Prosocial dynamics in multiagent systems

Santos, F.P.

DOI
10.1002/aaai.12143
Publication date
2024
Document Version
Final published version
Published in
AI Magazine

Received: 5 July 2023 | Accepted: 6 October 2023

DOI: 10.1002/aaai.12143

HIGHLIGHT

Prosocial dynamics in multiagent systems

Fernando P. Santos

Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Correspondence
Fernando P. Santos, Informatics Institute, University of Amsterdam, Science Park 900, 1098 XH Amsterdam, The Netherlands.
Email: f.p.santos@uva.nl

Funding information
Fundação para a Ciência e a Tecnologia; James S. McDonnell Foundation

Abstract

Meeting today's major scientific and societal challenges requires understanding dynamics of prosociality in complex adaptive systems. Artificial intelligence (AI) is intimately connected with these challenges, both as an application domain and as a source of new computational techniques: On the one hand, AI suggests new algorithmic recommendations and interaction paradigms, offering novel possibilities to engineer cooperation and alleviate conflict in multiagent (hybrid) systems; on the other hand, new learning algorithms provide improved techniques to simulate sophisticated agents and increasingly realistic environments. In various settings, prosocial actions are socially desirable yet individually costly, thereby introducing a social dilemma of cooperation. How can AI enable cooperation in such domains? How to understand long-term dynamics in adaptive populations subject to such cooperation dilemmas? How to design cooperation incentives in multiagent learning systems? These are questions that I have been exploring and that I discussed during the New Faculty Highlights program at AAAI 2023. This paper summarizes and extends that talk.

INTRODUCTION

Prosociality is puzzling (Gintis 2003): prosocial individuals contribute to benefiting others, yet they must often incur a cost to do so. Why do such altruistic behaviors exist and are not outcompeted by selfish ones? (Pennisi 2005) And how to harness artificial intelligence applications to sustain prosociality within systems of artificial learning agents and humans? (Paiva, Santos, & Santos 2018). Solving the puzzle of prosociality is an essential endeavor to tackle some of the most pressing challenges that our society faces.

Understanding the roots of cooperation, and the institutions, social norms, and artifacts that might sustain it, is fundamental in various domains—from climate change (Bisaro & Hinkel 2016) and responsible use of natural resources (Dietz, Ostrom, & Stern 2003) to pandemic control (Traulsen, Levin, & Saad-Roy 2023). In international relations, cooperation is still fundamental to prevent *arms races, nuclear proliferation, and military escalation*, as noted already in the 80s (Axelrod 1984). The efforts to comprehend human prosociality are long-standing yet unsettled.

Beyond human groups, understanding prosocial behavior is fundamental in multiagent systems. In these systems, multiple computational agents, with a varying degree of autonomy, attempt to fulfill their goals while interacting with other artificial agents (Wooldridge 2009). If agents can learn and adapt over time, it is important to understand how to design interaction rules and learning protocols that incentivize cooperation and guarantee satisfactory long-term rewards—fulfilling the previously named *prescriptive* agenda of noncooperative game theory in multiagent learning (Shoham, Powers, & Grenager 2007). Prosociality can here be measured as the probability

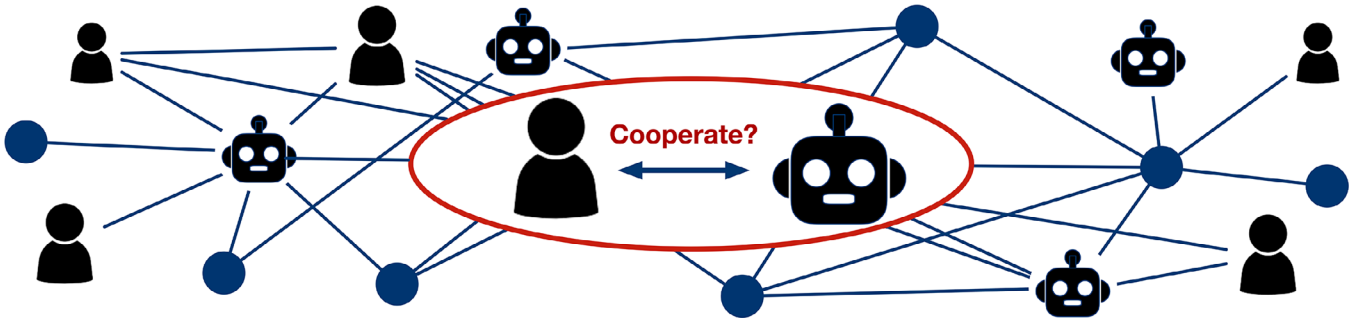

FIGURE 1 Humans and artificial intelligent applications form, nowadays, complex systems. Understanding dynamics of prosociality in multiagent systems can benefit from the application of tools used in fields such as population dynamics and network science.

that agents learn to use a strategy leading to high collective benefits, even if sacrificing individual payoffs. Although systems of artificial agents can be directly designed to cooperate with others, the problem of designing prosocial systems remains under decentralized control, where each agent—eventually representing different humans or organizations—aims at independently maximizing long-term payoffs.

The problems of cooperation in multiagent systems and human societies are no longer independent. Humans coexist with artificial agents, both in the physical world and on online platforms. The challenge of understanding human cooperation is today entangled with the challenge of designing artificial agents and algorithms that facilitate prosocial interactions both online and offline (Crandall et al. 2018; Oliveira et al. 2021; Akata et al. 2020; Guo et al. 2023). Moreover, understanding human cooperation can provide invaluable knowledge on how to design artificial cooperation (and vice versa).

Understanding dynamics of prosociality in multiagent systems can benefit from the application of tools typically used in complex adaptive systems (see Figure 1). Such tools can contribute to apprehend how simple interventions (e.g., agents with a modified behavior, new interaction rules, or new sources of information) can affect the long-term macro dynamics in a system composed by many learning agents. Apart from understanding which actions agents are likely to take—and subsequent probabilities of cooperation among agents—one can also grasp the dynamics leading to such states, how long the process will take, when to intervene, and whether behaviors can ever become stable. This analysis can benefit from methods borrowed from theoretical ecology and population dynamics.

In this paper, written in the context of the AAAI 2023 New Faculty Highlights program, I summarize five decision-making domains where, I believe, a combination of tools at the interface of AI, multiagent systems, and population dynamics can improve our abilities to design increasingly prosocial systems. This paper focuses on prosociality in the context of (1) **reputation** systems, (2) **recommender** systems, (3) **hybrid** systems, (4) **classification** systems, and (5) multisector **urban** systems—summarized in Figure 2. Although seemingly unrelated, these five domains share commonalities: they constitute areas where understanding the interrelated dynamics of humans and agents' behavior is essential; and they constitute domains where achieving socially desirable outcomes requires solving social dilemmas of cooperation and prosociality.

Prosociality in reputation systems

Reputation systems are a fundamental mechanism to elicit trust among strangers and a backbone of e-commerce, crowdsourcing marketplaces, and sharing economic platforms (Resnick et al. 2000). Reputation systems also play a central role in multiagent system when artificial agents must select trustworthy partners or adapt based on information about opponents' prior interactions (Pinyol & Sabater-Mir 2013). In the realm of evolutionary biology, reputations are a central mechanism to explain cooperation through indirect reciprocity (Nowak & Sigmund 2005). In this regard, a fundamental challenge is understanding which rules to assign reputation are more likely to elicit long-term stable cooperation (Ohtsuki & Iwasa 2004).

Indirect reciprocity has been identified as a key mechanism to explain the evolution of cooperation among humans (Nowak & Sigmund 2005). Agents are assumed to adopt strategies determining which action to employ (be cooperative or not) when interacting with another agent. Importantly, the decision of which action to select depends on reputations; agents can restrict cooperation to those that have a specific reputation. After each interaction, the reputations of interacting agents are updated. This update follows a social norm defining which actions should lead to

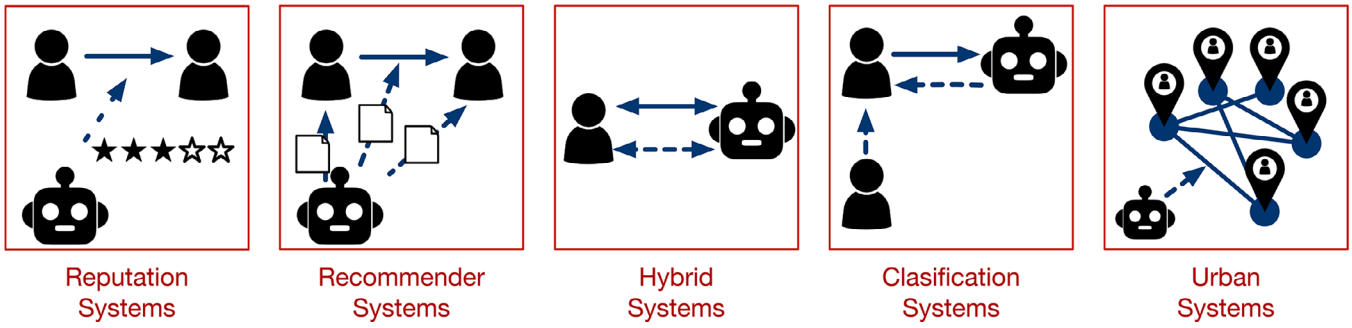

FIGURE 2 Five domains where understanding prosocial dynamics is beneficial to designing collective systems where humans co-exist with artificial intelligence applications. Dashed arrows represent information transmission; full arrows represent interactions where agents can decide to cooperate (i.e., act prosocially) or defect. (1) In **reputation systems**, humans decide to cooperate or defect with each other and, after that, their reputation is updated and eventually spread on online platforms. (2) In **recommender system**, AI affects information sources humans are exposed to, which in turn can affect their decision to cooperate. (3) In **hybrid systems**, humans and social artificial agents directly decide to cooperate or defect with each other, based on information and signals exchanged. (4) In classification systems, humans can use information provided by transparent algorithms or human peers to change their features and change the outcome of a classification algorithm. (5) In **urban systems,** AI is used to plan and design city infrastructure, and to offer citizens new services and recommendations; adopting new technologies and shifting to new paradigms depends on multisector decisions and the willingness of stakeholders to act prosocially.

a good reputation. In this sense, indirect reciprocity norms resemble injunctive norms studied in social psychology, which postulate the behaviors one is expected to follow (Bicchieri 2005).

Determining which social norms lead to higher levels of cooperation under indirect reciprocity is computationally challenging. The number of potential norms increases exponentially with the number of bits needed to define an interaction (Santos, Pacheco, & Santos 2021), and the ultimate cooperative levels of norms depend on a dynamical process where strategies co-evolve with reputations in potentially large populations. The challenges of identifying cooperative norms are augmented in group-structured populations, a setting where assigning reputations can depend on both prior actions and group identities (Smit & Santos 2023; Romano, Balliet, & Wu 2017; Whitaker, Colombo, & Rand 2018). Besides computational complexity, the study of indirect reciprocity norms calls for the formalization of cognitive complexity (Santos, Santos, & Pacheco 2018; Santos, Pacheco, & Santos 2021). Even assuming the simple setting of binary actions and binary reputations, norms considered can encode very complex judgments, whose applicability in real settings involving humans is questionable. Formalizing complexity in indirect reciprocity—and, in general, reputation systems—allows us to search for reputation assignment rules and strategies that maximize prosociality while keeping simplicity and interpretability.

Reputations can enable cooperation. Yet reputation systems can themselves require selfless information sharing, relying on users' prosociality. Sharing one's experiences on online platforms about interactions with others requires time and effort. If sharing reputations is costly, cooperation under indirect reciprocity involves a second-order social dilemma, whereby sharing reputations itself needs to be incentivized (Sasaki, Okada, & Nakai 2016; Santos, Pacheco, & Santos 2018).

Prosociality in recommender systems

Recommender systems are, nowadays, one of the most impactful and widespread applications of artificial intelligence (Ricci, Rokach, & Shapira 2021). In their essence, recommended systems suggest items that users are likely to find relevant. Items can be objects to purchase, music, videos, jobs, news, or even other users to connect with on online social platforms. In a world where information is shared at unprecedent rates, recommender systems are an important tool to cope with information overload. Recommender systems are advantageous to producers and users alike: the first can improve the outreach of items produced and ultimately add value to their business; the latter can identify new products, discover interesting items, and satisfy their needs more expeditiously.

Recommender systems suggest yet another domain where humans co-exist with artificial intelligence algorithms and fully understand their co-evolving dynamics can benefit from applying population dynamics methods (Piao et al. 2023; Santos 2023). Grasping the impacts of recommender systems on human societies also requires capturing how these systems impact prosociality. This is evident in applications such as news recommendation

and link recommendation algorithms on online social networks, which can impact how information spreads and, consequently, the perceived costs/benefits of cooperation and collective action.

The challenges of incentivizing cooperation to solve some of our most pressing societal problems can be captured by simple economic games such as nonlinear public goods games. These are interactions where attaining collective success requires that a critical mass of cooperators exist. As cooperation involves a cost, reaching the minimal number of cooperators required for cooperative efforts to become consequential is not an easy task. This is the challenge, for instance, when countries are called to cooperate by reducing CO_2 emissions (Milinski et al. 2008; Santos & Pacheco 2011) or when individuals are asked to cooperate by wearing masks to prevent a virus from spreading (Traulsen, Levin, & Saad-Roy 2023). In these domains, underestimating the cooperative efforts of individuals around us might impact our own willingness to cooperate—in fact, humans often reveal to be conditional cooperators (Fischbacher, Gächter, & Fehr 2001). This raises the question of how social perception biases can affect cooperation in nonlinear public goods dilemmas. In a previous work, we have shown that perception biases leading to false uniqueness or false consensus effects can hamper cooperation and collective action (Santos, Levin, & Vasconcelos 2021). Recommender systems that filter information one has access to—in particular, about opinions of others—can exacerbate such effects; these recommender systems should be evaluated not only in terms of creating echo chambers, information cocoons, or filter bubbles, but also regarding our willingness to behave prosocially.

Besides filtering information, recommendation algorithms can directly affect the way social networks evolve by directly recommending who should be connected with whom (Su, Sharma, & Goel 2016). These link recommendation algorithms can possibly exacerbate the community structure of networks affecting levels of polarization and radicalization (Santos, Lelkes, & Levin 2021). Networks have, in turn, a direct connection with the evolution of cooperative behavior (Rand, Arbesman, & Christakis 2011; Santos, Pacheco, & Lenaerts 2006; Shirado & Christakis 2020), which suggest that social recommenders on social media can also affect our prosocial dynamics.

Prosociality in hybrid systems

It is clear nowadays that humans co-exist with algorithms, as previous examples also evidence in the domain of online platforms. But humans are increasingly interacting with social artificial agents, with varying degrees of autonomy. These social agents can be simple social media bots (Ferrara et al. 2016) or embodied socially interactive agents (Lugrin 2021). The latter are autonomous agents that can perceive their environment, including people or other agents, decide how to interact, and express attitudes, emotions, engagement, or even empathy. Also in this domain, it is fundamental to understand how to design agents that behave prosocially and sustain human prosociality (Paiva et al. 2021).

Prosociality in hybrid populations composed of humans and artificial social agents depends on humans' willingness to adapt their behavior according to the behavior of an artificial opponent (and vice versa). It is not, however, clear that humans will choose artificial partners and reciprocate cooperative actions similarly to what they do when interacting with other humans. Cooperation with artificial agents depends on trust and transparency (Han, Perret, & Powers 2021; Ishowo-Oloko et al. 2019). In the short term, experiments in environments where humans interact with robotic partners and virtual agents can reveal whether humans' reciprocal behaviors are such that we can expect cooperation stability (Santos et al. 2020; Santos et al. 2019; de Melo, Santos, & Terada 2023).

To infer how prosocial behaviors will develop in the long run, one can resort to agent-based simulations and population dynamics models. These models illustrate the long-term effects of introducing, in a population of adaptive learning agents (like humans) a subset of agents with a predetermined behavior. These behaviors can be engineered in such a way that a small fraction of agents can trigger long-lasting prosocial behaviors (Santos et al. 2019).

Prosociality in classification systems

Artificial Intelligence applications are, currently, used in many consequential applications, especially when they are used to classify humans. Classification algorithms are used, for example, in loan applications, fraud detection, college admission, or automated recruitment tools. In this context, algorithms should be increasingly transparent, allowing subjects to understand how and why algorithmic decisions were performed and eventually offering the possibility of recourse. Humans' adaptation after algorithmic decisions can be relevant to revert unfair decisions and allow people to improve their condition. On the other hand, individuals might adapt in malicious ways by, for example, manipulating the information provided. The challenge of designing classification algorithms that are robust to strategic manipulation by rational agents is studied in the field of *strategic classification* (Hardt et al. 2016).

The study of prosociality in large populations of adaptive agents can also be informative in the context of strategic

classification. When subject to the results of an algorithmic decision, individuals can choose to improve their condition—thereby incurring high effort to improve their chances of future success—or choose to game the system—for example, by providing false information or strategically adapt features in ways that do not cause future success (Kleinberg & Raghavan 2020; Miller, Milli, & Hardt 2020; Barsotti, Koçer, & Santos 2022). Improving means that individuals are required to pay a high cost to adapt and thereby concede classifiers the benefit of keeping high accuracy. Gaming means that individuals will pay a low individual cost, however reducing the accuracy level of the classifier. As in the case of altruistic cooperation, strategic classification suggests a social dilemma which, to be solved, requires prosocial agents.

In another direction, the way individuals strategically adapt to algorithms might depend on information collected from peers and from online platforms (Ghalme et al. 2021; Bechavod et al. 2022; Barsotti, Koçer, & Santos 2022). Disclosing truthful information for this purpose entails a second-order social dilemma, just as the challenge of costly reputation sharing previously discussed: individuals are required to spend time and effort (i.e., spend a cost) to offer others valuable information about their experiences, which hopefully contribute to others' possibility of algorithmic recourse (Karimi et al. 2022).

Prosociality in urban systems

Planning more livable and inclusive cities also constitutes a domain where we can benefit from a better understanding prosocial dynamics in scenarios where citizens co-exist with artificial intelligence applications (Stein & Yazdanpanah 2023). Prosociality is relevant when people decide to recycle, consume resources responsibly, take good care of public urban spaces, or take an active role in their communities (Santos & Bloembergen 2019; Hsu et al. 2020; Arana-Catania et al. 2021; Hsu et al. 2022). The connection between prosociality, AI, and urban systems is also evident in the case of route recommender systems, where following AI recommendations might lead to detrimental outcomes such as higher pollution levels (Cornacchia et al. 2022): will citizens be willing to accept algorithmic recommendations that are not individually optimal, yet contribute to the collective good?

At the planning level, understanding dynamics of decision-making between different sectors in a city (citizens, public sector, private sector) can shed light on the challenges to implement new initiatives or to adopt new technologies (Santos et al. 2016; Encarnação et al. 2016). A key example is the adoption of green technologies such as developing infrastructure for electric vehicles (Encarnação et al. 2018). Also here, understanding how to harness incentives to trigger prosocial behaviors is fundamental. Often, multiple sectors have competing goals, and unlocking new projects that benefit citizens might require that a particular stakeholder (e.g., public or private sector) incurs a cost to initiate a transition to a more desirable state (Encarnação et al. 2018). It is fundamental to understand which sector has a more decisive role, and how to harness the right incentives to guarantee sustained urban transitions.

Artificial intelligence applications can also be used to search the large space of possible options when deciding how to improve public services such as public transportation. When designing new public transportation transit schedules, routes, or lines, city planners might face fairness dilemmas: adding a new line might unequally favor different communities in a city (Michailidis, Ghebreab, & Santos 2023). When expanding public transit offer in an inclusive way, it might be necessary for a majority group to accept a higher cost to improve urban mobility to marginalized groups. The connection between prosociality, AI, and mobility in urban systems also extends to the domain of residential mobility (Bara, Santos, & Turrini 2023; Michailidis et al. 2023): preventing urban segregation might imply that individuals behave prosocially and support interventions that facilitate interactions with diverse communities.

CONCLUSION

This paper, written in the context of the AAAI-23 New Faculty Highlights program, features our previous research in the domain of prosocial dynamics in multiagent systems. Besides revisiting past work, this paper suggests a base for a future research agenda on advancing our tools and knowledge on how to design artificial intelligence applications that sustain prosociality across decision-making domains. Artificial Intelligence relates to the challenge of sustaining prosocial action. As presented in this paper, as an application field and as source of computational techniques. Second, AI suggests new interaction paradigms that involve groups of artificial agents and humans, offering new possibilities to engineer cooperation in multiagent (hybrid) systems. On the other hand, new learning algorithms provide improved techniques to simulate sophisticated agents and analyze increasingly realistic systems where cooperation is paramount.

The works showcased in this paper resort to a combination of techniques at the interface of multiagent systems and complex systems. In particular, the findings presented result from applying (evolutionary) game theory, multiagent reinforcement learning, network science,

and, more broadly, agent-based simulations. New techniques, inspired by the new paradigms of deep learning, graph representation learning, and foundation models, are promising in the domain of prosocial dynamics (Hughes et al. 2018; Dafoe et al. 2021). Extending current methods to cope with agents and human communities' heterogeneity can certainly offer fruitful new research lines (Merhej et al. 2022). Finally, understanding cooperation dynamics can be relevant to the own process of governing and regulating AI (Han et al. 2020; Han et al. 2022).

While this survey focuses on works applying techniques commonly used in computer science, the topic of cooperation and prosociality is naturally multidisciplinary. Advancing our knowledge of prosocial artificial systems can benefit from the input of biology, anthropology, psychology, philosophy, behavioral economics, to name some examples. Evolutionary theory, economic experiments, and anthropological case studies shed light on why and how humans cooperate, providing a basis to anticipate how contemporary technology might impact human prosociality (Skyrms 2004; Henrich & Henrich 2007). Ultimately, understanding cooperation in artificial systems can only be accomplished through cooperation between multiple fields.

ACKNOWLEDGMENTS

The work presented in this paper was supported by FCT-Portugal, the James S. McDonnell Foundation, and the Netherlands Innovation Center for AI (ICAI). The author is thankful to Sennay Ghebreab and anonymous reviewers for enriching comments.

CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict.

ORCID

Fernando P. Santos https://orcid.org/0000-0002-2310-6444

REFERENCES

Akata, Zeynep, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, and Holger Hoos. 2020. "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence." *Computer* 53(8): 18–28.

Arana-Catania, Miguel, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. "Citizen Participation and Machine Learning for a Better Democracy." *Digital Government: Research and Practice* 2(3): 1–22.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bara, Jacques, Fernando P. Santos, and Paolo Turrini. 2023. "The Role of Space, Density and Migration in Social Dilemmas." In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*. ACM.

Barsotti, Flavia, Rüya G. Koçer, and Fernando P. Santos. 2022. "Transparency, Detection and Imitation in Strategic Classification." In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI*, vol. 2022.

Bechavod, Yahav, Chara Podimata, Steven Wu, and Juba Ziani. 2022. "Information Discrepancy in Strategic Learning." In *Proceedings of the International Conference on Machine Learning (ICML 2022)*, 16911715. PMLR.

Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bisaro, Alexander, and Jochen Hinkel. 2016. "Governance of Social Dilemmas in Climate Change Adaptation." *Nature Climate Change* 6(4): 354–59.

Cornacchia, Giuliano, Matteo Böhm, Giovanni Mauro, Mirco Nanni, Dino Pedreschi, and Luca Pappalardo. 2022. "How Routing Strategies Impact Urban Emissions." In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–4.

Crandall, Jacob W., Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A. Goodrich, and Iyad Rahwan. 2018. "Cooperating with Machines." *Nature Communications* 9(1): 233.

Dafoe, Allan, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. "Cooperative AI: Machines Must Learn to Find Common Ground." *Nature* 593(7857): 3336.

Dietz, Thomas, Elinor Ostrom, and Paul C. Stern. 2003. "The Struggle to Govern the Commons." *Science* 302(5652): 1907–12.

Encarnação, Sara, Fernando P. Santos, Francisco C. Santos, Vered Blass, Jorge M. Pacheco, and Juval Portugali. 2016. "Paradigm Shifts and the Interplay between State, Business and Civil Sectors." *Royal Society Open Science* 3(12): 160753.

Encarnação, Sara, Fernando P. Santos, Francisco C. Santos, Vered Blass, Jorge M. Pacheco, and Juval Portugali. 2018. "Paths to the Adoption of Electric Vehicles: An Evolutionary Game Theoretical Approach." *Transportation Research Part B: Methodological* 113: 24–33.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. "The Rise of Social Bots." *Communications of the ACM* 59(7): 96–104.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters* 71(3): 397–404.

Ghalme, Ganesh, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. 2021. "Strategic Classification in the Dark." In *Proceedings of the International Conference on Machine Learning (ICML 2022)*, 36723681. –PMLR.

Gintis, Herbert. 2003. "Solving the Puzzle of Prosociality." *Rationality and Society* 15(2): 155–87.

Guo, Hao, 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): 108179.

Han, The A., Tom Lenaerts, Francisco C. Santos, and Luis Moniz Pereira. 2022. "Voluntary Safety Commitments Provide an Escape from Over-Regulation in AI Development." *Technology in Society* 68: 101843.

Han, The A., Luis M. Pereira, Francisco C. Santos, and Tom Lenaerts. 2020. "To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race." *Journal of Artificial Intelligence Research* 69: 881921. https://doi.org/10.1613/jair.1.12225.

Han, The A., Cedric Perret, and Simon T. Powers. 2021. "When to (or Not to) Trust Intelligent Machines: Insights from an Evolutionary Game Theory Analysis of Trust in Repeated Games." *Cognitive Systems Research* 68: 111–24.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. "Strategic Classification." In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 111–22.

Henrich, Natalie, and Joseph P. Henrich. 2007. *Why Humans Cooperate: A Cultural and Evolutionary Explanation*. Oxford: Oxford University Press.

Hsu, Yen-Chia, Jennifer Cross, Paul Dille, Michael Tasota, Beatrice Dias, Randy Sargent, Ting-Hao Huang, and Illah Nourbakhsh. 2020. "Smell Pittsburgh: Engaging Community Citizen Science for Air Quality." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10(4): 1–49.

Hsu, Yen-Chia, Himanshu Verma, Andrea Mauri, Illah Nourbakhsh, and Alessandro Bozzon. 2022. "Empowering Local Communities Using Artificial Intelligence." *Patterns* 3(3): 100449.

Hughes, Edward, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, and Raphael Koster. 2018. "Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas." In Advances in Neural Information Processing Systems, 31.

Ishowo-Oloko, Fatimah, 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): 517–21.

Karimi, Amir-Hossein, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. "A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations." *ACM Computing Surveys* 55(5): 1–29.

Kleinberg, Jon, and Manish Raghavan. 2020. "How Do Classifiers Induce Agents to Invest Effort Strategically?" *ACM Transactions on Economics and Computation (TEAC)* 8(4): 1–23.

Lugrin, Birgit. 2021. "Introduction to Socially Interactive 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*, 1–20. New York: Association for Computing Machinery.

de Melo, Celso M., Francisco C. Santos, and Kazunori Terada. 2023. "Emotion Expression and Cooperation under Collective Risks." *iScience* 26: 108063.

Merhej, Ramona, Fernando P. Santos, Francisco S. Melo, and Francisco C. Santos. 2022. "Cooperation and Learning Dynamics under Wealth Inequality and Diversity in Individual Risk." *Journal of Artificial Intelligence Research* 74: 733–64.

Michailidis, Dimitris, Sennay Ghebreab, and Fernando P. Santos. 2023. "Balancing Fairness and Efficiency in Transport Network Design through Reinforcement Learning." In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2532–34.

Michailidis, Dimitris, Mayesha Tasnim, Sennay Ghebreab, and Fernando P. Santos. 2023. "Towards Reducing School Segregation by Intervening on Transportation Networks." In Citizen-Centric Multiagent Systems 2023 (CMAS'23), 4.

Milinski, Manfred, Ralf D. Sommerfeld, Hans-Jürgen Krambeck, Floyd A. Reed, and Jochem Marotzke. 2008. "The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change." *Proceedings of the National Academy of Sciences* 105(7): 2291–94.

Miller, John, Smitha Milli, and Moritz Hardt. 2020. "Strategic Classification is Causal Modeling in Disguise." In *International Conference on Machine Learning*, 69176926. PMLR.

Nowak, Martin A., and Karl Sigmund. 2005. "Evolution of Indirect Reciprocity." *Nature* 437(7063): 1291–98.

Ohtsuki, Hisashi, and Yoh Iwasa. 2004. "How Should We Define Goodness?—Reputation Dynamics in Indirect Reciprocity." *Journal of Theoretical Biology* 231(1): 107–20.

Oliveira, Raquel, 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: 106547.

Paiva, Ana, 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. New York: Association for Computing Machinery.

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

Pennisi, Elizabeth. 2005. "How Did Cooperative Behavior Evolve?" *Science* 309(5731): 93–93.

Piao, Jinghua, Jiazhen Liu, Fang Zhang, Jun Su, and Yong Li. 2023. "Human–AI Adaptive Dynamics Drives the Emergence of Information Cocoons." *Nature Machine Intelligence* 5: 1–11. https://doi.org/10.1038/s42256-023-00731-4.

Pinyol, Isaac, and Jordi Sabater-Mir. 2013. "Computational Trust and Reputation Models for Open Multi-Agent Systems: A Review." *Artificial Intelligence Review* 40(1): 1–25.

Rand, David G., Samuel Arbesman, and Nicholas A. Christakis. 2011. "Dynamic Social Networks Promote Cooperation in Experiments with Humans." *Proceedings of the National Academy of Sciences* 108(48): 19193–98.

Resnick, Paul, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. "Reputation Systems." *Communications of the ACM* 43(12): 45–48.

Ricci, Francesco, Lior Rokach, and Bracha Shapira. 2021. "Recommender Systems: Techniques, Applications, and Challenges." In *Recommender Systems Handbook*, 1–35. New York: Springer.

Romano, Angelo, Daniel Balliet, and Junhui Wu. 2017. "Unbounded Indirect Reciprocity: Is Reputation-Based Cooperation Bounded by Group Membership?" *Journal of Experimental Social Psychology* 71: 59–67.

Santos, Fernando P. 2023. "How to Break Information Cocoons." *Nature Machine Intelligence* 5: 1–2.

Santos, Fernando P., and Daan Bloembergen. 2019. "Fairness in Multiplayer Ultimatum Games through Moderate Responder

Selection." In *Artificial Life Conference Proceedings*, 187–94. Cambridge, MA: MIT Press. One Rogers Street 02142-1209, USA

Santos, Fernando P, Sara Encarnação, Francisco C. Santos, Juval Portugali, and Jorge M. Pacheco. 2016. "An Evolutionary Game Theoretic Approach to Multi-Sector Coordination and Self-Organization." *Entropy* 18(4): 152.

Santos, Fernando P., Yphtach Lelkes, and Simon A. Levin. 2021. "Link Recommendation Algorithms and Dynamics of Polarization in Online Social Networks." *Proceedings of the National Academy of Sciences* 118(50): e2102141118.

Santos, Fernando P., Simon A. Levin, and Vítor V. Vasconcelos. 2021. "Biased Perceptions Explain Collective Action Deadlocks and Suggest New Mechanisms to Prompt Cooperation." *iScience* 24(4): 102375.

Santos, Fernando P., Samuel F. Mascarenhas, Francisco C. Santos, Filipa Correia, Samuel Gomes, and Ana Paiva. 2019. "Outcome-Based Partner Selection in Collective Risk Dilemmas." In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19). International Foundation for Autonomous Agents and Multiagent Systems, 1556–64.

Santos, Fernando P., 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): 1–29.

Santos, Fernando P., 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, 6146–53.

Santos, Fernando P., Jorge M. Pacheco, and Francisco C. Santos. 2018. "Social Norms of Cooperation with Costly Reputation Building." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

Santos, Fernando P., Jorge M. Pacheco, and Francisco C. Santos. 2021. "The Complexity of Human Cooperation under Indirect Reciprocity." *Philosophical Transactions of the Royal Society B* 376(1838): 20200291.

Santos, Fernando P., Francisco C. Santos, and Jorge M. Pacheco. 2018. "Social Norm Complexity and Past Reputations in the Evolution of Cooperation." *Nature* 555(7695): 242–45.

Santos, Francisco C., and Jorge M. Pacheco. 2011. "Risk of Collective Failure Provides an Escape from the Tragedy of the Commons." *Proceedings of the National Academy of Sciences* 108(26): 10421–25.

Santos, Francisco C., Jorge M. Pacheco, and Tom Lenaerts. 2006. "Cooperation Prevails When Individuals Adjust Their Social Ties." *PLoS Computational Biology* 2(10): e140.

Sasaki, Tatsuya, Isamu Okada, and Yutaka Nakai. 2016. "Indirect Reciprocity Can Overcome Free-Rider Problems on Costly Moral Assessment." *Biology Letters* 12(7): 20160341.

Shirado, Hirokazu, and Nicholas A. Christakis. 2020. "Network Engineering Using Autonomous Agents Increases Cooperation in Human Groups." *iScience* 23(9): 101438.

Shoham, Yoav, Rob Powers, and Trond Grenager. 2007. "If Multi-Agent Learning Is the Answer, What Is the Question?" *Artificial Intelligence* 171(7): 365–77.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Smit, Jacobus, and Fernando P. Santos. 2023. "Learning Fair Cooperation in Systems of Indirect Reciprocity." In Adaptive Learning Agents Workshop 2023 - AAMAS.

Stein, Sebastian, and Vahid Yazdanpanah. 2023. "Citizen-Centric Multiagent Systems." In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1802–7. https://dl.acm.org/doi/abs/10.5555/3545946.3598843

Su, Jessica, Aneesh Sharma, and Sharad Goel. 2016. "The Effect of Recommendations on Network Structure." In *Proceedings of the 25th International Conference on World Wide Web*, 1157–67.

Traulsen, Arne, Simon A. Levin, and Chadi M. Saad-Roy. 2023. "Individual Costs and Societal Benefits of Interventions during the COVID-19 Pandemic." *Proceedings of the National Academy of Sciences* 120(24): e2303546120.

Whitaker, Roger M., Gualtiero B. Colombo, and David G. Rand. 2018. "Indirect Reciprocity and the Evolution of Prejudicial Groups." *Scientific Reports* 8(1): 13247.

Wooldridge, Michael. 2009. *An Introduction to Multiagent Systems*. Chichester: John Wiley & Sons.

How to cite this article: Santos, Fernando P. 2024. "Prosocial dynamics in multiagent systems." *AI Magazine* 45: 131–38. https://doi.org/10.1002/aaai.12143

AUTHOR BIOGRAPHY

Fernando P. Santos is an Assistant Professor at the Informatics Institute of the University of Amsterdam. His research lies at the interface of AI and complex systems. He is interested in understanding cooperation and collective dynamics in multiagent systems, and in designing fair/prosocial AI. Fernando completed his PhD in Computer Science and Engineering at Instituto Superior Técnico (University of Lisbon) and was a James S. McDonnell postdoctoral fellow at Princeton University.