The Impact of Artificial Agents in Human Cooperation Through Indirect Reciprocity

Doctoral Consortium

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

ABSTRACT

Indirect reciprocity (IR), where reputations indicate with whom to cooperate or defect, is one of the key mechanism for supporting prosocial behavior among unrelated individuals. This mechanism has been studied in the context of human-human interactions. However, as artificial intelligence systems continue to be deployed, the introduction of artificial agents (AAs) in society has the potential to fundamentally alter reputations' assignment and spread, affecting cooperation dynamics. AAs are fundamentally different from humans, and can vary in their characteristics: they can have a wide social reach, be centralized (e.g., chatbots) or decentralized (e.g., local LLMs), be physical or virtual, and more. Despite this, like humans, AAs can also assign and spread reputations, and cooperate or defect. My thesis focuses on creating a framework to study the possible impacts that artificial agents have on human cooperation through IR, using both theoretical models and user studies.

KEYWORDS

Indirect Reciprocity; Reputations; Hybrid Populations; Mixed-Motive Game; Cooperation; Consensus; Evolutionary Game Theory

ACM Reference Format:
Alexandre S. Pires. 2025. The Impact of Artificial Agents in Human Cooperation Through Indirect Reciprocity: Doctoral Consortium. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Human cooperation is a fundamental tool in society, allowing us to not only achieve common goals, but to ensure greater individual well-being. Cooperation can be formalized as an individual (donor) spending a cost c (energy, time, money), to provide a benefit b to another individual (receiver) [36]. As the cost of cooperating is always on the side of the donor, its rational choice is always to defect, even when $b > c$, despite cooperation providing a greater overall benefit — this is known as a *social dilemma* [15].

To mitigate this dilemma, humans evolved and developed many mechanisms that promote prosocial behavior [19]. One particular mechanism is that of reputations, which help establish cooperation among unrelated individuals [21]: even if two individuals have never directly interacted before, they might have observed or heard about the interactions of the other (e.g., via gossiping [9]). In an interaction, there is, therefore, an incentive to cooperate in order to maintain a good reputation, ensuring future cooperation [2]. Similarly, reputations are used to determine who we cooperate with, as those with bad reputations are more likely to exploit us. This mechanism is known as **Indirect Reciprocity (IR)**.

With the advent of artificial intelligence (AI), there has been much effort to develop artificial agents (AAs) that not only can interact among themselves, but also with humans, resulting in a *hybrid population* (humans and AAs) [1, 7, 11]. As these agents can be both virtual (e.g. chatbots [5]) or physical (e.g. robots [30, 40]), be centralized or independent, human-like or machine-like, and much more, a vast new research direction has opened to understand their potential impact on prosocial behavior [6, 26].

There is a vast body of literature on **IR** and its role in human cooperation [20, 25]. A large focus lies in studying social norms, the rules that dictate what reputation an individual is assigned following an interaction. Furthermore, its emergence and evolution [21, 27, 41, 43], complexity [34], stability [23, 24], and relationship with culture and morality [14, 22] have been widely studied in many contexts. Besides reputation assignment, it is also crucial to understand how reputations spread, and how that spreading affects cooperation. Models of **IR** typically consider one of two categories of reputation spreading: **public reputations** [20, 32, 37], where all individuals are assumed to share assessments, usually as a consequence of rapid gossip; or **private reputations** [12, 17, 29], where each individual has its own view of every other individual, and no gossip is present. More recently, some models [18] are capable of partial gossip to more closely follow the real-world.

Similar to humans, AAs are also capable of discriminating and assigning reputations following their own social norms [3, 39], and of spreading reputations. In addition, it has also been shown that humans judge AAs differently than they judge other humans [16], suggesting that social norms are also dependent on the type of individuals involved. While frameworks exist to assess **IR** and cooperation among AAs [13], understanding how human cooperation will be impacted by AAs, as well as how to develop AAs that promote, or at least do not damage human cooperation is therefore a pressing matter [1, 8]. In my thesis, I aim to address the following questions, via a combination of theoretical and empirical methods:

- Q1: Under which scenarios can AAs promote, sustain or hinder human cooperation through **IR**?
- Q2: Can a theoretical framework be developed to predict the resulting impact of AAs in human cooperation through **IR**?

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).

2 ARTIFICIAL AGENTS IN PUBLIC REPUTATIONS

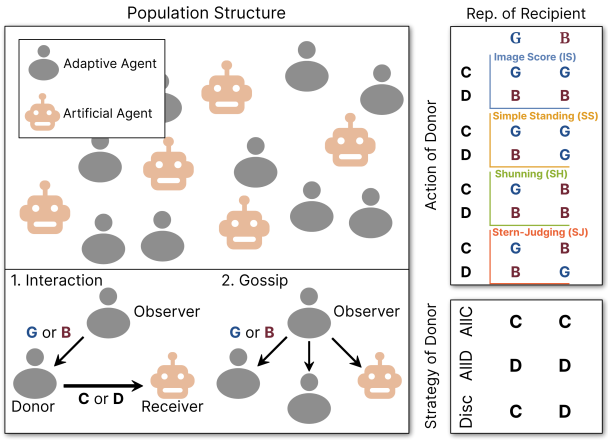

Figure 1: We consider a hybrid population consisting of adaptive agents (representing humans), in gray, and artificial agents, in orange. Interactions involve three agents: a Donor, Receiver and Observer(s). The Donor chooses whether to cooperate or defect (C or D) with the Receiver. The interaction is witnessed by 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, becoming public knowledge. When reputations are private, this step is partial or non-existent, thus disagreements can exist. Top right: Reputations are assigned following a common social norm that determines the next reputation of a donor given the reputation of the recipient and the action of the donor. Bottom right: The three strategies considered.

My first work [28] proposes an evolutionary game theoretical model to explore hybrid populations when reputations are public. In this scenario, each individual is assigned a binary reputation: **Good, G** or **Bad, B**. Following previous models of **IR** [22, 23, 33], each agent is capable of using one of various strategies: *ALLC*, which always cooperates, *ALLD*, which always defects, and the discriminator strategy *DISC*, which cooperates with good individuals and defects with bad. Adaptive agents (representing humans) [36], will adopt strategies with a probability that is proportional to the payoff of each strategy. AAs, on the other hand, are defined as *hard-coded* agents, whose strategy is preselected and constant in time [31, 38]. This assumption does not mean that AAs do not employ learning algorithms – although the execution of actions can be complex, we abstract this complexity and focus instead on the resulting strategy of the AA. An illustration of the model is presented in Figure 1.

We study the prevalence of cooperation [33] by measuring the probability that, at any timestep, we observe cooperation. By varying the proportion of AAs relative to humans, the strategy of AAs, and the social norm used, we made a first assessment of the potential impact of AAs in human cooperation through **IR**. Our primary conclusions are that for cooperation to be promoted, AAs must employ a *DISC* strategy, rewarding cooperators while punishing defectors. AAs that cooperate unconditionally instead incentivize humans to defect, undermining cooperation. Furthermore, *ALLD* AAs are capable of completely dissolving cooperation, raising concerns about the resilience of cooperation. We also study the impact of biases against AAs in the form of a fixed reputation: we show how a negative view of AAs is capable of undermining their benefit, but a positive view can lead to a greater boost in cooperation.

3 (CENTRALIZED) ARTIFICIAL AGENTS IN PRIVATE REPUTATIONS

My second work studies private reputations. Here, cooperation is notably more difficult to achieve due to disagreements between individuals. The primary method to sustain cooperation using reputations is to defect against (punish) exploiters and cooperate with (reward) cooperators, via the *DISC* strategy. However, without gossip, it is difficult to agree if a defection was an exploitation or a punishment, which generates further disagreements. This is known the *punishment dilemma*. This work thus focuses on understanding how AAs can help mitigate this dilemma. Additionally, recent work has shown that humans judge AAs primarily by their actions [16], as opposed to humans, which are judged by their intentions. This implies that different social norms are used depending on the types of agent involved, which we model by adapting [18] to distinguish between the reputations of each type of agent.

We highlight that the major factor enabling the punishment dilemma is that the average reputations of *DISC* and *ALLC* are close to that of *ALLD*, leading *DISC* agents to ineffectively punish defectors. However, a *DISC* AA is capable of promoting cooperation between humans by increasing this reputation gap, effectively mitigating the punishment dilemma. This stems from the AAs being both capable of more interactions than humans, and of AAs being judged differently, allowing humans to observe more interactions of others with AAs, and thus distinguish *ALLD* from cooperators.

4 ONGOING AND FUTURE WORK

More detailed models: My prior models, although insightful, lack a connection with recent advancements in AI, particularly LLMs [5]. LLMs, in contexts like chatbots, are unique, as they can be sources of influence [4] to individuals without directly participating in social dilemmas. Furthermore, LLMS are highly accessible and possess their own social norms [35]. As such, I am studying the social norms of current LLM models, and their effects in cooperation.

Experimental work: My past work used abstractions of AAs and humans, and although **IR** has been experimentally verified in human populations [10, 42], gaps between theoretical and empirical work still exist. To this end, I am developing user studies, based on measuring human-assigned reputations in hybrid interactions, to understand what social norms are used between humans and AAs, and to clarify the role of **IR** in hybrid populations.

Fundamental work in IR: Modelling **IR** in hybrid populations is challenging due to the many asymmetries between humans and AAs, such as reputation assignment and spreading, social norms, network structure, interactions, and proportion in the population. In addition, the accessibility of AAs can range from an agent per human, many in the population, or even a single agent partially or fully accessible by the entire population. To capture these scenarios, I am developing fundamental advancements in **IR** modelling.

REFERENCES

[1] 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.

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

[3] 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.

[4] Hui Bai, Jan Voelkel, Johannes Eichstaedt, and Robb Willer. 2023. Artificial Intelligence Can Persuade Humans on Political Issues. Preprint (Version 1), Research Square. https://doi.org/10.21203/rs.3.rs-3238396/v1 Posted on September 7, 2023.

[5] 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).

[6] Jacob W Crandall, Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, et al. 2018. Cooperating with machines. *Nature Communications* 9, 1 (2018), 233.

[7] 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.

[8] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. *arXiv arXiv:2012.08630* (2020). Publisher: arXiv Version Number: 1.

[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] Martin Dufwenberg, Uri Gneezy, Werner Güth, and EEC van Damme. 2001. Direct versus indirect reciprocity: An experiment. *Homo Oeconomicus-Journal of Behavioral and Institutional Economics* 18, 1/2 (2001), 19–30.

[11] 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.

[12] 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.

[13] Karen K Fullam, Tomas B Klos, Guillaume Muller, Jordi Sabater, Andreas Schlosser, Zvi Topol, K Suzanne Barber, Jeffrey S Rosenschein, Laurent Vercouter, and Marco Voss. 2005. A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems.* 512–518.

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

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

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

[17] 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.

[18] 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.

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

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

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

[22] 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.

[23] 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.

[24] 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.

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

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

[27] 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.

[28] 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

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

[30] Stephanie Rosenthal, Joydeep Biswas, and Manuela M Veloso. 2010. An effective personal mobile robot agent through symbiotic human-robot interaction.. In *Autonomous Agents and Multi-Agent Systems*, Vol. 10. 915–922.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems* 36 (2024).

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

[37] 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.

[38] 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.

[39] Aron Vallinder and Edward Hughes. 2024. Cultural Evolution of Cooperation among LLM Agents. *arXiv preprint arXiv:2412.10270* (2024).

[40] 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.

[41] 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.

[42] Erez Yoeli, Moshe Hoffman, David G Rand, and Martin A Nowak. 2013. Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences* 110, supplement_2 (2013), 10424–10429.

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