Evolutionary Multi-agent Reinforcement Learning in Group Social Dilemmas

B. Mintz[1] and F. Fu[1]

Mathematics Department, Dartmouth College

(*Electronic mail: brian.a.mintz@dartmouth.edu.)

Reinforcement learning (RL) is a powerful machine learning technique that has been successfully applied to a wide variety of problems. However, it can be unpredictable and produce suboptimal results in complicated learning environments. This is especially true when multiple agents learn simultaneously, which creates a complex system that is often analytically intractable. Our work considers the fundamental framework of Q-learning in Public Goods Games, where RL individuals must work together to achieve a common goal. This setting allows us to study the tragedy of the commons and free rider effects in AI cooperation, an emerging field with potential to resolve challenging obstacles to the wider application of artificial intelligence. While this social dilemma has been mainly investigated through traditional and evolutionary game theory, our approach bridges the gap between these two by studying agents with an intermediate level of intelligence. Specifically, we consider the influence of learning parameters on cooperation levels in simulations and a limiting system of differential equations, as well as the effect of evolutionary pressures on exploration rate in both of these models. We find selection for higher and lower levels of exploration, as well as attracting values, and a condition that separates these in a restricted class of games. Our work enhances the theoretical understanding of evolutionary Q-learning, and extends our knowledge of the evolution of machine behavior in social dilemmas.

This study applies a powerful, widely used artificial intelligence framework to a social dilemma to study the issue of AI coordination. By using an evolutionary approach, we expand the understanding of the complex dynamics that this system exhibits.

―――――――――――――

I. INTRODUCTION:

The world has recently seen a surge of innovations powered by advanced artificial intelligence technologies, such as Large Language Models and self-driving cars, promising to fundamentally reshape many aspects of the world. As progress in these continues, we will see such systems deployed more extensively throughout the world. Already this has created complex systems of interacting agents with different goals and patterns of behavior. Understanding the theoretical basis of these systems will be crucial for successful implementations, and require new approaches with interdisciplinary ideas[1]. In particular, a pressing open question is how to ensure models act cooperatively while performing their given task well[2]. This is related to the issue of aligning the incentives used in training AI models with those of the broader society.

Reinforcement learning (RL) is one prominent framework for AI that has been successfully applied to many challenging problems due to its exceptionally general approach[3]. Whereas many machine learning techniques have constraints on the types of problems they can address, this approach can be applied to a wide variety of problems. The key idea behind reinforcement learning is simply that actions that have a positive payoff will be repeated more, and those with a negative payoff will be repeated less[4,5]. This model has its basis in models of animal psychology, but has since found a series of cutting-edge applications due to its flexibility. Reinforcement learning has been applied to a wide variety of problems from traditional games like Go, to stock price predictions, the management of energy systems, and controlling chaotic dynamics[6–8]. This approach can be unpredictable, especially when multiple agents learn simultaneously, as this creates a dynamic environment[9–13]. Such Multi-Agent Reinforcement Learning (MARL) systems have likewise seen many applications such as navigating groups of autonomous vehicles and distributing resources through communication systems[14,15]. This field draws attention from researchers studying complex systems, engineers seeking to improve algorithms, and policy makers trying to manage these technologies effectively.

Recently, there has been growing interest in combining reinforcement learning with evolutionary algorithms, a similarly general approach with simple motivation. Evolutionary algorithms use mutation and selection to solve complex problems[16–19]. Reinforcement learning has a natural connection to this method, as learning in a single agents is equivalent to evolutionary dynamics between the strategies that agent can use with a particular form of mutation[10,20]. Thus evolution between agents themselves can be seen as a form of multi-level selection, which has long been investigated by evolutionary theorists. The majority of the work in this intersection focuses on comparing and improving different algorithms, as opposed to understanding its theory[21–27]. Consequently, a wide variety of MARL architectures have been proposed, with studies determining when each is optimal for various applications[28–31].

Our work seeks to expand the theoretical understanding of these evolutionary MARL systems, eventually allowing a more principled approach to choosing or designing algorithms for MARL. Specifically, we investigate the effect of selection on a parameter governing the degree of exploration agents follow, and how this depends on the game determining agent interactions. Due to its relative analytic tractability, we focus on one of the foundational reinforcement learning models, Q-learning. We combine this with a variety of ideas from Evolu-

tionary theory to study dynamics within this system. To investigate cooperative AI, we focus on the evolution of learning in a classical social dilemma, the public goods game[32,33]. These are a canonical example of conflicting incentives, the interests of the self and those of the collective, and are suitably broad to encompass a wide variety of scenarios.

Most prior work on evolution in social dilemmas have studied simple, often static, strategies[34–37]. This is extended to more realistic agents in our work. Likewise, evolutionary MARL generalizes techniques like particle swarm optimization and simulated annealing to explore more efficiently[38,39]. Previous study into reinforcement learning in social dilemmas has mainly focused on the Prisoner's Dilemma[40–48]. However this game is limited to interactions between two individuals, as with most theoretical studies of MARL, limiting the complexity that can be found but also reducing the potential applicability of the findings. A few studies have considered MARL in the public goods game, investigating the effect of learning models and how to optimize cooperation[49–52]. The majority look to use MARL to explain test subjects' behavior in experiments,[53–59] Our study extends these preceding works by revealing a rich variety of evolutionary dynamics between learning agents in a group social dilemma.

II. MODEL

In this work, we expand the MARL framework to include evolutionary dynamics between agents. Inspired by evolutionary game theory, we investigate the population dynamics in a model where agents reproduce and die concurrently with learning. We study this through extensive agent based simulations of a stochastic model, and an evolutionary analysis of a corresponding deterministic system of ordinary differential equations.

Our model uses the foundational Q-learning algorithm to perform reinforcement learning. In general, each agent keeps a table of values for each possible state of the system and possible choice of action. These are updated according to

$$Q(s_{t+1}, a_{t+1}) = Q(s_t, a_t) + \alpha \left[r_{t-1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$
(1)

where s_t and a_t are the state / action at time t, α is the learning rate that determines how quickly values are updated, r_{t-1} is the reward received, and γ is the discount rate governing the extent the agent cares about future rewards. In essence, this generalizes the notion of keeping a weighted average of the rewards for a particular action, and also considers the change in state an action will cause. Actions can then be selected in a variety of ways, we use the Boltzmann function to determine policies, as it is better suited to mathematical analysis. With this method, actions are selected at random proportionally to the exponential of their payoff divided by a parameter T, the "temperature" of the agent, which controls the exploration of new strategies by the agent. As T approaches zero, only the action with the highest Q-value will be selected, corresponding to purely exploiting this best-known strategies. Temperature can vary with time or payoff relative to some aspiration

level, though it can also be a constant value. We focus on stateless Q-learning, as the variable group composition makes it challenging to specify meaningful states.

In this study, rewards are determined by the public goods game. This is a natural setting for arbitrary numbers of agents to interact, and has a long history of being used to understand group social dilemmas like the tragedy of the commons and free rider effects. Each agent, of a group of N, chooses whether to pay a cost of one to contribute, or not contribute, referred to as defection. The total contribution is then scaled by a reward function $f(x)$ and redistributed evenly. That is, the payoff of individual j is

$$\pi_j(c_1, ...c_N) = \frac{1}{N} f \left(\sum_{i=1}^{N} c_i \right) - c_i$$
(2)

where c_i is the contribution, zero or one, of individual i. Classically, the reward function can be linear $f(x) = rx$ with $1 < r < N$, or contain more complicated nonlinear effects. For this work, we include the division by N in the definition of $f(x)$ to more easily interpret it as the reward per individual, simplifying comparison between values of N. We'll represent these functions at the possible discrete levels of contribution with vectors $[f(0), f(1), f(2), f(3), ..., f(N)]$, since the intermediate values cannot be realized so are irrelevant. This allows for more precise control of the rewards, though it has the drawback of being harder to extend to larger groups.

Our first approach is an agent based simulation of a variation on a classical stochastic model of evolution, the Moran (death-birth) process[60]. A finite number of N individuals interact to receive payoffs. At each time step, every individual has an independent probability r of dying, then being replaced by the offspring of another member of the population, possibly with some mutation. The individual who gives birth for this replacement is selected proportionally to their fitness $f_i = e^{\beta \pi_i}$, the exponential of their average payoff π_i over all previous interactions. Using the exponential of fitness is a standard technique to avoid complications from negative fitness. The parameter β gives the strength of selection, in this study we mainly consider $\beta = 1$ for simplicity. Concurrently with the replacement, individuals perform Q-learning under Boltzmann Selection, with shared parameters T, α, and γ, to receive rewards from the public goods game with reward function $f(x)$. This function could be linear such as $f(x) = kx$, which is most commonly considered, or contain non-linear effects such as $f(x) = b_0 x + b_1 x^2$. Sample trajectories of this simulation are shown in Fig. 1. Death occurs uniformly at random at each iteration of the classical Moran process, so on average individuals only learn for N iterations before being replaced. Our probabilistic modification gives agents more time to learn, an expected $1/r$ iterations, since the independent death probability means the iterations individuals learn for follows a geometric distribution with parameter r. Intuitively, the replacement rate parameter r balances between learning, when it is low, and evolution, when it is high. We can then use this to estimate the fixation probability that a mutant will replace the resident type. These determine the evolutionary trajectories under rare mutations. In particular, they can

determine whether there will be positive or negative selection, or other effects like attractors and repellors of the dynamics.

Complementing this agent based simulation, our second model applies the evolutionary technique of adaptive dynamics using a limiting system of differential equations[61–64]. This is described more fully in the Appendix.

III. RESULTS

There is a significant effect from noise in the stochastic model, due to randomness in the action choices, which agents are selected to die, and which replace them. Because of this, careful consideration must be given to the model parameters to obtain meaningful results. In particular, if the rewards are large relative to the discount rate and temperature, it is possible that agents will prematurely fix their strategy and not explore sufficiently. Further, evolutionary dynamics are inherently more volatile in smaller populations, so larger groups or more trials are necessary. Controlling for these, we further account for the effect of the other parameters in this simulation by holding all others fixed before varying the parameter of interest, repeating this process for several combinations of the other parameters to ensure the results are robust. Figure 2 plots the average probability of contribution depending on the learning rate and discount factor, showing a large learning rate and small discount factor are ideal in these cases. In other games, it is possible large discount factors or low learning rates are better. The two games shown use linear reward functions $f(x) = kx$ with $k = 0.9$ and 1.1 so the jumps in reward for an additional contribution are slightly above and below the cost of contribution of one, so learning should always favor contributing in the former case, and not contributing in the latter. Despite this, they can result in similar levels of cooperation, as the payoff is mostly determined by the actions of the other group members given the large group size. The combined effects of learning and evolution, given by the temperature and replacement rate parameters, are presented in Figure 3. Depending on the temperature, we can see positive or neutral effects from replacement rate in a particular game. Across the replacement rates, we can see increasing temperature initially increases the contribution level, then decreases it to 0.5 as agents purely explore. This suggests intermediate temperature values perform best, consistent with previous findings. Consequently, even within a single game, different effects from these forces are possible.

Our deterministic model reveals a wide range of possible selection effects on temperature for different reward functions. Specifically, we find examples where temperature experiences positive and negative selection, as well as attraction to an intermediate value. To understand how the game influences this, we study the direction of selection across all possible reward functions. Specifically, while the invasion fitness is positive, we repeatedly take mutant values slightly above resident values. This gives a sense for the most likely evolutionary trajectory of the temperature. Since there are $n + 1$ values in the reward function for a group of n players, we restrict to small cases to assist with finding patterns. In particular, we

consider reward functions of the form $[0, j_0, j_0 + j_1, m]$ where $0 \leq j_0 \leq m$ and $0 \leq j_1 \leq m - j_0$. These correspond to a constraint where there is an upper limit m on the per individual reward, and that the reward function is weakly increasing. Since there are only two parameters for a fixed m, we can plot the final temperature over the j_0, j_1-plane, shown in Figure 4. When $m > 3$ we see there is a clear separation between regimes, selection is positive when $j_0 + j_1 = f(2)$ is above a threshold depending on m. For smaller values of m, selection is predominantly negative. The distinction is that for $m > 3$, it is possible for all jumps to be above the cost of contribution one, where contribution would always be beneficial. This corresponds to the region $j_0 > 1$ and $j_1 > 1$, which does experience consistent selection on temperature. The computation of fixation probabilities generally supports these results, though computational constraints limited further investigation into these.

IV. DISCUSSION

Our results suggest that the relationship between learning parameters and cooperation levels is quite sophisticated, and warrants further study. The temperature of agents can evolve in a variety of ways depending on the environment. This suggests that combination of evolutionary algorithms with reinforcement learning can be used to optimize cooperation. Our approach of computing fixations probabilities presents and more interpretable than traditional evolutionary algorithms, providing a broader picture of the possible evolutionary trajectories that might be taken. Further, analytic approaches that average over possible interaction can improve the efficiency of these investigations and allow for more robust results through reducing stochastic effects.

While a variety of reinforcement models have been proposed, we focus on a the foundational Q-learning due to its theoretical guarantees and relative analytic tractability. Our analytic approach to studying these dynamics makes a few strong assumptions. We replace the reward an action receives with the average over all possible interactions, losing a large amount of data. Consequently, agents with the same temperature are expected to follow the same learning dynamics, often contrasting simulations where they can easily diverge based on initial actions. Future work could allow a more complete description of the possible states of the system, extending previous mean-field approaches that characterize the evolution of the probability distribution of strategies in the population[65–67]. This would undoubtedly require a far more sophisticated mathematical model, and complicate the analysis. Another significant factor is the lack of states in our reinforcement learning model. This could simply be the number who cooperated in the previous round, or an average over the last several rounds. However this is a rather inaccurate measure of the true state, the strategies of all agents.
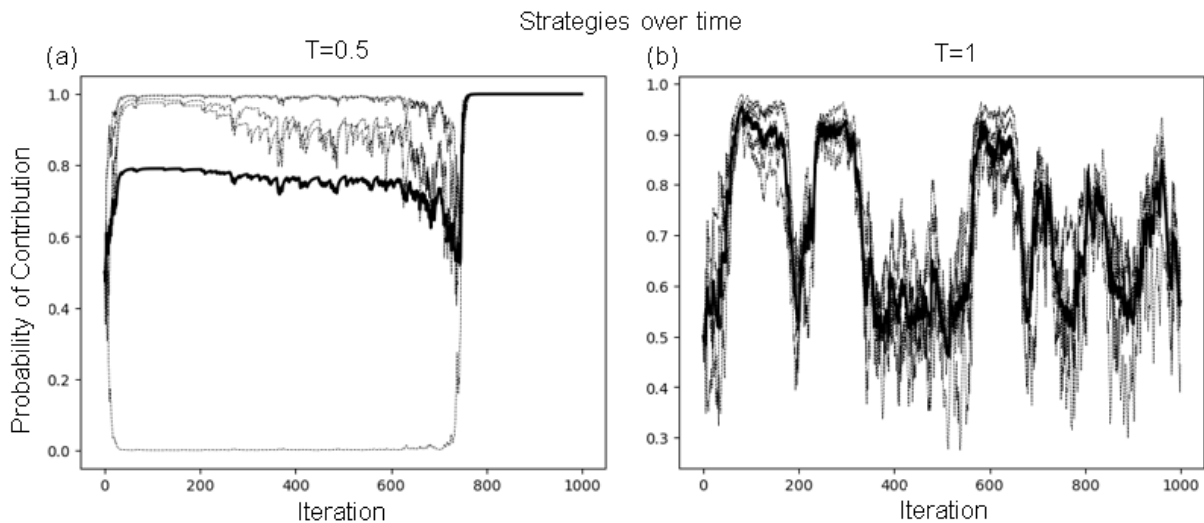
FIG. 1. **Stochastic learning dynamics with symmetric temperature**. These plots show the trajectories of strategies in the agent-based simulation over time as dotted lines, with the average strategy in bold, where the rewards are [0, 0, 0, 2, 4, 6], $N = 5$, $\gamma = 0$, $\alpha = 0.1$, $r = 0$, and $T = 0.5$ in panel (a) and $T = 1$ in panel (b). By varying the temperature, a range of behaviors are possible. For low temperatures, relative to learning rate and rewards, agents enter a self-reinforcing cycle where they choose the most beneficial action repeatedly. For large temperatures, the strategies fail to converge. We see good alignment with the predictions of the ODE model, that strategies cluster together when the temperatures are the same.
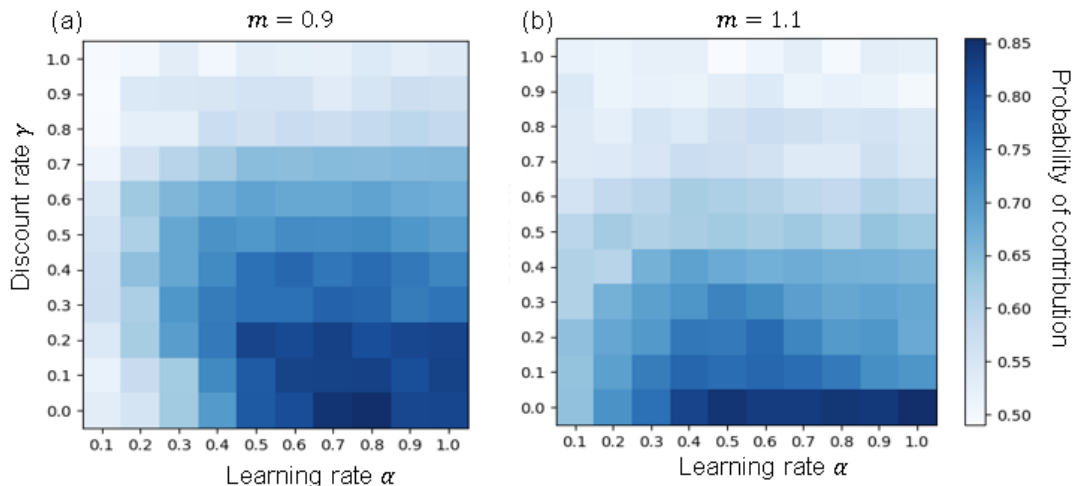

FIG. 2. **The optimal learning parameters vary with the reward function**. This plot shows the average, over 100 runs, strategy in the group after 500 iterations where the horizontal axis is the learning rate and vertical axis is the discount factor, both between zero and one. Further, $r = 0$ so there is no replacement, the temperature is $T = 0.5$, the population consists of five agents, and the reward function is linear $f(x) = kx$ with $k = 0.9$ on the left, and 1.1 on the right. In these cases the jumps are a constant of k, so in the first it is always slightly better to defect, despite this agents contribute 85% of the time with a higher learning rate and low discount rate. Similarly, on the right it is always slightly better to contribute, and a wider range of learning rates achieve a similar probability of contributing. Because of the nonzero temperature, it is impossible to achieve perfect cooperation, a strategy of one, and the achieved values are approximately the largest possible given this temperature and the possible rewards for each action.

V. CONCLUSION

In this work, we extended the MARL framework to allow for reproduction, introducing evolutionary pressures between agents. While previous studies have studied evolution on individuals with static strategies, or fixed groups with variable strategies, our model combines these two dynamics to create a more realistic system than either alone. By doing so, we are able to broaden the applications of evolutionary game theory, and bring theoretical insights into a complex learning model. We apply evolutionary techniques to study the dynamics in the temperature of agents, a key parameter governing their degree of exploration. In particular, we use agent based simulations to estimate the fixation probability of mutations in this param-
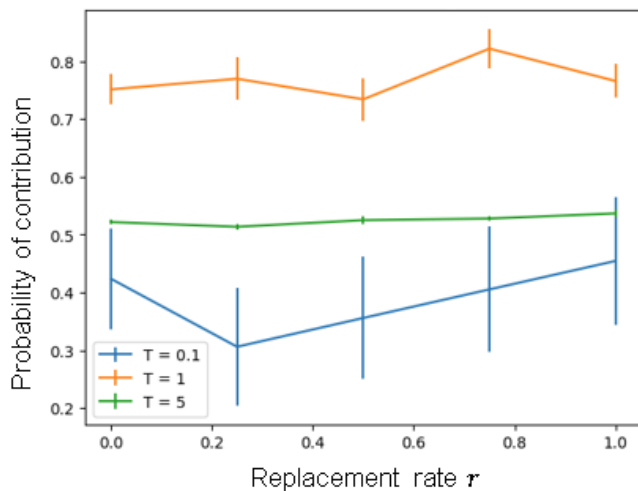
FIG. 3. **Learning and evolution can have varying effects on contribution levels**. By varying the learning rate T and replacement probability r, one can tune the relative strength of learning and evolution. This plots the average, plus or minus one standard error, over 20 runs of group level of contribution after 1000 iterations. These simulation are initialized with a single temperature, with no mutation in temperature, so selection is only acting on the strategies.
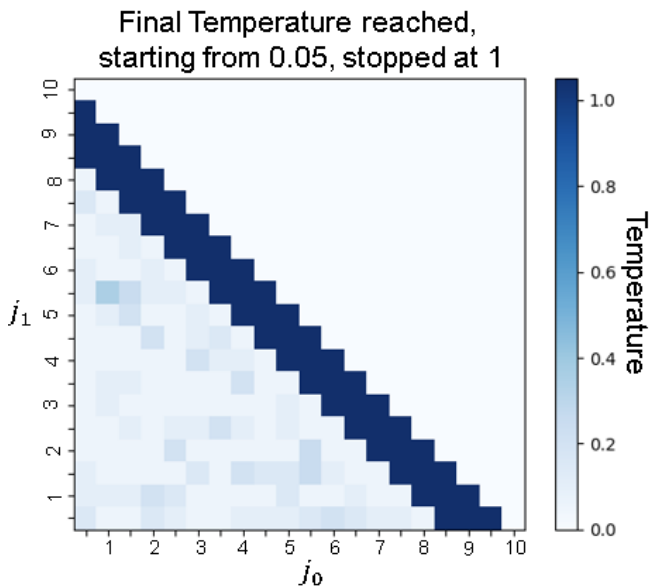

FIG. 4. **The reward function can lead to positive or negative selection**. This plot represents the most likely outcome of the evolutionary dynamics in the temperature parameter, starting from $T = 0.05$ and up to $T = 1$, over the space of possible reward functions for the three player game, found through the adaptive dynamics approach described in the appendix. Letting m be the maximum reward for when all individuals contribute, we can specify the function as $[0, j_0, j_0 + j_1, m]$ where j_0 and j_1 are the jumps in reward when an additional individual contributes, if zero or one other had already contributed. Assuming the reward function is increasing, we have $0 \leq j_0 \leq m$ and $0 \leq j_1 \leq m - j_0$, so only the values in the lower triangle are considered. We see there is a clear transition to larger final temperatures when $j_0 + j_1$ exceeds a threshold depending on m, in this case $m = 10$.

eters. These probabilities determine the evolutionary trajectories under the assumption of rare mutations. This assumption also allows us to apply adaptive dynamics using a system of ordinary differential equations to remove stochastic effects.

Through extensive simulations, we explored the intricate relation between learning parameters and cooperation. In particular, temperature and replacement rate could have a variety of effects depending on other parameters like the reward function determining the type of public goods game being played. Depending on this, the temperature could evolve up or down, or to an intermediate levels. By studying a restricted class of these games, we conjectured a condition that determines which type of selection the temperature will undergo.

There are multiple possible extensions of this work. Firstly, this framework could be applied to a number of other games, such as the Iterated Prisoners Dilemma or coordination games. In particular, a round number of continuation probability could be used to determine how long the group interacts for, allow for more or less learning to occur. By restricting to simpler cases, like small population sizes, we could derive further theoretical results. For example, the Moran process we simulated in model one could be explicitly represented to analytically determine the fixation probabilities. While the analytical results are often limited to simple cases, we could extend the simulations to capture more complicated effects, for example by including mutation in the parameter values. We could also consider a model where the initial strategy is also genetically determined, which would allow a more effective comparison between learning agents with nonzero temperature, and those who never change their strategy. We could also investigate the full replicator dynamics using the system of differential equations to determine finesses. This would provide a clearer picture of the dynamics between individuals with different temperatures. Future study in this direction has the potential to greatly enhance our understanding of the complex dynamics in Multi-agent Reinforcement Learning.

Our approach expands the theoretical understanding of combining genetic algorithms with reinforcement learning. While evolution can and has produced remarkable solutions to many challenging problems, some situations can be ill-suited to this approach. Our investigation highlights the fact that careful consideration must be given to minimize the stochastic effects that can overpower selection. Such understanding is crucial as we see a myriad of applications of reinforcement learning. Further, these results provide initial implications for the coordination of AI systems, which are trained to optimize their own performance, yet must work successfully with other individuals. As such technologies continue to develop and connect with more aspects of our world, understanding the interfaces between models of different capacities will become increasingly important.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix

By assuming mutations are sufficiently rare, Adaptive Dynamics simplifies evolution to competition between two types: the resident, and a mutant which either fixates to become the new resident or becomes extinct before the next mutant emerges. Invasion fitness is the difference in payoffs $E(m,r) - E(r,r)$ between the mutant trait m and resident trait r when the mutant is initially rare, and is often used as a proxy for the fixation probability. Indeed, if this is negative, the mutant will experience negative selection and likely die out. Typically adaptive dynamics also assumes small levels of mutation to use the gradient to determine dynamics, but the lack of a closed form for our dynamics makes this infeasible, freeing up consideration to non-local mutation, as in the above model. Here, we assume interaction continues for long enough for the dynamics to reach equilibrium, and remains there long enough that the equilibrium payoff approximates the average payoff accumulated throughout the whole interaction, the quantity determining fitness in the first model. Since mutation is rare, we assume the group of N consists of one individual having temperature m while the others have temperature r, and find the equilibrium numerically by solving the system over a sufficiently long time range, estimated on a case by case basis. Specifically, we use the initial condition $x(0) = 1/2$ since we assume the group forms without any prior information, so each agent initially follows a uniformly random strategy. Essentially, this approach separates the timescales between learning and evolution, performing selection based on the equilibrium reached.

To derive our system of differential equations, we follow[10] and consider stateless Q-learning with no discounting ($\gamma = 0$), simplifying the Q-value update equation to

$$Q_i(t+1) = Q_i(t) + \alpha[r_i(t) - Q_i(t)] \tag{A.1}$$

where $r_i(t)$ is the average reward of choosing action i at time t, and α and $Q_i(t)$ are as before. If actions are chosen with the Boltzmann mechanism with temperature T, then the probability x_i of choosing action i is $x_i(t) = \frac{\exp(Q_i(t))/T}{\sum_i \exp(Q_i(t))/T}$. Taking the time derivative and rearranging and scaling time by α/T, we find

$$\frac{\dot{x}_i}{x_i} = \left[r_i - \sum_k x_k r_k\right] - T\sum_k x_k \ln\frac{x_i}{x_k} \tag{A.2}$$

This first term corresponds to increasing the probability of choosing an action that has an above average payoff, and the second corresponds to the energy in a statistical-mechanical system. In fact, this shows that some reinforcement learning methods can be viewed as an evolutionary process within an agent between actions. When there are just two actions, we can summarize an agent's strategy by a single number x, the probability of choosing the first action. This gives the now single equation

$$\frac{\dot{x}}{x} = [r_1 - (xr_1 + (1-x)r_2)] - T\left(x\ln\frac{x}{x} + (1-x)\ln\frac{x}{1-x}\right) \tag{A.3}$$

or equivalently

$$\dot{x} = x(1-x)\left(r_1 - r_2 - T\ln\frac{x}{1-x}\right) \tag{A.4}$$

In the game we study the actions are contribution and defection, and the difference in expected payoffs is

$$r_1 - r_2 = \sum_{S\subseteq\{1,\dots,N-1\}} [f(|S|+1) - 1 - f(|S|)]\prod_{i\in S} x_i \prod_{i\notin S}(1-x_i) \tag{A.5}$$

the average difference in payoffs for contribution and defection over all possible subsets S of individuals contributing, weighted by the probability of that outcome. These dynamics alone can exhibit a large degree of complexity, Fig. 5 depicts the loci of equilibria over the space of possible reward function for just three players, and a given temperature. Finally, the full system we consider has separate equations for x_m and x_r, the strategies of those with the mutant and resident temperatures respectively, and the x_i in $r_1 - r_2$ are either x_m or x_r as appropriate.

[1] M. Tsvetkova, T. Yasseri, N. Pescetelli, and T. Werner, "A new sociology of humans and machines," Nature Human Behaviour **8**, 1864–1876 (2024).

[2] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, (2021).

[3] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," Journal of artificial intelligence research **4**, 237–285 (1996).

[4] C. J. Watkins and P. Dayan, "Q-learning," Machine learning **8**, 279–292 (1992).

[5] C. Szepesvári, *Algorithms for reinforcement learning* (Springer nature, 2022).

[6] A. G. Barto, P. S. Thomas, and R. S. Sutton, "Some recent applications of reinforcement learning," in *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems* (2017).

[7] A. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," Renewable and Sustainable Energy Reviews **137**, 110618 (2021).

[8] M. A. Bucci, O. Semeraro, A. Allauzen, G. Wisniewski, L. Cordier, and L. Mathelin, "Control of chaotic systems by deep reinforcement learning," Proceedings of the Royal Society A **475**, 20190351 (2019).

[9] S. Leonardos and G. Piliouras, "Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory," Artificial Intelligence **304**, 103653 (2022).

[10] A. Kianercy and A. Galstyan, "Dynamics of boltzmann q learning in two-player two-action games," Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **85**, 041145 (2012).

[11] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," IEEE transactions on cybernetics **50**, 3826–3839 (2020).

[12] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers, "Evolutionary dynamics of multi-agent learning: A survey," Journal of Artificial Intelligence Research **53**, 659–697 (2015).

[13] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," Artificial Intelligence **136**, 215–250 (2002).

[14] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," Artificial Intelligence Review , 1–49 (2022).

[15] W. Du and S. Ding, "A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications," Artificial Intelligence Review **54**, 3215–3238 (2021).
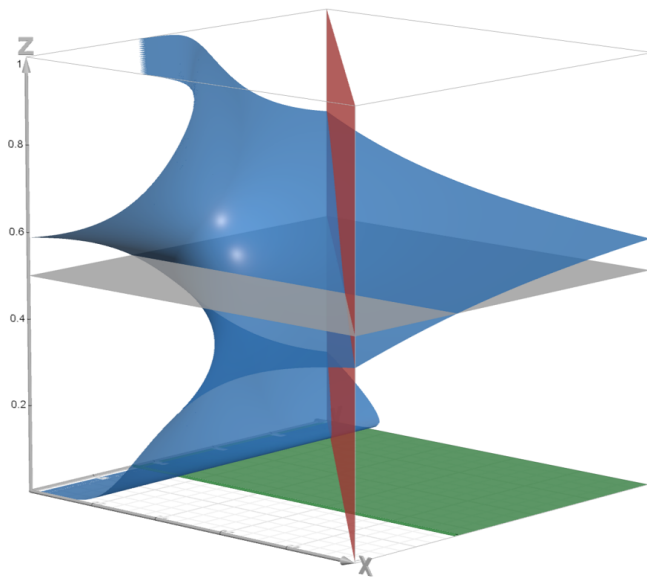
FIG. 5. **Null-manifold of symmetric learning dynamics over the space of three player Public Goods Games**. This plots the equilibria of the learning dynamics where $j_0 = x$, $j_1 = y$, and z is the strategy, assuming $j_2 = m - j_0 - j_1$. Specifically, this plot uses $m = 3$ and $T = 0.1$. The red plane delineates the rejoin $j_0 + j_1 \leq m$, and the green region is the subset where the initial rate of change of the strategy is positive. In this case, the maximum equilibrium contribution level, that is reached from an initial strategy of 0.5, occurs when y is on this boundary, and x is small. Note a large range of x, from zero to 0.1, have approximately the same level of contribution. Additionally, these values are close to having a negative initial change in the strategy, likely making them unstable for the actual dynamics, and thus possibly resulting in less frequent contribution.

[16]R. Miikkulainen and S. Forrest, "A biological perspective on evolutionary computation," Nature Machine Intelligence **3**, 9–15 (2021).

[17]B. Li, Z. Wei, J. Wu, S. Yu, T. Zhang, C. Zhu, D. Zheng, W. Guo, C. Zhao, and J. Zhang, "Machine learning-enabled globally guaranteed evolutionary computation," Nature Machine Intelligence , 1–11 (2023).

[18]K.-S. Tang, K.-F. Man, S. Kwong, and Q. He, "Genetic algorithms and their applications," IEEE signal processing magazine **13**, 22–37 (1996).

[19]L. Haldurai, T. Madhubala, and R. Rajalakshmi, "A study on genetic algorithm and its applications," Int. J. Comput. Sci. Eng **4**, 139–143 (2016).

[20]K. Tuyls, K. Verbeeck, and T. Lenaerts, "A selection-mutation model for q-learning in multi-agent systems," in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems* (2003) pp. 693–700.

[21]A. Sehgal, H. La, S. Louis, and H. Nguyen, "Deep reinforcement learning using genetic algorithm for parameter optimization," in *2019 Third IEEE International Conference on Robotic Computing (IRC)* (IEEE, 2019) pp. 596–601.

[22]S. Dominic, R. Das, D. Whitley, and C. Anderson, "Genetic reinforcement learning for neural networks," in *IJCNN-91-Seattle International Joint Conference on Neural Networks*, Vol. 2 (IEEE, 1991) pp. 71–76.

[23]D. E. Moriarty, A. C. Schultz, and J. J. Grefenstette, "Evolutionary algorithms for reinforcement learning," Journal of Artificial Intelligence Research **11**, 241–276 (1999).

[24]F. Liu and G. Zeng, "Study of genetic algorithm with reinforcement learning to solve the tsp," Expert Systems with Applications **36**, 6995–7001 (2009).

[25]X. Wu, Q. Zhu, W.-N. Chen, Q. Lin, J. Li, and C. A. C. Coello, "Evolutionary reinforcement learning with action sequence search for imperfect information games," Information Sciences , 120804 (2024).

[26]M.-J. Kim, J. S. Kim, and C. W. Ahn, "Evolving population method for real-time reinforcement learning," Expert Systems with Applications **229**, 120493 (2023).

[27]H. Bai, R. Cheng, and Y. Jin, "Evolutionary reinforcement learning: A survey," Intelligent Computing **2**, 0025 (2023).

[28]Q. Zhu, X. Wu, Q. Lin, L. Ma, J. Li, Z. Ming, and J. Chen, "A survey on evolutionary reinforcement learning algorithms," Neurocomputing **556**, 126628 (2023).

[29]Y. Lin, F. Lin, G. Cai, H. Chen, L. Zou, and P. Wu, "Evolutionary reinforcement learning: A systematic review and future directions," arXiv preprint arXiv:2402.13296 (2024).

[30]Z. Tan and K. Li, "Differential evolution with mixed mutation strategy based on deep reinforcement learning," Applied Soft Computing **111**, 107678 (2021).

[31]Z. Liu, B. Chen, H. Zhou, G. Koushik, M. Hebert, and D. Zhao, "Mapper: Multi-agent path planning with evolutionary reinforcement learning in mixed dynamic environments," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2020) pp. 11748–11754.

[32]A. Szolnoki and M. Perc, "Conditional strategies and the evolution of cooperation in spatial public goods games," Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **85**, 026104 (2012).

[33]F. C. Santos, M. D. Santos, and J. M. Pacheco, "Social diversity promotes the emergence of cooperation in public goods games," Nature **454**, 213–216 (2008).

[34]C. Hauert, M. Holmes, and M. Doebeli, "Evolutionary games and population dynamics: maintenance of cooperation in public goods games," Proceedings of the Royal Society B: Biological Sciences **273**, 2565–2571 (2006).

[35]S. Kurokawa and Y. Ihara, "Emergence of cooperation in public goods games," Proceedings of the Royal Society B: Biological Sciences **276**, 1379–1384 (2009).

[36]B. Mintz and F. Fu, "Social learning and the exploration-exploitation trade-off," Computation **11**, 101 (2023).

[37]B. Mintz and F. Fu, "The point of no return: Evolution of excess mutation rate is possible even for simple mutation models," Mathematics **10**, 4818 (2022).

[38]D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," Soft computing **22**, 387–408 (2018).

[39]M.-P. Song and G.-C. Gu, "Research on particle swarm optimization: a review," in *Proceedings of 2004 international conference on machine learning and cybernetics (IEEE Cat. No. 04EX826)*, Vol. 4 (IEEE, 2004) pp. 2236–2241.

[40]W. Barfuss and J. M. Meylahn, "Intrinsic fluctuations of reinforcement learning promote cooperation," Scientific Reports **13**, 1309 (2023).

[41]T. W. Sandholm and R. H. Crites, "Multiagent reinforcement learning in the iterated prisoner's dilemma," Biosystems **37**, 147–166 (1996).

[42]S. S. Izquierdo, L. R. Izquierdo, and N. M. Gotts, "Reinforcement learning dynamics in social dilemmas," Journal of Artificial Societies and Social Simulation **11**, 1 (2008).

[43]M. Kaisers and K. Tuyls, "Faq-learning in matrix games: Demonstrating convergence near nash equilibria, and bifurcation of attractors in the battle of sexes." in *Interactive Decision Theory and Game Theory* (2011).

[44]N. Masuda and M. Nakamura, "Numerical analysis of a reinforcement learning model with the dynamic aspiration level in the iterated prisoner's dilemma," Journal of theoretical biology **278**, 55–62 (2011).

[45]M. W. Macy and A. Flache, "Learning dynamics in social dilemmas," Proceedings of the National Academy of Sciences **99**, 7229–7236 (2002).

[46]L. Wang, D. Jia, L. Zhang, P. Zhu, M. Perc, L. Shi, and Z. Wang, "Lévy noise promotes cooperation in the prisoner's dilemma game with reinforcement learning," Nonlinear Dynamics **108**, 1837–1845 (2022).

[47]C. Yu, M. Zhang, F. Ren, and G. Tan, "Emotional multiagent reinforcement learning in spatial social dilemmas," IEEE transactions on neural networks and learning systems **26**, 3083–3096 (2015).

[48]A. Oroojlooy and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," Applied Intelligence **53**, 13677–13722 (2023).

[49]O. Leimar and J. M. McNamara, "Learning leads to bounded rationality and the evolution of cognitive bias in public goods games," Scientific reports **9**, 16319 (2019).

[50]H. H. Nax and M. Perc, "Directional learning and the provisioning of public

goods," Scientific reports **5**, 1–6 (2015).

[51] X. Wang, Z. Yang, G. Chen, and Y. Liu, "Enhancing cooperative evolution in spatial public goods game by particle swarm optimization based on exploration and q-learning," Applied Mathematics and Computation **469**, 128534 (2024).

[52] L. Wang, L. Fan, L. Zhang, R. Zou, and Z. Wang, "Synergistic effects of adaptive reward and reinforcement learning rules on cooperation," New Journal of Physics **25**, 073008 (2023).

[53] A. Iwasaki, S. Imura, S. H. Oda, I. Hatono, and K. Ueda, "Does reinforcement learning simulate threshold public goods games?: a comparison with subject experiments," IEICE TRANSACTIONS on Information and Systems **86**, 1335–1343 (2003).

[54] Y. Horita, M. Takezawa, K. Inukai, T. Kita, and N. Masuda, "Reinforcement learning accounts for moody conditional cooperation behavior: experimental results," Scientific reports **7**, 39275 (2017).

[55] M. A. Janssen and T.-K. Ahn, "Learning, signaling, and social preferences in public-good games," Ecology and society **11** (2006).

[56] J. Arifovic and J. Ledyard, "Scaling up learning models in public good games," Journal of Public Economic Theory **6**, 203–238 (2004).

[57] C. Bühren, J. Haarde, C. Hirschmann, and J. Kesten-Kühne, "Social preferences in the public goods game–an agent-based simulation with econsim," Plos one **18**, e0282112 (2023).

[58] C. R. Cotla, "Learning in repeated public goods games-a meta analysis," Available at SSRN 3241779 (2015).

[59] W. Hichri and A. Kirman, "The emergence of coordination in public good games," The European Physical Journal B **55**, 149–159 (2007).

[60] M. A. Nowak, *Evolutionary dynamics: exploring the equations of life* (Harvard university press, 2006).

[61] U. Dieckmann, J. A. Metz, M. Sabelis, and K. Sigmund, "Adaptive dynamics of infectious diseases," Pursuit of virulence management , 460–463 (2002).

[62] O. Diekmann *et al.*, "A beginner's guide to adaptive dynamics," Banach Center Publications **63**, 47–86 (2004).

[63] X. Chen and F. Fu, "Outlearning extortioners: unbending strategies can foster reciprocal fairness and cooperation," PNAS nexus **2**, pgad176 (2023).

[64] Å. Brännström, J. Johansson, and N. Von Festenberg, "The hitchhiker's guide to adaptive dynamics," Games **4**, 304–328 (2013).

[65] A. Galstyan, "Continuous strategy replicator dynamics for multi-agent q-learning," Autonomous agents and multi-agent systems **26**, 37–53 (2013).

[66] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International conference on machine learning* (PMLR, 2018) pp. 5571–5580.

[67] R. Carmona, M. Laurière, and Z. Tan, "Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning," The Annals of Applied Probability **33**, 5334–5381 (2023).

[68] W. Barfuss, "Dynamical systems as a level of cognitive analysis of multi-agent learning: Algorithmic foundations of temporal-difference learning dynamics," Neural Computing and Applications **34**, 1653–1671 (2022).

[69] M. Ueda, "Memory-two strategies forming symmetric mutual reinforcement learning equilibrium in repeated prisoners' dilemma game," Applied Mathematics and Computation **444**, 127819 (2023).

[70] Y. Usui and M. Ueda, "Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner's dilemma," Applied Mathematics and Computation **409**, 126370 (2021).

[71] J. E. Agudo and C. Fyfe, "Reinforcement learning for the n-persons iterated prisoners' dilemma," in *2011 Seventh International Conference on Computational Intelligence and Security* (IEEE, 2011) pp. 472–476.

[72] M. Harper, V. Knight, M. Jones, G. Koutsovoulos, N. E. Glynatsi, and O. Campbell, "Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma," PloS one **12**, e0188046 (2017).

[73] U. ManChon and Z. Li, "Public goods game simulator with reinforcement learning agents," in *2010 Ninth International Conference on Machine Learning and Applications* (IEEE, 2010) pp. 43–49.

[74] C. Hilbe, B. Wu, A. Traulsen, and M. A. Nowak, "Cooperation and control in multiplayer social dilemmas," Proceedings of the National Academy of Sciences **111**, 16425–16430 (2014).