Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

Review

Mathematical frameworks for the analysis of norms

Alessandro Sontuoso[1,2]

Research into society's informal rules of conduct, or norms, has recently experienced a surge, extending across multiple academic disciplines. Despite this growth, the theoretical modeling of norms often remains siloed within specific paradigms, as different disciplines tend to favor certain frameworks over others, thereby hindering the spread of innovative ideas. This article breaks through disciplinary barriers to explore recent advancements in the mathematical study of norms. It specifically focuses on cutting-edge theoretical research, structuring the discussion around four general frameworks: game theory, evolutionary game theory, agent-based modeling, and multi-agent reinforcement learning.

Addresses
[1] Department of Economics, City University of London, Northampton Sq., London EC1V 0HB, UK
[2] Smith Institute for Political Economy and Philosophy, Chapman University, One University Dr., Orange, CA 92866, USA

Corresponding author: Sontuoso, Alessandro (alessandro.sontuoso@city.ac.uk)

Keywords
Informal institutions, Social norms, Injunctive norms, Personal norms, Values, Game theory, Evolutionary game theory, Agent-based modeling, Multi-agent reinforcement learning, Machine learning.

Introduction

Society's informal rules of conduct, or norms, significantly influence individual decision-making. As we gain deeper insights into their interactive dynamics, the study of norms has emerged as a key field within and beyond the social sciences. Despite this increasing academic interest, the theoretical modeling of norms has often remained siloed within specific paradigms, as different disciplines tend to favor certain frameworks over others, thereby hindering the spread of innovative ideas. Breaking disciplinary barriers, this article explores recent advancements in the mathematical study of norms. It specifically focuses on the latest theoretical research, organizing the exposition into four wide-ranging approaches: game theory, evolutionary game theory, agent-based modeling, and multi-agent reinforcement learning.

While formalisms vary, social norms are commonly understood (per Bicchieri's account [1]) as group-specific behavioral regularities that arise from particular preferences for conformity: these preferences are contingent on group members' beliefs about what others *will* do, and what others think *ought to* be done ([2–4]).[1] With this in mind, each of the above four modeling approaches further draws on various intuitions about individuals' thought processes and motivations; as a result, each model incorporates distinct assumptions — often implicitly — into its formal specification. So, even though each framework aims to grasp the essence of social norms, their diverse assumptions often lead to distinct conclusions and, in fact, address different aspects of the complex world of norms. This article discusses some of the newest theoretical contributions, highlighting their broad assumptions, scope, benefits, and possible limitations.

Game theory: *modeling the conditions under which experienced agents follow stable norms*

Game-theoretic models of social norms focus on equilibrium states; in particular, here the emphasis is on the conditions under which a social norm implies a stable solution to a mixed-motive (social dilemma) game. Formally, a social norm is a behavioral pattern such that: (i) players choose best-responses to one another's strategies, given preferences for conformity to that pattern; and (ii) players hold correct 'empirical' beliefs about the others' actions, as well as correct 'normative' expectations as to the actions that others deem appropriate ([5]).

Thus, a social norm is a stable pattern where people in a given society expect and follow a certain way of behaving, based on their understanding of what others

[1] In practice, people generally exhibit varying degrees of conformity to those two types of expectations (i.e., what *will* and what *ought to* be done), or even unconditional conformity toward an ideal, leading to distinctions between social norms, descriptive norms, and moral norms. In what follows, 'norms' will be used in an inclusive sense, with the specific type indicated as appropriate. For a detailed analytical discussion of the differences between classes of norms, see Bicchieri [1].

Current Opinion in Psychology 2024, **60**:101930

This review comes from a themed issue on Norm Change (2024)

Edited by Cristina Bicchieri, Michele Gelfand and Giulia Andrighetto

For complete overview about the section, refer Generation COVID: Coming of Age Amid the Pandemic (2024)

Available online 16 October 2024

https://doi.org/10.1016/j.copsyc.2024.101930

will do and what others believe is appropriate. As such, game-theoretic models of social norms typically assume a *belief-dependent* utility function, meaning that individuals' utilities are affected by their beliefs about how their actions align with the group's standards. As per the scope of this journal, the following discussion will primarily focus on research published in the past two years; readers should consult [6] for a comprehensive survey of belief-dependent motivations, collectively referred to as psychological game theory ('PGT'). In this regard, it is important to note that PGT models vary considerably, and not all are designed to capture norms in the sense described above. For instance, some early PGT applications — known as 'social preferences' — assume that individuals care about conforming to a unique (exogenously given) standard of behavior (e.g., egalitarianism), irrespective of the population or context. However, social norms are properly understood as contingent on expectations that vary with the population and context ([1,5]).

Accordingly, a recent series of models account for heterogeneity in what is deemed appropriate, so that an individual anticipates that standard 'codes of conduct' may differ across people or situations ([7,8]). Among the latest approaches, [9] proposes a model to predict decisions in dictator games by explicitly integrating empirical beliefs, normative expectations, and personal values (otherwise known as personal normative beliefs): an advantage of this model is the ability to show how the level of consensus in personal values affects dictator behavior. A related research stream focuses on *image concerns* in the face of heterogeneity in what is deemed appropriate (without explicitly modeling empirical beliefs). In particular, [10] presents a model of social norms that regulate the expression of opinions: in this case, an individual faces a binary decision about which opinion to express, balancing a desire for authenticity with the need to be perceived as having the right values by the relevant audience; social norms regarding opinion expression emerge as equilibria of a signaling game. In a similar vein, [11] studies a simple setting in which one has to make a morally contentious choice: here, preferences are determined by consumption utility, personal values (norms), and social image, where the latter is derived from either 'respect' or 'approval'; the model shows that different forms of image concerns may lead to opposite choices when personal norms vary across participants (see also [12]). Lastly, [13] introduces a model where one's preferences across time periods are dynamically shaped by one's chosen actions and previously adopted 'worldviews' (norms): per this model, the agent's optimal behavior corresponds to the Markov-perfect equilibrium of an intertemporal game in which each 'present self' chooses its 'future self'.

To sum up, the focus of game-theoretic frameworks of social norms is on belief-dependent preferences that, under the relevant conditions, allow for a stable pattern where individuals within a society anticipate and adhere to some (endogenous) standards of behavior. Specifically, these models account for variously defined preferences for conformity to local codes of conduct, which directly impact equilibrium states (under the usual assumptions that individuals choose mutual best-responses and hold correct beliefs). Looking ahead, as highlighted by some of these models, the malleability of people's normative beliefs is an area that merits further investigation ([14,15]). Another promising research avenue involves formalizing the flow (and absorption) of noisy information: so far, much of this work has focused on *peer effects* arising from the acquisition of descriptive information ([16–18]), but future studies should consider explicitly integrating normative information as well.

Evolutionary game theory: *modeling the dynamic processes that lead to norm emergence*

Evolutionary game theory ('EGT') frameworks provide insights into how behaviors may evolve over time, thereby complementing classical game theory. Indeed, even though a Nash equilibrium can be interpreted as a potential stable point resulting from a dynamic adaptive process (e.g., the *culmination* of a trial-and-error learning process), classical game theory does not generally model such *dynamics*. EGT, by contrast, directly addresses this gap by formalizing the dynamics of adaptation, that is, the processes by which the distribution of strategies within a large population may change in response to the environment.[2] EGT is thus especially well-suited for modeling the emergence of informal institutions — such as social norms — where social order gradually arises from repeated interactions among boundedly rational individuals ([19]). The discussion below focuses on the latest theoretical research; for surveys of earlier results, readers are referred to Refs. [20–22].

A prominent stream of research examines preferences influenced by similarity effects (i.e., utility increases when one's choice aligns with the population's average behavior), as well as other motivations independent of similarity, whether material or psychological. Given this broad class of preferences, [23] contrasts 'discrete norms' with 'continuous norms', where the set of possible actions is either discrete or continuous (think of a dress code versus tipping etiquette). With discrete

[2] EGT frameworks typically assume a large population of agents who are repeatedly and randomly matched in pairs (or groups) to play a game. The composition of the population evolves according to the replicator equation or other evolutionary dynamics, which determine how the proportion of agents employing each strategy changes over time based on their relative success. For instance, in the replicator dynamic, strategies that yield payoffs higher than the population average become more prevalent, while less successful strategies decline. These dynamics illustrate how advantageous behaviors can spread throughout the population, thereby modeling the evolution of strategies over time.

norms, the pressure for similarity leads to multiple stable equilibria, indicating that locally common behaviors can successfully persist over time. With continuous norms, instead, choices converge on a unique equilibrium: in this case, similarity-independent factors determine equilibrium outcomes, irrespective of the model's initial conditions. (For further results on similarity, see Refs. [24,25].). Relatedly, see Ref. [26] for insights into how cross-generational environmental stability tends to promote a belief in the value of tradition, leading to 'cultural persistence'; for more on cultural transmission, see Ref. [27].

Turning to a different line of research, [28] defines social norms as correlated equilibria: yet, unlike correlated equilibria in classical game theory (where all the players know the joint distribution of 'recommendations' that may be generated by a predefined coordination device), this model posits that social norms can arise as correlated equilibria through evolutionary dynamics. A 'subjective norm' prescribes a strategy to a player conditional on a privately observed event, and provides a probability distribution of prescriptions that the opponent may be following. When a subjective norm is both individually rational and evolutionarily stable, correlated beliefs can emerge, allowing players to coordinate on a social norm.

[29] provides a model of 'indirect reciprocity', whereby an individual considers their peers' experiences in deciding how to interact with someone ([30]). Specifically, players adopt strategies that may depend on the reputations of others, with reputations being determined by a 'social norm' that dictates how observed behaviors are judged. Unlike earlier models, here the population is partitioned into 'gossip groups': each group adheres to a distinct social norm for judging reputations; strategies and social norms that generate higher payoffs are more likely to gain popularity via (biased) imitation. Under certain conditions, the model predicts that the population will converge on a social norm that penalizes those who cooperate with individuals of poor social standing. (For more on norms and punishment, see Refs. [31–33].).

In other research, [34,35] develop a set of models that integrate multiple components into an individual's utility function, often considered separately in prior literature. Besides a material payoff, the models account for various forms of social pressure (such as the urge to conform to an external authority, follow peer behavior, or meet peer expectations), along with a cognitive dissonance component. After choosing an action and observing peer behavior, each individual revises their personal norms and beliefs, with the revision following DeGroot-like dynamics. These models show that equilibrium outcomes are driven by the interplay between individuals' desire to maximize material payoffs and the actions endorsed by the external authority.

Summing up, evolutionary game theory (EGT) models of social norms focus on the dynamic processes that gradually lead to the adoption of certain behaviors; hence, EGT is particularly apt for investigating how norms emerge, spread, and stabilize over time via repeated interactions, learning, and adaptation. Also, due to its focus on large populations, EGT can intuitively explain frequency effects and path dependence in norm adoption, illustrating how certain outcomes are driven by the model's initial conditions. Yet, it is worth noting that — in some cases — the generalizability of the results is limited in that small variations in the models' parameters can cause substantially different outcomes, yielding near unfalsifiable predictions.

Agent-based modeling: *exploring complex patterns emerging from heterogeneous agents*

Agent-based modeling ('ABM') approaches involve 'computational objects' (e.g., individuals, groups, or any other purposeful entities) that interact in *space* and *time* according to predefined inductive rules. While related to (evolutionary) game-theoretic models, ABM differs in allowing for much greater heterogeneity within both the population and the environment: in particular, agents may vary in their network locations and cognitive abilities, and are designed to variously respond to available information and incentives. In short, while classical and evolutionary game-theoretic models of norms respectively emphasize which equilibria are possible and via what dynamic processes, ABM frameworks are not restricted by considerations of analytic tractability or provability. Thus, ABM allows researchers to explore complex emergent patterns that may or may not lead to equilibrium. (For extensive surveys of early results, see Refs. [36–38].).

Among recent contributions, [39] proposes a model for prisoner's dilemma games in which agents interact on a static network, motivated by material payoffs, personal values, and normative expectations; the model also incorporates parameters for bounded rationality, learning, memory decay, social pressure, and long-term adaptation dynamics. Results suggest that cooperative systems thrive at the cusp of instability, teetering between equilibrium and chaos (i.e., near the critical point of a phase transition): this appears to enable agents to maintain high levels of group coordination while remaining responsive to external perturbations. For additional results on the cycles of cooperation and defection, see Ref. [40].

[41] examines networked agents whose beliefs are represented as probability distributions over various

levels of agreement with some issue. The interaction between two agents is characterized by a positive or negative link weight: a negative weight indicates a tendency for two individuals to disagree on issues, so when one individual updates a belief, the neighbor will revise it in the opposite direction. Further, when a particular belief does not align with one's belief system or social group, one may experience cognitive dissonance; this can be mitigated by modifying one's beliefs or by severing connections to agents that contradict one's firmly held views. This framework can be used to assess belief dynamics in experiments.

In related work, [42] proposes a model to analyze societies with varying levels of susceptibility to social pressure and varying tendencies to form new acquaintances ('extraversion'), capturing how individuals influence and are influenced by others within their social networks. Simulations suggest that the majority opinion consolidates faster in societies exhibiting greater influenceability; also, unpopular beliefs are more likely to spread in societies featuring greater influenceability and lower extraversion (i.e., sparser networks). In other research, [43] investigates the effects of various forms of polarization in a multiplayer coordination game with Pareto-rankable equilibria: although this model does not explicitly define norms, its results show that the segregation of the network into clusters can hinder information flow and thus efficient coordination, suggesting that segregation may impede the spread of virtuous norms.

Overall, agent-based modeling approaches enable the exploration of complex systems that may otherwise be intractable or may lie at the threshold between chaos and equilibrium. While some such computational frameworks produce messy outputs that lead to few generalizable insights, many others uncover novel findings, which can inform the derivation of analytic results (e.g., via game-theoretic methods).

Multi-agent reinforcement learning: *optimizing via trial and error*

Multi-agent reinforcement learning (MARL) is a thriving field at the intersection of game theory and computer science that studies the behavior of agents interacting repeatedly in a shared environment. MARL research involves some form of Markov decision processes: at each time step, the environment is in a particular state, and agents select an action available in that state; the environment then stochastically transitions to a new state (with the probability of the new state depending on the chosen actions), agents receive an individual reward, and so on. Through trial and error, agents aim to find the optimal 'policy' that maximizes individual rewards over time. Unlike classical game theory, which focuses on identifying ideal policies to be implemented by experienced agents, MARL approaches to social norms analyze how agents gradually learn the environment and discover optimal policies via a *trial-and-error process*.[3] Thus, MARL can be used to investigate how prosocial norms progressively emerge in a shared environment. (For expansive surveys of early research, see Refs. [44–47].).

In recent work, [48] examines partially observable general-sum Markov games: in each game state, agents navigate a two-dimensional grid where their actions include collecting berries and possibly punishing other agents. As they partially observe the environment, agents need to understand the relationship between their actions, observations, and rewards; in particular, they need to identify the one type of poisonous berries (out of several types of harmless berries), which are associated with delayed negative rewards. To that end, each agent is endowed with a 'deep neural network', a machine learning algorithm that maps observations to actions predicted to maximize rewards. The model shows that when agents are rewarded for punishing others who collect poisonous berries, long-term rewards improve. Remarkably, when agents are rewarded for punishing even the collection of one specific type of harmless berries (besides the poisonous berries), long-term rewards improve further. This suggests that 'silly rules', which seemingly lack benefits, provide agents valuable practice in enforcing beneficial norms (against poisonous berries), ultimately yielding better long-term rewards.

Turning to related issues, [49] analyzes how public sanctioning helps people learn social norms through decentralized multi-agent reinforcement, whereby each agent forms a prediction about whether society might approve or disapprove of the observed behavior. In a somewhat similar vein, [50] studies reputation mechanisms via reinforcement learning (see also [51]). Lastly, [52] introduces a novel algorithm that adaptively learns moral values via limited human evaluative feedback.

In short, multi-agent reinforcement learning offers an extra toolbox for investigating social norms, as it enables the analysis of agents that gradually learn how to interact optimally within a shared environment. Looking ahead, open questions remain on how to develop algorithms capable of generalizing normative considerations effectively across populations or situations, thereby expanding the models' predictive scope in diverse settings. Incidentally, addressing these questions could lead to breakthroughs in creating AI applications that integrate better into human-centric environments.

[3] Note that agents in EGT are typically not consciously optimizing; rather, strategies that lead to higher payoffs ('fitness') become more common through a natural selection-like process, determined by some evolutionary dynamics. By contrast, agents in MARL explicitly optimize their policies based on rewards received from their actions.

Conclusion

This article has presented a systematic analysis of alternative mathematical approaches to studying norms, discussing some of the latest theoretical models and emphasizing their assumptions, scope, benefits, and possible limitations.

Even though game theory, evolutionary game theory, agent-based modeling, and multi-agent reinforcement learning often share common ground and intersect in practice, I have classified the literature into these four groups based on certain attributes for clarity of exposition. In particular, in the game theory section, I explored models with rational agents aiming to maximize their utility, and I showed how specific belief-dependent preferences can explain the perpetuation of social norms. In the evolutionary game theory (EGT) section, I focused on models that define evolutionary dynamics and generally derive equilibria analytically (in contrast to agent-based models relying on simulations). The agent-based modeling (ABM) section concentrated on papers that explicitly incorporate spatial structures, and allow for greater heterogeneity within both the population and the environment; these papers utilize computational simulations to explore how local interactions affect the emergence of norms, without necessarily relying on closed-form solutions. In the multi-agent reinforcement learning (MARL) section, I examined models that involve some form of Markov decision processes and employ theoretical notions of equilibrium, analyzing interactions between learning agents in dynamic environments.

While each framework seeks to capture the essence of social norms, they rely to some extent on distinct assumptions in their formal models, each highlighting different aspects of the complex landscape of norms. In this regard, it is worth noting that different academic disciplines often gravitate toward certain frameworks over others, which can lead to an overemphasis on specific aspects and siloed knowledge. To conclude, future research should aim for greater integration across disciplines to develop a more comprehensive understanding of how norms emerge, spread, change, and stabilize within societies.

Author contribution

Alessandro Sontuoso is solely responsible for conceptualization, writing, and revision of the manuscript.

Declaration of competing interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

References of particular interest have been highlighted as:

* of special interest
** of outstanding interest

1. Bicchieri Cristina: *The grammar of society: the nature and dynamics of social norms*. Cambridge University Press; 2006.
2. Bicchieri Cristina, Muldoon Ryan, Sontuoso Alessandro: *Social norms*. The Stanford encyclopedia of philosophy; 2023.
3. Gelfand Michele J, Gavrilets Sergey, Nunn Nathan: **Norm dynamics: interdisciplinary perspectives on social norm emergence, persistence, and change**. *Annu Rev Psychol* 2024, **75**: 341–378.
4. Andrighetto Giulia, Gavrilets Sergey, Gelfand Michele, Mace Ruth, Vriens Eva: **Social norm change: drivers and consequences**. *Philosophical Transactions of the Royal Society B* 2024, **379**, 20230023.
5. Bicchieri Cristina, Sontuoso Alessandro: **Game-theoretic accounts of social norms: the role of normative expectations**. In *Handbook of experimental game theory*. Edward Elgar Publishing; 2020:241–255.
6. Battigalli Pierpaolo, Dufwenberg Martin: **Belief-dependent motivations and psychological game theory**. *J Econ Lit* 2022, **60**: 833–882.
7. Bicchieri Cristina, Sontuoso Alessandro: *I cannot cheat on you after we talk*. The Prisoner's Dilemma; 2015:101–114.
8. Kimbrough Erik O, Vostroknutov Alexander: *A theory of injunctive norms*. 2023. Available at: SSRN 3566589.
9. ** D'Adda, Giovanna, Dufwenberg Martin, Passarelli Francesco, Tabellini Guido: **Social norms with private values: theory and experiments**. *Game Econ Behav* 2020, **124**:288–304.
10. * Golman Russell: **Acceptable discourse: social norms of beliefs and opinions**. *Eur Econ Rev* 2023, **160**, 104588.
11. te Velde VL: **Heterogeneous norms: social image and social pressure when people disagree**. *J Econ Behav Organ* 2022, **194**:319–340.
12. Bursztyn Leonardo, Egorov Georgy, Haaland Ingar, Rao Aakaash, Roth Christopher: **Justifying dissent**. *Q J Econ* 2023, **138**:1403–1451.
13. Bernheim B Douglas, Braghieri Luca, Martínez-Marquina Alejandro, Zuckerman David: **A theory of chosen preferences**. *Am Econ Rev* 2021, **111**:720–754.
14. Bicchieri Cristina, Dimant Eugen, Sonderegger Silvia: **It's not a lie if you believe the norm does not apply: conditional norm-following and belief distortion**. *Game Econ Behav* 2023, **138**: 321–354.
15. Bénabou Roland, Tirole Jean: **Mindful economics: the production, consumption, and value of beliefs**. *J Econ Perspect* 2016, **30**:141–164.
16. Boucher Vincent, Rendall Michelle, Ushchev Philip, Zenou Yves: **Toward a general theory of peer effects**. *Econometrica* 2024, **92**:543–565.
17. Ushchev Philip, Zenou Yves: **Social norms in networks**. *J Econ Theor* 2020, **185**, 104969.
18. Charness Gary, Naef Michael, Sontuoso Alessandro: **Opportunistic conformism**. *J Econ Theor* 2019, **180**:100–134.
19. Skyrms Brian: *Evolution of the social contract*. 2nd ed. Cambridge University Press; 2014.

20. Pan Xinyue, Gelfand Michele, Nau Dana: **Integrating evolutionary game theory and cross-cultural psychology to understand cultural dynamics**. *Am Psychol* 2021, **76**:1054.

21. Young H Peyton: **The evolution of social norms**. *Annual Review of Economics* 2015, **7**:359–387.

22. Alexander J McKenzie: *Evolutionary game theory*. The Stanford Encyclopedia of Philosophy; 2021.

23. * Yan Minhua, Mathew Sarah, Boyd Robert: **"Doing what others do" does not stabilize continuous norms**. *PNAS nexus* 2023, **2**:1–11.

24. Arefin Md Rajib, Tanimoto Jun: **Coupling injunctive social norms with evolutionary games**. *Appl Math Comput* 2024, **466**, 128463.

25. Efferson Charles, Ehret Sönke, von Flüe Lukas, Vogt Sonja: **When norm change hurts**. *Philosophical Transactions of the Royal Society B* 2024, **379**, 20230039.

26. Giuliano Paola, Nunn Nathan: **Understanding cultural persistence and change**. *Rev Econ Stud* 2021, **88**:1541–1581.

27. Della Lena Sebastiano, Panebianco Fabrizio: **Cultural transmission with incomplete information**. *J Econ Theor* 2021, **198**, 105373.

28. ** Morsky Bryce, Akçay Erol: **Evolution of social norms and correlated equilibria**. *Proc Natl Acad Sci USA* 2019, **116**: 8834–8839.

29. Kessinger Taylor A, Tarnita Corina E, Plotkin Joshua B: **Evolution of norms for judging social behavior**. *Proc Natl Acad Sci USA* 2023, **120**, e2219480120.

30. Schmid Laura, Chatterjee Krishnendu, Hilbe Christian, Nowak Martin A: **A unified framework of direct and indirect reciprocity**. *Nat Human Behav* 2021, **5**:1292–1302.

31. Morsky Bryce, Plotkin Joshua B, Akçay Erol: **Indirect reciprocity with Bayesian reasoning and biases**. *PLoS Comput Biol* 2024, **20**, e1011979.

32. Pandula Neel, Akçay Erol, Morsky Bryce: **Indirect reciprocity with abductive reasoning**. *J Theor Biol* 2024, **580**, 111715.

33. Li Xueheng, Molleman Lucas, van Dolder Dennie: **Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation**. *Evol Hum Behav* 2021, **42**:469–479.

34. Gavrilets Sergey, Richerson Peter J: **Authority matters: propaganda and the coevolution of behaviour and attitudes**. *Evolutionary Human Sciences* 2022, **4**, e51.

35. Gavrilets Sergey: **Coevolution of actions, personal norms and beliefs about others in social dilemmas**. *Evolutionary Human Sciences* 2021, **3**, e44.

36. Andrighetto Giulia, Vriens Eva: **A research agenda for the study of social norm change**. *Philosophical Transactions of the Royal Society A* 2022, **380**, 20200411.

37. Axtell Robert L, Doyne Farmer J: **Agent-based modeling in economics and finance: past, present, and future**. *J Econ Lit* 2022:1–101.

38. Zhang Haifeng, Vorobeychik Yevgeniy: **Empirically grounded agent-based models of innovation diffusion: a critical review**. *Artif Intell Rev* 2019, **52**:707–741.

39. * Realpe-Gómez John, Andrighetto Giulia, Nardin Luis Gustavo, Montoya Javier Antonio: **Balancing selfishness and norm conformity can explain human behavior in large-scale prisoner's dilemma games and can poise human groups near criticality**. *Phys Rev* 2018, **97**, 042321.

40. Roy Sourav, Chowdhury Sayantan Nag, Kundu Srilena, Sar Gourab Kumar, Banerjee Jeet, Rakshit Biswambhar, Mali Prakash Chandra, Perc Matjaž, Ghosh Dibakar: **Time delays shape the eco-evolutionary dynamics of cooperation**. *Sci Rep* 2023, **13**, 14331.

41. Galesic Mirta, Olsson Henrik, Dalege Jonas, Does Tamara Van Der, Stein Daniel L: **Integrating social and cognitive aspects of belief dynamics: towards a unifying framework**. *J R Soc Interface* 2021, **18**, 20200857.

42. Muthukrishna Michael, Schaller Mark: **Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change**. *Pers Soc Psychol Rev* 2020, **24**:103–120.

43. Vasconcelos Vítor V, Constantino Sara M, Dannenberg Astrid, Lumkowsky Marcel, Weber Elke, Levin Simon: **Segregation and clustering of preferences erode socially beneficial coordination**. *Proc Natl Acad Sci USA* 2021, **118**, e2102153118.

44. Gronauer Sven, Diepold Klaus: **Multi-agent deep reinforcement learning: a survey**. *Artif Intell Rev* 2022, **55**:895–943.

45. Morris-Martin Andreasa, De Vos Marina, Padget Julian: **Norm emergence in multiagent systems: a viewpoint paper**. *Aut Agents Multi-Agent Syst* 2019, **33**:706–749.

46. Du Yali, Leibo Joel Z, Islam Usman, Willis Richard, Sunehag Peter: **A review of cooperation in multi-agent learning**. *arXiv preprint arXiv:2312.05162* 2023.

47. Yang Yaodong, Wang Jun: **An overview of multi-agent reinforcement learning from game theoretical perspective**. *arXiv preprint arXiv:2011* 2020, 00583.

48. ** Köster Raphael, Hadfield-Menell Dylan, Everett Richard, Weidinger Laura, Hadfield Gillian K, Leibo Joel Z: **"Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents." Proceedings of the national academy of sciences** 2022, **119**(3), e2106028118.

49. Vinitsky Eugene, Köster Raphael, Agapiou John P, Duéñez-Guzmán Edgar A, Vezhnevets Alexander S, Leibo Joel Z: **A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings**. *Collectiv Intell* 2023, **2**, 26339137231162025.

50. Anastassacos Nicolas, García Julian, Hailes Stephen, Musolesi Mirco: **Cooperation and reputation dynamics with reinforcement learning**. *arXiv preprint arXiv:2102.07523* 2021.

51. McKee Kevin R, Hughes Edward, Zhu Tina O, Chadwick Martin J, Koster Raphael, Garcia Castaneda Antonio, Beattie Charlie, Graepel Thore, Botvinick Matt, Leibo Joel Z: **A multi-agent reinforcement learning model of reputation and cooperation in human groups**. *arXiv preprint arXiv:2103.04982* 2021.

52. Shi Zijing, Fang Meng, Chen Ling, Du Yali, Wang Jun: **Human-guided moral decision making in text-based games**. In *The 38th annual AAAI conference on artificial intelligence*; 2024.

Further information on references of particular interest

9. ** Proposes a game-theoretic model to predict behavior in dictator games by explicitly integrating empirical beliefs, normative expectations, and personal values.

10. * Presents a model of social norms regulating the expression of opinions: per this model, an individual faces a binary decision about which opinion to express, balancing the desire for authenticity with the need to be perceived as having the right values by the relevant audience.

23. * Contrasts 'discrete norms' with 'continuous norms' (where the set of possible actions is either discrete or continuous): with discrete norms, the pressure for similarity leads to multiple stable equilibria; with continuous norms, instead, choices converge on a unique equilibrium.

28. ** Builds on the literature that studies how social norms can arise as correlated equilibria through evolutionary dynamics.

39. * Develops an agent-based model for prisoner's dilemma games in which individuals interact on a static network, motivated by material payoffs, personal values, and normative expectations.

48. ** Introduces a multi-agent reinforcement learning model where individuals engage in a foraging task: the model suggests that 'silly rules', which seemingly lack benefits, provide agents valuable practice in enforcing beneficial norms, ultimately yielding better long-term rewards.