations assigned to cooperators and defectors, successfully mitigating the punishment dilemma. This allows for both better punishments of defectors, and greater rewarding of cooperators. This stems from the AA being able to assess humans through direct interactions at a greater scale than humans.**3)** More consensus does not imply higher cooperation and distinguishability. The AA is capable of both increases and decreases in consensus depending on the social norm. **4)** The performance of the AA in promoting cooperation and consensus can be both hindered or boosted by the simplified human judgment, depending on the social norm. In both cases, it is still possible to achieve widespread cooperation by the presence of the AA. In general, more permissive judgments of the AA permit more positive impact. And **5)** If interactions with the AA are frequent enough and the AA is capable of identifying defectors, humans will ultimately delegate reputation-based discrimination to the AA, allowing humans to adopt unconditional cooperation. This dynamic has natural connections with delegation in human-AI systems [12] and deserves further exploration.

It is essential that social interactive agents are designed considering their impacts in long-term human cooperation [30, 46]. We hope that our study inspires future Human-AI experimental research (e.g., see [5]), considering reputation dynamics, social norms and the cultural environment where AAs are placed. Many factors have proved relevant to trigger prosociality in humans through AAs [44], however, aspects such as characterizing the social norms in place during hybrid interactions, having AAs identify and adapt to social norms, or the role of transparency [24] when judging or receiving judgments from AAs remain unexplored. More fundamentally, experimental frameworks to analyze **IR** and compare it to theoretical models remain a challenge. Our work aims to provide a theoretical baseline for future Human-AI interaction studies under **IR** and to inform new models.

Finally, throughout this work we implied the capacity of AAs to actively discriminate agents based on their assigned reputations [65]. We clarify that this position was taken to study the consequences of such systems and not as a support for algorithmic discrimination. We restrict the scope of our results to the context of the donation game under **IR**, where reputations solely convey the cooperative nature of individuals and no other characteristic.

ACKNOWLEDGMENTS

We thank the ELLIS Unit Amsterdam for funding. F.P.S acknowledges funding from the Dutch Research Council (NWO): this publication is part of the project with file number OCENW.M.22.322 of the research programme Open Competitie ENW which is (partly) financed by the NWO.

REFERENCES

[1] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. 2020. Elessar: Ethics in Norm-Aware Agents.. In *Autonomous Agents and Multi-Agent Systems*, Vol. 20. 16–24.

[2] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, and et al. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (Aug. 2020), 18–28.

[3] Richard Alexander. 2017. *The Biology of Moral Systems.* Routledge.

[4] Nicolas Anastassacos, Julian Garcia, Stephen Hailes, Mirco Musolesi, et al. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021).* 115–123.

[5] Hüseyin Aydın, Kevin Dubois-Godin, Libio Goncalvez Braz, Floris den Hengst, Kim Baraka, Mustafa Mert Çelikok, Andreas Sauter, Shihan Wang, and Frans A Oliehoek. 2025. SHARPIE: A Modular Framework for Reinforcement Learning and Human-AI Interaction Experiments. *arXiv preprint arXiv:2501.19245* (2025).

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv arXiv:2108.07258* (2021).

[7] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th Autonomous Agents and Multi-Agent Systems.* 507–513.

[8] 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 (2021), 33–36.

[9] Terence D Dores Cruz, Isabel Thielmann, Simon Columbus, Catherine Molho, Junhui Wu, et al. 2021. Gossip and reputation in everyday life. *Philosophical Transactions of the Royal Society B* 376, 1838 (2021), 20200301.

[10] David Elliott, Walter Keen, and Lei Miao. 2019. Recent advances in connected and automated vehicles. *Journal of Traffic and Transportation Engineering* 6, 2 (2019), 109–131.

[11] Marc Esteva, Juan-Antonio Rodriguez-Aguilar, Carles Sierra, Pere Garcia, and Josep L Arcos. 2001. On the formal specification of electronic institutions. In *Agent Mediated Electronic Commerce: The European AgentLink Perspective.* Springer, 126–147.

[12] Elias Fernández Domingos, Inês Terrucha, Rémi Suchon, Jelena Grujić, Juan C. Burguillo, Francisco C. Santos, and Tom Lenaerts. 2022. Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific Reports* 12, 1 (May 2022), 8492.

[13] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

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

[15] Yuma Fujimoto and Hisashi Ohtsuki. 2023. Evolutionary stability of cooperation in indirect reciprocity under noisy and private assessment. *Proceedings of the National Academy of Sciences* 120, 20 (2023), e2300544120.

[16] Francesca Giardini and Rafael Wittek. 2019. *The Oxford Handbook of Gossip and Reputation.* Oxford University Press.

[17] Nathan Griffiths and Sarabjot Singh Anand. 2012. The impact of social placement of non-learning agents on convention emergence.. In *Autonomous Agents and Multi-Agent Systems*, Vol. 12. Citeseer, 1367–1368.

[18] Hao Guo, Chen Shen, Shuyue Hu, Junliang Xing, Pin Tao, Yuanchun Shi, and Zhen Wang. 2023. Facilitating cooperation in human-agent hybrid populations through autonomous agents. *iScience* 26, 11 (2023).

[19] Alyssa Hanson, Nichole D. Starr, Cloe Emnett, Ruchen Wen, Bertram F. Malle, and Tom Williams. 2024. The Power of Advice: Differential Blame for Human and Robot Advisors and Deciders in a Moral Advising Context *(HRI '24).* Association for Computing Machinery, New York, NY, USA, 240–249. https://doi.org/10.1145/3610977.3634942

[20] Bradley Hayes, Daniel Ullman, Emma Alexander, Caroline Bank, and Brian Scassellati. 2014. People help robots who help others, not robots who help themselves. In *The 23rd IEEE international symposium on robot and human interactive communication.* IEEE, 255–260.

[21] Chris Haynes, Michael Luck, Peter McBurney, Samhar Mahmoud, Tomáš Vítek, and Simon Miles. 2017. Engineering the emergence of norms: a review. *The Knowledge Engineering Review* 32 (2017), e18.

[22] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How Humans Judge Machines.* MIT Press.

[23] 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 (2018), 12241–12246.

[24] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1, 11 (2019), 517–521.

[25] Mari Kawakatsu, Taylor A Kessinger, and Joshua B Plotkin. 2024. A mechanistic model of gossip, reputations, and cooperation. *Proceedings of the National Academy of Sciences* 121, 20 (2024), e2400689121.

[26] Taylor A Kessinger, Corina E Tarnita, and Joshua B Plotkin. 2023. Evolution of norms for judging social behavior. *Proceedings of the National Academy of Sciences* 120, 24 (2023), e2219480120.

[27] Marcus Krellner and The Anh Han. 2022. Pleasing Enhances Indirect Reciprocity-Based Cooperation Under Private Assessment. *Artificial Life* 27, 3–4 (2022), 246–276.

[28] Marcus Krellner and The Anh Han. 2023. We both think you did wrong–How agreement shapes and is shaped by indirect reciprocity. *arXiv arXiv:2304.14826* (2023).

[29] Birgit Lugrin, Catherine Pelachaud, and David Traum (Eds.). 2021. *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* (1 ed.). Vol. 37. ACM.

[30] Birgit Lugrin, Catherine Pelachaud, and David Traum. 2022. *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application.* ACM.

[31] Bertram F Malle, Stuti Thapa Magar, and Matthias Scheutz. 2019. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Robotics and well-being* (2019), 111–133.

[32] Bertram F Malle and Matthias Scheutz. 2020. Moral competence in social robots. In *Machine ethics and robot ethics.* Routledge, 225–230.

[33] Sebastián Michel-Mata, Mari Kawakatsu, Joseph Sartini, Taylor A Kessinger, Joshua B Plotkin, and Corina E Tarnita. 2024. The evolution of private reputations in information-abundant landscapes. *Nature* (2024), 1–7.

[34] Andreasa Morris-Martin, Marina De Vos, and Julian Padget. 2019. Norm emergence in multiagent systems: a viewpoint paper. *Autonomous Agents and Multi-Agent Systems* 33 (2019), 706–749.

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

[36] Martin A Nowak and Karl Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393, 6685 (1998), 573–577.

[37] Martin A Nowak and Karl Sigmund. 2005. Evolution of indirect reciprocity. *Nature* 437, 7063 (2005), 1291–1298.

[38] Hisashi Ohtsuki and Yoh Iwasa. 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* 231, 1 (2004), 107–120.

[39] Hisashi Ohtsuki and Yoh Iwasa. 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* 239, 4 (2006), 435–444.

[40] Hisashi Ohtsuki and Yoh Iwasa. 2007. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *Journal of Theoretical Biology* 244, 3 (2007), 518–531.

[41] Hisashi Ohtsuki, Yoh Iwasa, and Martin A Nowak. 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* 457, 7225 (2009), 79–82.

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

[43] Isamu Okada, Tatsuya Sasaki, and Yutaka Nakai. 2017. Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. *Scientific Reports* 7, 1 (2017), 9737.

[44] Raquel Oliveira, Patrícia Arriaga, Fernando P. Santos, Samuel Mascarenhas, and Ana Paiva. 2021. Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior* 114 (Jan. 2021), 106547.

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

[46] Ana Paiva, Filipa Correia, Raquel Oliveira, Fernando Santos, and Patrícia Arriaga. 2021. Empathy and prosociality in social agents. In *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.* 385–432.

[47] Ana Paiva, Fernando Santos, and Francisco Santos. 2018. Engineering prosociality with autonomous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[48] Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. 2024. Human-AI coevolution. *Artificial Intelligence* (2024), 104244.

[49] 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 (2021), 23581.

[50] Alexandre S. Pires and Fernando P. Santos. 2024. Artificial Agents Facilitate Human Cooperation Through Indirect Reciprocity. In *Frontiers in Artificial Intelligence and Applications.* IOS Press. https://doi.org/10.3233/FAIA240869

[51] Alexandre S. Pires and Fernando P. Santos. 2025. Supporting Information and Code - Artificial Agents Mitigate the Punishment Dilemma of Indirect Reciprocity. https://doi.org/10.5281/zenodo.14870987

[52] Arunas L Radzvilavicius, Alexander J Stewart, and Joshua B Plotkin. 2019. Evolution of empathetic moral evaluation. *Elife* 8 (2019), e44269.

[53] Fernando P. Santos, Samuel Mascarenhas, Francisco C. Santos, Filipa Correia, Samuel Gomes, and Ana Paiva. 2020. Picky losers and carefree winners prevail in collective risk dilemmas with partner selection. *Autonomous Agents and Multi-Agent Systems* 34, 2 (Oct. 2020), 40.

[54] 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.

[55] 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 (2016), 37517.

[56] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. 2016. Social norms of cooperation in small-scale societies. *PLoS Computational Biology* 12 (2016), e1004709.

[57] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. 2018. Social norm complexity and past reputations in the evolution of cooperation. *Nature* 555, 7695 (2018), 242–245.

[58] Bastin Tony Roy Savarimuthu and Stephen Cranefield. 2011. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems* 7, 1 (2011), 21–54.

[59] Matthias Scheutz and Bertram F Malle. 2021. May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare* (2021), 89–102.

[60] Laura Schmid, Pouya Shati, Christian Hilbe, and Krishnendu Chatterjee. 2021. The evolution of indirect reciprocity under action and assessment generosity. *Scientific Reports* 11, 1 (2021), 17443.

[61] Sandip Sen and Stéphane Airiau. 2007. Emergence of norms through social learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence* (Hyderabad, India) *(IJCAI'07)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1507–1512.

[62] Gopal Sharma, Hao Guo, Chen Shen, and Jun Tanimoto. 2023. 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 (2023), 20230301.

[63] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. 2023. Practices for governing agentic AI systems. *Research Paper, OpenAI, December* (2023).

[64] Hirokazu Shirado, Shunichi Kasahara, and Nicholas A. Christakis. 2023. Emergence and collapse of reciprocity in semiautomatic driving coordination experiments with humans. *Proceedings of the National Academy of Sciences* 120, 51 (Dec. 2023), e2307804120.

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

[66] Karl Sigmund. 2010. *The Calculus of Selfishness*. Princeton University Press.

[67] Martin Smit and Fernando P. Santos. 2024. Learning Fair Cooperation in Mixed-Motive Games with Indirect Reciprocity. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 220–228. https://doi.org/10.24963/ijcai.2024/25 Main Track.

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

[69] Inês Terrucha, Elias Fernández Domingos, Francisco C. Santos, Pieter Simoens, and Tom Lenaerts. 2024. The art of compensation: How hybrid teams solve collective-risk dilemmas. *PLoS One* 19, 2 (2024), e0297213.

[70] Arne Traulsen, Martin A Nowak, and Jorge M Pacheco. 2006. Stochastic dynamics of invasion and fixation. *Physical Review E* 74, 1 (2006), 011909.

[71] Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. 2024. A new sociology of humans and machines. *Nature Human Behaviour* 8, 10 (2024), 1864–1876.

[72] Satoshi Uchida. 2010. Effect of private information on indirect reciprocity. *Physical Review E* 82, 3 (2010), 036111.

[73] Satoshi Uchida and Tatsuya Sasaki. 2013. Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos, Solitons & Fractals* 56 (2013), 175–180.

[74] Satoshi Uchida and Karl Sigmund. 2010. The competition of assessment rules for indirect reciprocity. *Journal of Theoretical Biology* 263, 1 (2010), 13–19.

[75] Nicolaas Godfried Van Kampen. 1992. *Stochastic processes in physics and chemistry*. Vol. 1. Elsevier.

[76] Manuela Veloso, Joydeep Biswas, Brian Coltin, and Stephanie Rosenthal. 2015. CoBots: robust symbiotic autonomous mobile service robots. In *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina) *(IJCAI'15)*. AAAI Press, 4423–4429.

[77] Daniel Villatoro, Sandip Sen, and Jordi Sabater-Mir. 2010. Of social norms and sanctioning: A game theoretical overview. *International Journal of Agent Technologies and Systems (IJATS)* 2, 1 (2010), 1–15.

[78] Jason Xu, Julian Garcia, and Toby Handfield. 2019. Cooperation with bottom-up reputation dynamics. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 269–276.

[79] H Peyton Young. 2015. The evolution of social norms. *Economics* 7, 1 (2015), 359–387.