PNAS SPECIAL FEATURE: INTRODUCTION OPEN ACCESS

Collective artificial intelligence and evolutionary dynamics

Udari Madhushani Sehwag[a,1], Alex McAvoy[b,c], and Joshua B. Plotkin[d,e]

Collective behavior is ubiquitous and highly structured in the natural world, allowing individuals to coordinate and cooperate in pursuit of common aims. The field of evolutionary game theory helps explain how structured collective behavior emerges in humans and other animals. But results from evolutionary game theory are typically restricted to simple and stylized problems. To be sure, simple models have been incredibly useful for understanding natural systems, developing hypotheses, and designing experiments to test hypotheses. At the same time, the field of multiagent research in AI has recently seen explosive growth, allowing researchers to model collective behavior in very complex domains using agents trained with reinforcement learning. There is a broad qualitative similarity between the problems addressed in these two fields, so that bridging them may provide theoretical guarantees about algorithms in reinforcement learning while extending the reach of evolutionary game theory. Synergy between these fields should help us to understand and reliably engineer collective behavior in complex domains.

Evolutionary game theory has already stimulated algorithmic designs, formal connections with reinforcement learning, and theoretical and numerical analysis of cooperation in mean-field and networked populations. However, theoretical results are mostly restricted to simple environmental settings, such as repeated matrix games. In contrast, the multiagent reinforcement learning community has developed algorithmic designs and computational experiments in more complex environments, which include temporal and spatial variability (1). The complexity associated with these environments leaves the field lacking a theoretical grounding for its empirical observations or conjectures. And so now is an opportune time to explore *i)* whether theoretical guarantees and algorithmic designs derived for simple environments can be extended to more complex environments and *ii)* how theoretical guarantees with minimally restrictive assumptions can support experimental observations and conjectures.

This special feature draws together work from game theory, AI, and population dynamics to shed light on key problems at their interface–such as how cooperative behaviors can emerge among self-interested individuals and how agents can learn to coordinate with partners whom they have not encountered before. These are pressing problems in a world that is increasingly governed by the sharing of information among autonomous systems, and between humans and artificial agents. AI and other communities are growing rapidly, but their work remains largely segregated by academic disciplines. The goal of this special feature is to stimulate discussion and exchange of cutting-edge developments between these domains, taking a step toward bridging the gap between AI and other communities that are grappling with multiagent systems.

The special feature begins with a broad perspective from Barfuss et al. (2), which explores the concept of "collective cooperative intelligence." The authors emphasize the need to bridge complex systems science and multiagent reinforcement learning to better understand and foster large-scale cooperation. The authors observe that complex systems science (which includes evolutionary game theory) provides a theoretical foundation for emergent collective behavior, while multiagent reinforcement learning offers computational tools to model and optimize cooperation; and they propose an integrated approach to study how intelligent agents navigate complex environments to achieve sustainable cooperation. This article concludes with five open research strands related to "collective reinforcement learning dynamics," which is a potential bridge between complex systems science and multiagent reinforcement learning, focused on equations, rather than algorithms, to model learning. These strands include both basic theory and analyses from dynamical systems to integrate cognitive and microscopic processes into models of population-based learning.

Narrowing the scope to a particular model of population-based learning, Bielawski et al. (3) study the impact of heterogeneous learning rates in multiagent population congestion games. Congestion games describe scenarios in which individual actions can have negative impacts on the population (e.g., use of a roadway or communication network). The paper takes a dynamical systems approach to a classical model in evolutionary game theory (4), focusing on large (infinite) unstructured populations, with updates based on a technique (multiplicative weight update method) commonly used in machine learning and optimization. The authors demonstrate that while traditional Nash equilibria exist, learning dynamics can deviate toward chaotic and unstable behaviors, increasing social costs compared to game-theoretic predictions. Despite microscopic chaos, however, the time-averaged macroscopic behavior converges to the unique Nash equilibrium. These results highlight a

Author affiliations: [a]Department of Computer Science, Stanford University, Stanford, CA 94305; [b]School of Data Science and Society, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [c]Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [d]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104; and [e]Center for Mathematical Biology, University of Pennsylvania, Philadelphia, PA 19104

U.M.S. and J.B.P. are organizers of this Special Feature.

Author contributions: U.M.S., A.M., and J.B.P. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: udarim@stanford.edu.

Published June 16, 2025.

complicated relationship between individual learning dynamics and equilibrium concepts in game theory. An important point in this paper is that complex dynamics can arise from simple models, but understanding these models can still yield insights into more complex environments, where such direct analyses might be prohibitive.

Duéñez-Guzmán et al. (5) approach population dynamics with a slightly different lens, exploring the role of partner choice in learning agents. Rather than deterministic models in the limit of infinite populations (3, 4), these authors focus on small populations with complex spatial environments. Partner choice, in which individuals use cues or past behavior when choosing social interactions, is a long-standing topic in evolutionary theory, and it has been shown to promote cooperation in hunter-gatherer societies (6). Here, the authors consider an extended "boat race" environment in which agents race back and forth across a river to collect apples. Although this environment is both spatially- and temporally extended, it is based on a simple matrix game called the "stag hunt" (a social dilemma in collaborative hunting). The agents have perceptible features, including a color and a badge, and the authors explore how statistical discrimination emerges in the population. Using multiagent reinforcement learning experiments, the authors find that increasing the salience of outcome-relevant features helps agents rely less on spurious correlations, leading to fairer and more rewarding partner choices.

Köster et al. (7) investigate a related topic that has also been a longstanding focus of evolutionary theory: the emergence of in-group bias (8). In their modern reconsideration of this problem, the authors consider artificial agents trained using deep reinforcement learning. They show that even "tabula rasa" agents (i.e., those without inherent social biases) develop a preference for interacting with visually similar agents based on exposure and familiarity. These findings suggest that such biases arise as a byproduct of general cognitive processes without requiring intrinsic motivational mechanisms, but they can be mitigated through sufficient intergroup interaction. Interestingly, although this paper is ostensibly about multiagent reinforcement learning, it has connections to the theory of multilevel selection (9), where agents benefit from expressing altruistic behaviors within their group because it enhances the competitiveness of the group as a whole.

The work of Kleiman-Weiner et al. (10) zooms in on the cognitive processes at play in the expression of social behaviors like cooperation. Informed by both AI and human decision-making, the authors introduce the "Bayesian reciprocator," which integrates Bayesian theory of mind into evolutionary models of cooperation. Inspired by the observed effects of reciprocity, reputation, relationships, robustness to noise, forgiveness, and theory of mind, the Bayesian agent is able infer others' beliefs and strategies, allowing agents to cooperate conditionally based on inferred reciprocity, as opposed to perfect knowledge commonly assumed in traditional evolutionary game theory. Importantly, the Bayesian reciprocator is able to incorporate social preferences into its decision-making process. Social preferences have long been part of the field of behavioral economics (11, 12) and have recently been considered in multiagent reinforcement learning algorithms (13). Evolutionary simulations in dynamic game environments indicate that the Bayesian reciprocator is more robust and adaptable compared to standard automata strategies, significantly expanding the conditions under which cooperation can evolve.

The research article of McAvoy et al. (14) explores a complementary question: When can an agent elicit cooperation from an opponent who is purely self-interested? This work builds on a key finding about how to transform a repeated social dilemma into a take-it-or-leave-it ultimatum game via clever strategic play. The authors extend the theory of so-called "zero-determinant" strategies, originally developed for repeated games by Press and Dyson (15), to stochastic (Markov) games (16), which can be used to describe conflicts of interest in mixed-motivation reinforcement learning scenarios. One shortcoming of the classical theory is that it requires an explicit solution to identify such zero-determinant strategies; whereas McAvoy et al. (14) get around this problem by using reinforcement learning to find statistical solutions. Zero-determinant strategies allow an agent to unilaterally align incentives in two-agent games, promoting fairness and cooperation by effectively converting mixed-motivation interactions into cooperative interactions, without requiring coordination or central control. This study may be thought of as a statistical perspective on zero-determinant strategies in multistate settings relevant in multiagent reinforcement learning, which provides opportunities for evolutionary game theory to inform reinforcement learning and for reinforcement learning to inform evolutionary game theory.

The work of Terrucha et al. (17) explores how humans program artificial delegates to make strategic decisions in collective-risk dilemmas, where individuals must contribute to a public good to avoid collective loss. This paper is based on human experiments in which the participants choose delegates to carry out their actions as a kind of commitment device. The authors find that although delegation increases contributions to the public good, precision errors in the programmed agents (delegates) prevent optimal success, suggesting that algorithmic fine-tuning is necessary for effective cooperation in digital decision-making contexts. Intriguingly, the experiments suggest that agents who do not delegate end up reducing contributions early on, possibly in order to punish others who contribute little. These results imply that delegation, coupled with learning processes for fine-tuning, may help to sustain social goals and collective action.

Moving more squarely into the realm of evolutionary game theory, the perspective article from Garcia and Traulsen (18) takes a look at the impact of strategy choice in models of cooperation. The authors argue that the choice of strategy sets can significantly influence outcomes and may lead to misleading conclusions about the emergence of cooperation. They propose three guiding principles for systematically choosing strategies: *i)* ensure unbiased inclusion of all computationally equivalent strategies; *ii)* ground choices in explicit microeconomic models; and *iii)* connect assumptions to stylized empirical facts. The authors also suggest that methods from AI can enhance the robustness

and complexity of these models, which echoes sentiments found throughout the special feature.

In the final paper of the special feature, Tacchetti et al. (19) explore how to use AI in mechanism design. Whereas game theory typically posits an incentive structure and studies what strategic outcomes will arise, mechanism design flips this script, starting with a goal and finding an incentive structure (a "mechanism") to achieve that goal (20–22). The authors discuss how deep reinforcement learning can be used to optimize social and economic policies, such as auctions, taxation, and resource redistribution, for human benefit. This work highlights how AI can be used for mechanism design to create more efficient and equitable systems, as well as the challenges related to modeling human preferences, ethical considerations, and the interpretability of policies generated by AI.

Three major themes emerge from this special feature. The first is population dynamics and reinforcement learning, where reinforcement learning agents are seen as existing within a population, and population dynamics (common to artificial, human, and organismal societies) can influence learning. The papers of Barfuss et al. (2), Bielawski et al. (3), Duéñez-Guzmán et al. (5), and Köster et al. (7) exemplify this theme across diverse settings. Although reinforcement learning in stable partnerships is still highly relevant [nowadays, perhaps most notably in large language models (23–26)], many–if not most–agents exist in an ambient population, and the associated population dynamics can influence learning similar to how demographic stochasticity influences evolutionary dynamics (27). Importantly, many problems in reinforcement learning are in populations whose dynamics are not governed by a central controller or institution, which leads to interesting dynamical systems and is partly why Barfuss et al. (2) propose equations (rather than just algorithms) to describe learning.

The second emergent theme is game theory informed by (machine) learning. Principles of learning, including natural cognitive processes as well as machine learning, can augment the simple models of adaptation that are typically found in evolutionary game theory. The papers of Barfuss et al. (2), Kleiman-Weiner et al. (10), McAvoy et al. (14), and Terrucha et al. (17) are situated well within this theme, as is possibly that of Tacchetti et al. (19). Incorporating learning dynamics expands the applicability and expressiveness of game-theoretic models, even when the underlying models themselves are only abstractions of reality. This theme cuts across the entire field of game theory, which includes evolutionary game theory in populations, as well as more classical settings such as two-player games and even mechanism design.

The third theme is AI for cooperation and decision-making. A common thread in populations at all scales—whether organismal or human, natural, or artificial—as well as in the papers of this special feature, is that the goals and incentives vary not only between groups but also between individuals in a group and the group as a whole. While there is overlap between this theme and the others, a distinguishing characteristic here is that the focus is less descriptive and more about how AI can inform behaviors themselves. Falling into this theme are the papers of Barfuss et al. (2), Kleiman-Weiner et al. (10), McAvoy et al. (14), Garcia and Traulsen (18), and Tacchetti et al. (19). These studies reflect and may eventually interact with modern, large-scale advances in AI to better engineer cooperation, collaboration, and sustainability practices.

Finally, we would be remiss if we did not mention foundation models, which are transforming many aspects of the world around us and will likely intersect with the themes covered here. However, whereas the connections between foundation models and these themes is an interesting area for future research, at present much of the existing work in multiagent reinforcement learning aligns more directly with game theory. As a result, multiagent reinforcement learning stands to benefit substantially from game-theoretic insights (especially from evolutionary models), while simultaneously providing computational methods and tools to enrich and extend more classical ideas in game theory. In the future, as foundation models continue to evolve and impact AI broadly, their eventual integration with the themes explored here–collective behavior, cooperation, and population dynamics–may lead to significant theoretical and practical advancements.

The work developed in this special issue provides solid groundwork for study how artificial agents might operate in multiagent settings, potentially enhancing our understanding of complex interactions. As Barfuss et al. (2) highlight, the goal is not always just performance but also understanding, which is especially important given the many open problems at the intersection of reinforcement learning and evolutionary game theory. Thus, we have reason to hope that the complementary topics highlighted in this special feature will open pathways for future interdisciplinary research.

1. K. Zhang, Z. Yang, T. Başar, *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, D. Cansever, Eds. (Springer International Publishing, Cham, 2021), pp. 321–384.
2. W. Barfuss *et al.*, Collective cooperative intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319948122 (2023).
3. J. Bielawski, T. Chotibut, F. Falniowski, M. Misiurewicz, G. Piliouras, Heterogeneity, reinforcement learning and chaos in population games. *Proc. Natl. Acad. Sci.* **122**, e2319929122 (2023).
4. J. Hofbauer, K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge University Press, 1998).
5. E. A. Duéñez-Guzmán *et al.*, Perceptual interventions ameliorate statistical discrimination in learning agents. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319933122 (2023).
6. K. M. Smith, C. L. Apicella, Partner choice in human evolution: The role of cooperation, foraging ability, and culture in hadza campmate preferences. *Evol. Hum. Behav.* **41**, 354–366 (2020).
7. R. Köster, E. A. Duéñez-Guzmán, W. A. Cunningham, J. Z. Leibo, Tabula rasa agents display emergent in-group behavior. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319947122 (2023).
8. F. Fu *et al.*, Evolution of in-group favoritism. *Sci. Rep.* **2**, 460 (2012).
9. D. B. Cooney, S. A. Levin, Y. Mori, J. B. Plotkin, Evolutionary dynamics within and among competing groups. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2216186120 (2023).
10. M. Kleiman-Weiner, A. Vientós, D. G. Rand, J. B. Tenenbaum, Evolving general cooperation with a Bayesian theory of mind. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2400993122 (2024).
11. E. Fehr, K. M. Schmidt, A theory of fairness, competition, and cooperation. *Quart. J. Econ.* **114**, 817–868 (1999).
12. G. Charness, M. Rabin, Understanding social preferences with simple tests. *Quart. J. Econ.* **117**, 817–869 (2002).
13. E. Hughes *et al.*, "Inequity aversion improves cooperation in intertemporal social dilemmas" in *Adv. Neural Inf. Proc. Syst.*, S. Bengio *et al.*, Eds. (Curran Associates, Inc., 2018), vol. 31.
14. A. McAvoy *et al.*, Unilateral incentive alignment in two-agent stochastic games. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319927122 (2023).
15. W. H. Press, F. J. Dyson, Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10409–10413 (2012).
16. L. S. Shapley, Stochastic games. *Proc. Natl. Acad. Sci. U.S.A.* **39**, 1095–1100 (1953).
17. I. Terrucha *et al.*, Humans program artificial delegates to accurately solve collective-risk dilemmas but lack precision. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319942122 (2023).

18. J. Garcia, A. Traulsen, Picking strategies in games of cooperation. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319925122 (2023).
19. A. Tacchetti *et al*., Deep mechanism design: Learning social and economic policies for human benefit. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319949122 (2023).
20. L. Hurwicz, The design of mechanisms for resource allocation. *Am. Econ. Rev.* **63**, 1–30 (1973).
21. R. B. Myerson, Optimal auction design. *Math. Oper. Res.* **6**, 58–73 (1981).
22. E. Maskin, Nash equilibrium and welfare optimality. *Rev. Econ. Stud.* **66**, 23–38 (1999).
23. A. Vaswani *et al*., "Attention is all you need" in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)* , I. Guyon *et al*., Eds. (Curran Associates, Inc., 2017), pp. 6000–6010.
24. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* , J. Burstein, C. Doran, T. Solorio, Eds. (Association for Computational Linguistics, 2019), pp. 4171–4186.
25. J. Kaplan *et al*., Scaling laws for neural language models. arXiv [Preprint] (2020). http://arxiv.org/abs/2001.08361 (Accessed 17 May 2025).
26. L. Ouyang *et al*., Training language models to follow instructions with human feedback. arXiv [Preprint] (2022). http://arxiv.org/abs/2203.02155 (Accessed 17 May 2025).
27. G. W. A. Constable, T. Rogers, A. J. McKane, C. E. Tarnita, Demographic noise can reverse the direction of deterministic selection. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4745–E4754 (2016).