Cooperation between Independent Reinforcement Learners under Wealth Inequality and Collective Risks

Ramona Merhej
INESC-ID and Instituto Superior Tecnico
Universidade de Lisboa, Portugal
ISIR, Sorbonne University, France
merhejramona@gmail.com

Fernando P. Santos
Dept. of Ecology and Evolutionary Biology
Princeton University, USA
Informatics Institute, University of Amsterdam, NL
fpsantos@princeton.edu

Francisco S. Melo
INESC-ID and Instituto Superior Tecnico
Universidade de Lisboa, Portugal
fmelo@inesc-id.pt

Francisco C. Santos
INESC-ID and Instituto Superior Tecnico
Universidade de Lisboa, Portugal
franciscocsantos@tecnico.ulisboa.pt

ABSTRACT

We study how wealth inequality influences behavioral dynamics in groups of independent reinforcement learners facing a threshold public goods dilemma with uncertain returns. The game allows individuals to contribute or not to a common pool to reduce their chances of future losses. The non-linearity introduced by the threshold, the stochasticity introduced by the risk and the wealth heterogeneity of players result in a game setting with multiple equilibria. We find that the learners' dynamics in this case play a major role in determining the attained equilibrium point. Our results suggest that, under individual-based learning, wealth inequality can have sizable effects on the emerging collective behaviors, decreasing the overall chances of group success. Moreover, we compute the class-based Nash equilibria (i.e., where same wealth-class agents are assumed to play the same strategy) for this game and compare the performance of groups composed of independent learning agents with the performance obtained under the payoff maximizing class-based Nash equilibrium. We find that the learned strategies never really match optimal performance for all tested values of risk.

KEYWORDS

Social Dilemmas; Collective Risk Dilemmas; Public Good Games; Heterogeneous Agents; Reinforcement Learning

ACM Reference Format:
Ramona Merhej, Fernando P. Santos, Francisco S. Melo, and Francisco C. Santos. 2021. Cooperation between Independent Reinforcement Learners under Wealth Inequality and Collective Risks. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

1 INTRODUCTION

Social Dilemmas are situations where a conflict between personal and collective interests can lead rational individuals to adopt strategies that result in a sub-optimal solution for everyone involved [20]. Understanding the necessary conditions for cooperation to evolve in such contexts has prompt academic interest [2]. The problem becomes more compelling in dilemmas with multiple equilibria. In that context, not all Nash Equilibrium points can be equally cooperative or beneficial and, as a result, learning dynamics can lead to so-called coordination failures. The most commonly studied game of this type is the infinitely iterated prisoners' dilemma (IPD) with infinitely many Nash equilibria [39]. In the IPD, for example, both Always Defect and Tit-For-Tat [1] are strategies constituting a Nash Equilibrium when adopted by both players – with the former leading to substantially lower returns than the latter. The equilibrium reached in these games depends on the starting point of agents and their learning dynamics that define the evolution from there to the attained equilibrium. It is therefore relevant to, not only study the static properties and equilibrium points of such games, but also the process that learning agents use to reach different equilibria [14].

While the IPD is heavily studied in literature, games that are better representative of the real world often involve more than two players, have non-linear payoffs and may be asymmetric [9, 30, 41]. One interesting example is the general n-player threshold Public Goods Game (PGG) where possibly heterogeneous players need to merge their individual efforts to avoid a common risk. PGGs with collective risk are also known in the literature as Collective Risk Dilemma (CRD). In relation with the ongoing Covid-19 crisis, the game portrays the mass cooperation needed in following the prescribed, yet costly, safety measures (e.g. wearing a mask, using a tracing app, staying at home etc.) to avoid a disastrous spread of the disease. In this case, heterogeneity can come from people belonging to different risk classes according to their age or health conditions. Dilemmas involving risk and non-linear returns also exist in collective insurance arrangements, where heterogeneity can result from different levels of risk-exposure to natural hazards [44]. Another real life illustration of CRDs are the international negotiations to avoid the disastrous outcomes of climate change [30, 41, 60]. Here, heterogeneity appears as wealth inequality where different countries may have different contribution capacities. In our work we focus primarily on wealth inequalities between agents. We find this game interesting for our study both for its symbolic depiction of many real world problems and its attractive computational properties. The introduction of the threshold results in a non-linear payoff function while the risk factor adds stochasticity and uncertainty to the game. Initial inequality between players can highly alter the evolution of the game and we encourage its incorporation in games where a common good or disaster is shared.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

The aforementioned characteristics will establish a dilemma with multiple equilibrium points. The prevailing equilibrium will depend on the agents' learning dynamics, which is often modeled through multi-agent individual (reinforcement) [3, 34, 48, 51, 55, 56] or social (imitation) learning algorithms [41, 49]. Further details about the game are given in Section 3.

As mentioned earlier, in multi-equilibria games, learning dynamics can impact the equilibrium point that agents converge to. A rational agent would study such a game statically and decide from there on an optimal strategy to implement. Humans, on the other hand, are rationally bounded and often select their strategies dynamically, after a learning process. This has been confirmed again and again when humans in real life or experiments adopt far from optimal or rational strategies [12, 14, 26, 50]. Reinforcement Learning (RL) has been applied successfully to model human learning dynamics in a social dilemma [40]. The success of this simple algorithm in predicting human behaviors can be attributed to its characteristics that capture some of human learning traits. In particular, the model implements the Law of Effect [54] that reinforces the usage of previously successful actions and the Power Law of Practice [32] that generates a learning curve that is steep at the beginning and flattens with time.

Several variations of RL algorithms exist for the multi-agent scenario with some specifically designed to better perform in social dilemmas [13, 18]. However, these algorithms either require additional sharing of information between agents or hard-code the learning of an already known and desired strategy. We note that the goal of our paper is not to adapt RL algorithms to avoid social dilemmas or to choose the best performing one but rather to understand and describe what prevents or incites cooperation among adaptive learning agents. Since information sharing or the final desired policy may not always be at hand in real life, and especially in large populations, we focus on fully individual-based independent RL, an algorithm that relaxes this constraint and requires agents to only be aware of their own strategies and returns.

Additionally, RL allows for an easy implementation of stochastic policies that are desirable in our game. When involved in a threshold game, every contribution over the threshold is an unnecessary cost on society. A contribution of a fraction of the society can thus be more beneficial than a contribution of the whole society and hence more cooperation is not always advantageous. Implementing stochastic policies allows players to take turns in cooperating and defecting, creating fairer contribution arrangements (possibly leading to egalitarian average contributions over time). Suppose for example a threshold game where at every step of the game, 50% of the players need to contribute to achieve the target. Allowing the agents to follow stochastic strategies will distribute the responsibility of achieving the target on different players at every round. Instead, having agents converge to deterministic policies will establish players that either always cooperate or always defect and hence the same players will always incur the cooperation costs while the others will always get to free-ride. RL allows agents to *learn* such stochastic policies with no need to pre-engineer or pre-define a set of discretized possible stochastic strategies to select from.

Under this setting, we investigate how wealth inequality impacts cooperation and hence overall group achievements of a population involved in a threshold public good game with a collective risk. The final cooperation levels reached will depend on our chosen learning algorithm. We question whether independent individual-based learning agents can reach rational outcomes by comparing the performance of learned strategies to a statically extracted best (payoff maximizing) class-based Nash Equilibrium.

The paper is structured as follows: we begin in Section 2 by introducing related work from the literature. Following that, in Section 3, we give a formal definition of the threshold PGG game. In Section 4 we expose the difficulties of finding Nash-Equilibria points for the game described and define class-based Nash equilibrium points. We continue in Section 5 by describing the agents' learning dynamics. In Section 6 we present our results. We present the overall achievement of agents trained with RL and compare our results to the ones obtained with the payoff maximizing class-based Nash equilibrium. Finally we conclude in Section 7 with a recapitulation of our findings and a discussion on future works.

2 RELATED WORK

We present a descriptive study on heterogeneous agents that dynamically learn with reinforcement learning a solution for a multi-equilibria social dilemma. We investigate the impact of inequality on the performance of the agents' learned solutions. In a continuous PGG and under Evolutionary Game Theory (EGT) dynamics, strong inequality in wealth, productivity and benefits can inhibit cooperation [16]. In a threshold PGG and under EGT again, wealth inequality was found to help in achieving cooperation if rich/poor individuals can imitate each other regardless their wealth category [60]. Otherwise, wealth inequality can be strongly detrimental. An experimental study showed that wealth inequalities in a threshold PGG made cooperation, and hence collective success, harder to accomplish [53]. In collective risk game studies and under evolutionary dynamics, it was found that rich individuals contribute more than poor ones who only chose to cooperate if all rich players cooperated [5, 61]. However, experimental data did not always confirm these predictions and rich individuals were found to under-contribute while poor individuals over-contributed [5]. Thus, current results suggest that under different learning dynamics (real humans and EGT), inequality has sometimes disparate consequences. Reinforcement learning, where agents adapt their choices based on their past experiences, is yet another learning paradigm and the consequences of inequality under RL-dynamics may reveal distinctive peculiarities.

The majority of descriptive papers on social dilemmas are simulated with EGT dynamics. But RL and EGT follow two different learning paradigms. While in the former learning is social-based, in the latter it is individual-based. In fact, in EGT, strategies are learned through imitation of others based on relative performance, a process aptly referred to as social learning. In this context, if individuals face a cooperative dilemma, defectors are always better off than cooperators, and, as a result, will be imitated by others and their choice will spread throughout the population [33, 55]. On the other hand, in a Multi-Agent Reinforcement Learning (MARL) setting with independent learners [3, 4, 6, 8, 10, 11, 27, 28, 34, 40, 57, 58], agents learn individualistically, i.e. they aim at maximizing their return based on their own experience and disregard the returns of other players. Therefore what prevents cooperation in one setting might not necessarily translate to the other. Nonetheless, several relevant equivalences can be shown [3, 4, 57, 58] among these two learning paradigms.

These potential contrasting particularities can echo on the final learned policies. In two distinct studies [47] and [46] for example, agents play the same game but follow respectively evolutionary (social) and reinforcement learning (individualistic) dynamics. While the effect of key parameters and voting rules remains similar, agents can converge to dissimilar strategies depending on the learning paradigm. Individual and social learning can also alter the role of complex networks in coordination dilemmas [58].

Although the process of imitation learning that hinders cooperation under EGT is not present in RL, major challenges still inhibit the emergence of cooperation. In a study on sequential social dilemmas with deep RL [22], the authors identify coordination sub-problems. Coordination problems that prevent proper cooperation are quite common in MARL and are not only restricted to social dilemmas. Also, as in evolutionary games, RL agents are equally prone to develop complex cyclic dynamics and behavioral ecologies, depending on the complexity of the problem [35].

A survey on coordination problems in MARL [29] discusses some of its major challenges including the Pareto-selection problem, the non-stationarity problem, the shadowed equilibrium problem, the stochasticity problem etc. A lot of research has been dedicated to overcome these obstacles and aid agents in finding cooperative strategies. P. Kollock [20] classifies cooperation solutions into three categories: motivational, strategic and structural. Motivational solutions modify the objective or reward function of agents to promote cooperative behaviors. Notably, in RL, intrinsic rewards can be engineered and added to environmental rewards to help agents solve a sub-problem of the game and facilitate the emergence of coordination [24]. Strategic solutions start with a known and desired equilibrium and construct algorithms that converge to these specific points. Examples include an algorithm designed to always asymptotically behave as a Tit-for-Tat strategy by learning simultaneously a cooperative and a selfish Q-function and alternating between them to avoid exploitability [18]. The last and more general type of solutions are the structural ones. Structural solutions include centralized learning for example that can facilitate coordination and cooperation between agents. This type of framework provides agents with additional information during their training phase compared to information available at execution time. For instance, during training, allowing agents to observe additionally the states and actions of their opponents, allows them to better seize the dynamics of the game and hence perform better at execution time when this additional data is missing [25]. Other solutions consider opponent modeling [13, 17] where agents not only model the dynamics of the world but also the dynamics of agents in it.

In this paper, we examine the interactions between heterogeneous RL agents in a collective risk dilemma and observe how these can influence overall cooperation levels. Our goal is a descriptive one, that is, to understand emerging dynamics in a multi-agent learning setting and the particular impacts of inequality and risk; nonetheless, knowledge about emerging dynamics from this complex system may be useful to direct future works in developing new algorithms, along the solution classification detailed above [20], that guarantee cooperation in social dilemmas.

3 GAME DEFINITION AND NOTATIONS

A Public Good is a common resource shared between individuals regardless their contribution to generate it [20]. The challenges associated with managing these goods are captured in the so-called Public Goods Games (PGG). In PGGs, those who do not contribute are always better off than those who do, which creates an incentive for people to free-ride i.e. to profit from the common good that others contributed to create. On the other hand if all people decide to free-ride, no common good can exist which leads to a sub-optimal solution. In a PGG with a collective risk, the public good is modeled as the avoidance of a common disaster which benefits everyone even those who did not help in avoiding it. It is also referred to as a Collective Risk Dilemma (CRD) [9, 30, 41, 43, 59, 60].

Formally, in a CRD of N players, every player i is granted an initial endowment b_i. He can then choose to either invest nothing or a fraction c of it to a common pool. The benefits gained by investing in a common pool are modeled by the increased chances of avoidance of an otherwise common risk of probability r. Should the players manage to collect jointly a sum greater than a target threshold $\mathbf{t}$, then the disaster is avoided with certainty. Otherwise, with the disaster probability r, all players lose a fraction p of whatever they have left of their initial endowments. At the end of the game, the expected endowment of player i who started with b_{i,t_0} is

$$b_{i,t_\infty} =(1-c_i)b_{i,t_0} - rp(1-c_i)b_{i,t_0}\Theta\left(\mathbf{t} - \sum_{j=1}^{N} c_j b_{j,t_0}\right) \qquad (1)$$

where Θ is the Heaviside step function and c_i, c_j represent the binary choices of the different players of either contributing 0 or a fraction c to the pool. The total good, defined as the sum of all players' endowments, becomes

$$B_{t_\infty} = \sum_{j=1}^{N}(1-c_j)b_{j,t_0} - rp\sum_{j=1}^{N}(1-c_j)b_{j,t_0}\Theta\left(\mathbf{t} - \sum_{i=1}^{N} c_i b_{i,t_0}\right) \qquad (2)$$

For the game to be a social dilemma, total cooperation needs to result in higher payoffs than total defection. Looking at equation 2, this means that the threshold $\mathbf{t}$ first needs to be lower bounded by zero, otherwise total defection would also avoid the disaster and cooperation would be unnecessarily expensive. Second the threshold needs to be achievable with less than a total cooperation ($\mathbf{t} < cB_{t_0}$), otherwise the incentive to free-ride is eliminated and the dilemma is broken. As a result, we have $B^C_{t_\infty} = (1-c)B_{t_0}$ for total cooperation and $B^D_{t_\infty} = (1-rp)B_{t_0}$ for a total defection. To have a dilemma, $B^C_{t_\infty}$ must be larger than $B^D_{t_\infty}$ which translates to $c < rp$.

The threshold $\mathbf{t}$, the contribution fraction c, the risk value r, the damage p and the initial endowments b_i are objective variables that will influence the chosen strategies by the agents (i.e., contribution c_i). But another more subjective variable, the payoff, will also play a crucial role in the agents' learning. In game theory, a game is usually defined by its payoff matrix that represents the benefits of a joint action for a given player. We note that in real life, this (perceived) benefit, often called *utility*, is largely subjective and depends on how humans perceive or value a given return. Since the game presents multiplicative dynamics i.e. costs of cooperation and failure are relative to initial endowments, we define our payoffs as the log difference in endowments between two time-steps. This utility function captures what is described in economy as a diminishing marginal utility. In our case, this means that a loss of 70% of one's possessions, for example, is equally painful for any individual even if with different initial endowments, in absolute value, the losses are not equal. A large literature exists about utility functions and their representative meanings. It would be interesting to compare results obtained for a same game under different utility functions or even introduce heterogeneity between agents' perceived utility. For the moment however, this remains out of the scope of our paper and we consider a homogeneous log utility for all players.

We confirm that our chosen payoff function satisfies the second condition for a social dilemma. In any given round of a dilemma, a defector should be better off than a cooperator. For any two players i and j involved in the same game, if i cooperated and j did not, then j should receive a higher payoff than i. The log difference in endowments ensures that $\log \frac{b_{i,t_\infty}}{b_{i,t_0}} < \log \frac{b_{j,t_\infty}}{b_{j,t_0}}$ is satisfied for $c_i = c$ and $c_j = 0$ for all values of b_{i,t_0}, b_{j,t_0} (see equation 1). Hence, with our chosen utility function and a proper selection of c, r, p and $\mathbf{t}$, we can explore learning dynamics in the context of a social dilemma.

3.1 Introduction of wealth inequalities

We consider a population of finite size Z of which a fraction z_R is rich and holds a fraction w_R of the total riches W [60]. The remaining fraction $z_P = 1 - z_R$ of the population is poor and holds $w_P = 1 - w_R$ of the riches. The total wealth held by the rich/poor is equally distributed within the same wealth class. All poor players start with an equal initial endowment $b_P = \frac{W \times w_P}{Z \times z_P}$ and correspondingly all rich players start with the same endowment $b_R = \frac{W \times w_R}{Z \times z_R}$ where w_R and z_R are set such that $b_R > b_P$. Individuals are sampled from the population and organized into groups of size N. Such groups can contain $0, 1, ..., N$ rich individuals and respectively $N, N-1, ..., 0$ poor individuals. The individuals of the group now engage in a Collective Risk Dilemma. Participants can choose (with a certain probability) to contribute a constant fraction c of their endowment to the collective pool to help achieve a target threshold $\mathbf{t}$. We set $\mathbf{t}$ to be proportional to the contribution fraction c and to the average wealth in the population $b = \frac{W}{Z}$ with a factor of proportionality M such that $\mathbf{t} = Mcb$. The larger the value of M, the harder it is to reach the threshold. If the overall amount of contributions in the group is above that threshold, the target will be met and the disaster avoided. Otherwise and with probability r — the risk of occurrence of the collective disaster — individuals in the group will lose a fraction p of whatever they have. We assume that the players have a log utility function [37] and receive as rewards the difference in the log of their wealth before and after a game was played. Hence, a successful game will cost $x_C = \log\left(\frac{b_i - cb_i}{b_i}\right) = \log(1-c)$ for a cooperator and nothing for a defector since $x_D = \log\left(\frac{b_i}{b_i}\right) = 0$. Similarly, we can derive that a failure of avoiding the disaster will cost cooperators $\bar{x}_C = \log(1 - c - p(1-c))$ and defectors $\bar{x}_D = \log(1-p)$. The goal of each player is to find a probabilistic strategy π_i^*, representing the probability of player i choosing to cooperate, that maximizes his payoff. Table 1 summarizes the payoff matrix of the game.

Table 1: Payoff matrix of the game based on player's action and the outcome of the game.

Strategy	Successful game	Failed game
C	$x_C = \log(1-c)$	$\bar{x}_C = \log(1-c-p-pc))$
D	$x_D = 0$	$\bar{x}_D = \log(1-p)$

3.2 Numerical Values

We study the above game with a population of $Z = 200$ individuals. The rich represent $z_R = 20\%$ of the population and hold $w_R = 50\%$ of the total riches. The average wealth in the population is set to $b = 1$ yielding $W = Z$. The agents are involved in a game of $N = 6$ players with a threshold set to Mcb where $M = \frac{N}{2}$ and $c = 0.1$. The collective risk occurs with probability $r = 0.3$ if the threshold target is not achieved and the penalty paid in that case is $p = 0.7$ or 70% of the remaining wealth. The above used values satisfy the conditions necessary for a social dilemma (see Section 3)

4 CLASS-BASED NASH EQUILIBRIUM

As mentioned in Section 1, we wish to determine whether individual-based learning will lead agents to converge towards a rational equilibrium. To answer this question, we first study our game from a static perspective to try and extract its equilibrium points.

In general, computing Nash equilibria in large, general-sum games poses computational challenges [7]. To our best knowledge, most available algorithms such as [31, 38] obtain solution points for games where the payoff of a player is equal to the sum of the payoffs of his interactions with each player in the game. This is not the case in our threshold game where the joint payoff is not a linear combination of 2-player interactions. In the following, making use of the fact that the game is symmetric for players from the same wealth class, we define a *class-based Nash equilibrium* to transform our game into a 2-person matrix game.

Since we do not examine inequality emerging from co-existence or within a class, but rather between classes, we define as *class-based Nash equilibrium*, the Nash equilibrium point of the game *if* all players of the same wealth class are forced to follow the same strategy i.e. all rich players cooperate equally with a probability π_R an all poor players cooperate equally with a probability π_P. In other words, we impose and pre-condition our equilibrium on absolute equality and fairness within a given wealth class.

Consider a group of $N-1$ individuals and denote by n_R and n_P respectively, the number of rich and poor individuals within this group where $n_R \in \{0, 1, ..., N-1\}$ and $n_P = N-1-n_R$. Let n_R^c be the number of rich players that actually contribute to the pool i.e. $n_R^c \in \{0, 1, ..., n_R\}$ and n_P^c be the number of poor contributors in the group i.e. $n_P^c \in \{0, 1, ..., n_P\}$. Hence, a total number of $(n_R + 1) \times (n_P + 1)$ different pool contributions are possible.

The probability $P^{n_R}(n_R^c, n_P^c)$ with which each of these possible configurations occur in a group of n_R rich individuals follows a binomial law and depends on π_R and π_P.

$$P^{n_R}(n_R^c, n_P^c) = \binom{n_R}{n_R^c} \pi_R^{n_R^c} (1-\pi_R)^{n_R - n_R^c} \binom{n_P}{n_P^c} \pi_P^{n_P^c} (1-\pi_P)^{n_P - n_P^c} \tag{3}$$

Let i be the N^{th} player to join the group. Player i will now choose to contribute with probability π_R if he's rich or with probability

π_P if he's poor. Denote by A^D the action of defecting and not contributing, by A_R^C the contribution action of a rich individual and by A_P^C the contribution action of a poor one. Denote by $\mathcal{S}_D$ the set of configurations that achieve the threshold without the need of i's contribution. Mathematically, $\mathcal{S}_D = \{\forall\ (n_R^c, n_P^c) \in \{0, 1, ..., n_R\} \times \{0, 1, ..., n_P\} | n_R^c b_R c + n_P^c b_P c \geq Mbc\}$. Identically, denote by $\mathcal{S}_{A_P^C}$ the set of configurations that can achieve the threshold if i contributes and is poor and by $\mathcal{S}_{A_R^C}$ the set of configurations that can achieve the threshold if i contributes and is rich. The probability of the group achieving the threshold given that i chose action $a \in \{A^D, A_R^C, A_P^C\}$ is given by the sum of the probabilities of the events in $\mathcal{S}_D$, $\mathcal{S}_{A_P^C}$ and $\mathcal{S}_{A_R^C}$ respectively.

$$P^{n_R}(\mathbf{t}|a) = \sum_{(n_R^c, n_P^c) \in \mathcal{S}_a} P^{n_R}(n_R^c, n_P^c) \tag{4}$$

Since the game is probabilistic, the probability of a player avoiding or not a disaster given that he chose action a are given by

$$\begin{aligned} P^{n_R}(\text{success}|a) &= P^{n_R}(\mathbf{t}|a) + (1-r)P^{n_R}(\neg\mathbf{t}|a) \\ P^{n_R}(\text{failure}|a) &= 1 - P^{n_R}(\text{success}|a) \end{aligned} \tag{5}$$

We can now write the expected payoff functions of player i depending on whether he's rich or poor. Let $\mathcal{H}_R^{n_R}$ and $\mathcal{H}_P^{n_R}$ be the respective expected payoff functions of a rich and poor individual involved in a game with n_R rich players and where all rich follow strategy π_R and all poor follow strategy π_P. The expected payoff of an agent depends on whether the game was successful or not and whether he contributed or not to the common pool. We have

$$\begin{aligned} \mathcal{H}_R^{n_R}(\pi_R, \pi_P) =& \pi_R [P^{n_R}(\text{success}|A_R^C) x_C + P^{n_R}(\text{failure}|A_R^C)\bar{x}_C] + \\ & (1-\pi_R)[P^{n_R}(\text{success}|A^D) x_D + P^{n_R}(\text{failure}|A^D)\bar{x}_D] \end{aligned} \tag{6}$$

$$\begin{aligned} \mathcal{H}_P^{n_R}(\pi_R, \pi_P) =& \pi_P [P^{n_R}(\text{success}|A_P^C) x_C + P^{n_R}(\text{failure}|A_P^C)\bar{x}_C] + \\ & (1-\pi_P)[P^{n_R}(\text{success}|A^D) x_D + P^{n_R}(\text{failure}|A^D)\bar{x}_D] \end{aligned} \tag{7}$$

where x_C, $\bar{x}_C$, x_D and $\bar{x}_D$ are the payoffs described in Table 1.

Finally, since groups are sampled randomly, the expected payoff needs to account for the probability of an agent to find himself in a group with n_R rich individuals i.e.

$$\begin{aligned} \mathcal{H}_R(\pi_R, \pi_P) &= \sum_{n_R} \frac{\binom{Z_R-1}{n_R}\binom{Z-Z_R}{N-n_R-1}}{\binom{Z-1}{N-1}} \mathcal{H}_R^{n_R}(\pi_R, \pi_P) \\ \mathcal{H}_P(\pi_R, \pi_P) &= \sum_{n_R} \frac{\binom{Z_R}{n_R}\binom{Z-Z_R-1}{N-n_R-1}}{\binom{Z-1}{N-1}} \mathcal{H}_P^{n_R}(\pi_R, \pi_P) \end{aligned} \tag{8}$$

Both rich and poor players aim at maximizing their respective payoff functions $\mathcal{H}_R$ and $\mathcal{H}_P$. A Nash equilibrium (π_R^*, π_P^*) satisfies

$$\begin{aligned} \mathcal{H}_R(\pi_R^*, \pi_P^*) &\geq \mathcal{H}_R(\pi_R, \pi_P^*) \qquad \forall\ \pi_R \in [0,1] \\ \mathcal{H}_P(\pi_R^*, \pi_P^*) &\geq \mathcal{H}_P(\pi_R^*, \pi_P) \qquad \forall\ \pi_P \in [0,1] \end{aligned} \tag{9}$$

We have thus transformed the general N-player game into a two-person matrix game (rich and poor) in a larger action space. The joint poor player's pure action set is $\mathcal{A}_P = \{0C, 1C, \ldots, n_P C\}$ i.e. 0 to n_P poor players may cooperate and similarly, the joint rich player's is $\mathcal{A}_R = \{0C, 1C, \ldots, n_R C\}$. We look for algorithms that search for equilibrium points in 2-player matrix games but find them inapplicable to our game. In fact, the algorithms [23] search for any optimal probability distribution over the action space in the simplex. Since joint actions in our game emerge from a combination of individual actions following π_R and π_P, the probability distribution over the joint action space needs to follow a binomial distribution.

We therefore rely on a graphical method and discretize the domain of π_R and π_P into intervals of length $\epsilon = 0.001$. We calculate the corresponding payoff $\mathcal{H}_R$ and $\mathcal{H}_P$ over the space of possible (π_R, π_P). Referring to equations 6 and 7, we plot for every π_P, R's best response π_R^{BR} i.e. π_R^{BR} $s.t.$ $\mathcal{H}_R(\pi_R^{BR}, \pi_P)$ is maximized and similarly for every π_R, P's optimal response π_P^{BR}. The intersections of the hence formed lines represent class-based Nash equilibrium points i.e. strategies from which no player can deviate alone while increasing his payoff and assuming that within the same wealth class all players use the same strategy. We extract these points for different game configurations i.e. different risk values r. When several such points exist, we opt for the most performing one that we define as the one maximising the total expected payoffs of all players. We label this point the *best* class-based Nash Equilibrium. We use the performance obtained under the best class-based Nash equilibrium as the baseline to evaluate how rational the learned strategies with our algorithm are.

5 LEARNING ALGORITHM

The objective of our work is descriptive and we aim to understand the evolution of cooperation under reinforcement learning dynamics. To achieve this goal we train a population of independent RL learners with the Roth-Erev Algorithm [40]. We mean by independent that the learners do not model the presence of other players and perceive the emerging dynamics as part of their environment's dynamics. At update-step $k = 0$, before any learning is done, every player has an initial propensity to cooperate C or defect D. The propensity vector of player i holding the propensity values of these actions at a given update-step k is denoted $\mathbf{q}_i(k) = \left[q_{iC}(k), q_{iD}(k)\right]^T$. When player i engages in a game by selecting action A, he receives a payoff x from the game environment and updates his propensity vector according the the equations

$$\begin{aligned} q_{iA}(k+1) &= (1-\phi)q_{iA}(k) + x \\ q_{i\neg A}(k+1) &= (1-\phi)q_{i\neg A}(k) \end{aligned} \tag{10}$$

where ϕ is interpreted as a forgetting parameter needed to inhibit the propensities from growing to infinity. We leave out two other proposed extensions of the model, which are the extinction in finite time and the local experimentation. The extinction in finite time causes the propensities of some actions to go to 0 in finite time whereas we wish to maintain stochastic policies and the local experimentation parameter assumes that actions are ordered in a way such that close actions have close or similar payoffs, a characteristic that does not apply to our set of actions.

Because the game is a single state environment, this algorithm resembles the Q-learning algorithm [6] with the difference that at every update-step k, because of ϕ, the values of the propensities for all actions and not only the selected one, are updated.

We train asynchronously agents of a population learning with the update rule of equation 10. We choose asynchronous learning because we perceive it as more intuitive and natural. A comparison between synchronous and asynchronous learning in [45] showed no significant differences in results for players learning to play the Ultimatum Game. The procedure is summarized in Algorithm 1. At every update-step k, a group of N agents is selected randomly from the population of Z agents. The agents in this group engage in the game described in Section 3. Every player j in the group chooses randomly one of his available actions following probabilities $\mathbf{p}_j(k)$ that are derived by normalizing his propensity vector $\mathbf{q}_j(k)$. The selected actions and the game risk factor r determine whether or not the game is successful (i.e. if agents avoided the disaster). The payoffs for each agent are calculated according to Table 1 after which all agents in the group update their propensity vectors. This is repeated for a total of K update-steps. While training, we keep track of the number of times every agent in the population has been selected in a vector $\mathbf{u}$. Since the algorithm does not guarantee that all agents are chosen equally as many times, we define K', the minimum number of update-steps every agent needs to have performed before training is done. If after K total update-steps, some agent still hasn't performed at least K' updates, then training continues until this condition is satisfied.

Algorithm 1: Roth-Erev RL algorithm in an adaptive population with asynchronous updates of propensities.

```
Init: K total number of update-steps, K' minimum number of
      updates per agent
for $i \leftarrow 1$ to $Z$, population size do
    $\mathbf{q}_i(0) \leftarrow$ random initialization;
    $u_i \leftarrow 0$            /* tracks number of updates */
for $k \leftarrow 1$ to $K$ do
    1. sample random group $G$ of size $N$;
    2. sample actions $A_j \sim \mathbf{p}_j(k)$ for $j \in G$ (Eq. 11);
    3. evaluate game success;
    4. calculate payoff of $j \in G$ (Tab. 1);
    5. update $\mathbf{q}_j$ (Eq. 10);
    6. $u_j \leftarrow u_j + 1$ for $j \in G$;
    7. $u_{min} \leftarrow min(\mathbf{u})$
while $u_{min} < K'$ do
    repeat steps 1. to 7.
```

5.1 Parameters of the model

Following the description of the Roth-Erev algorithm, if all players start with the same initial propensities, the learning model depends on the strength of these propensities $Q_0 = \sum_A q_{iA}(0)$ which will determine the rate of learning and on the ratio $\frac{q_{iC}(0)}{q_{iD}(0)}$. A higher Q_0 will diminish the effect of the payoffs allowing for slower learning and more experimentation. The original authors suggest to initialize Q_0 to be in the order of the average payoff [40]. We choose to begin with a Q_0 that is 20 times the order of the average payoff to increase initial experimentation and decrease the chance of having players stuck in local minima caused by the high stochasticity of the game.

Since payoffs are negative or zero, we use the $Softmax$ to normalize the propensity vector. At any step k of the learning process, player i will select action A with probability

$$p_{iA}(k) = \frac{\exp(q_{iA}(k))}{\sum_{A \in \{C,D\}} \exp(q_{iA}(k))} \tag{11}$$

with $q_{iA}(0)$ sampled from a normal distribution $\mathcal{N}(\mu = -10, \sigma = 1)$. This will generate players with a random initial slight preference to defect or cooperate such that $\log_e\left(\frac{q_{iC}(0)}{q_{iD}(0)}\right) \sim \mathcal{N}(\mu' = 0, \sigma' = \frac{2\sigma}{\mu})$ (derived by following the log domain transformation of Katz [19]). We train the population for a total number of update-steps $K = 2,500,000$ and impose a minimum number of updates $K' = 30,000$ for every agent. The forgetting parameter is set to $\phi = 0.001$. All simulations are repeated for 5 runs.

6 RESULTS

In the following we display the results obtained for populations of agents learning to play the game introduced in Section 3 with the RL algorithm of Section 5. Unless otherwise indicated, the numerical values given in Section 3.2 and Section 5.1 are used.

The effectiveness of the learned strategies is evaluated based on a population's capability of achieving the required target threshold $\mathbf{t}$ when its agents are implementing given strategies. We refer to this capability as the performance or the overall population achievement η [9, 10, 59, 60]. We estimate η using a Monte Carlo method. Given a game and a strategy for each player of a population, we split our population into groups of N players. Within each group, the players choose to cooperate or not following the given strategies. The percentage of successful groups that actually reach the target threshold defines the performance of the population. This value being a random variable, we repeat and average the results over 10^6 simulations.

6.1 Effect of introducing wealth inequalities

For our first experiment, two types of populations are trained under varying risk value factors r. The first population P_1 is made of heterogeneous agents presenting initial wealth inequalities (see Section 3.1). The second population only counts homogeneous agents with an initial wealth equal to the average wealth of P_1. The results are plotted in Figure 1a.

As expected, and similar to results found under EGT [41], the performance of a population with and without inequalities increases with the risk factor r. Agents are more willing to cooperate and achieve the target if the consequences of failure are larger.

However, while under evolutionary dynamics diversity in a society has proven to increase cooperation rates [42] and wealth inequalities specifically were proven to encourage cooperation in some configurations [60], we find that wealth inequality decreases the overall achievement of a population. This is in conformity with some experimental results such as in [53], where wealth inequalities were found to inhibit group achievement. Looking closer to see what hindered cooperation and hence performance, we find two main responsible factors: the low initial endowment b_P that poor players receive and the group size N.

In fact, for the chosen group size $N = 6$, a contribution by a poor agent represents around 20% of the needed threshold $\mathbf{t}$ for success. This low impact of the cooperative action in reaching the threshold,

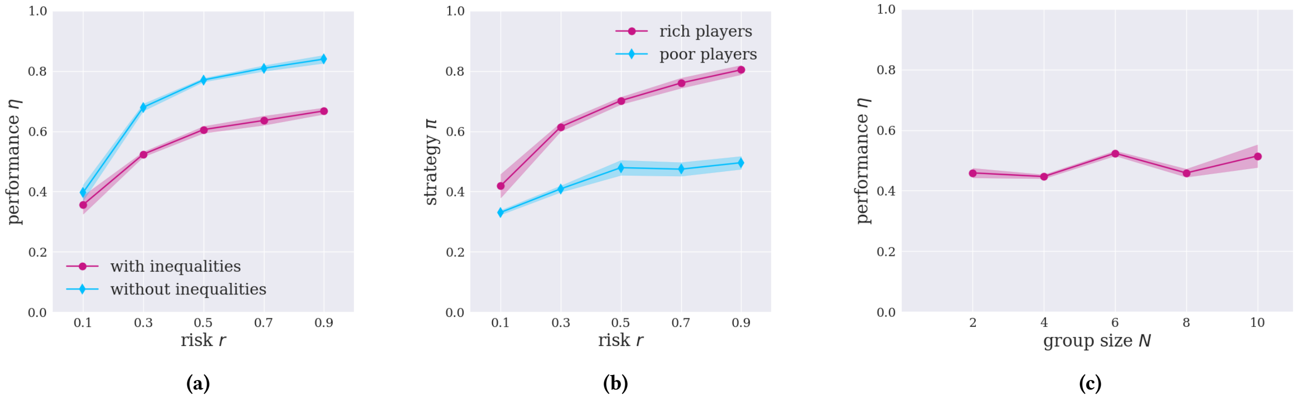

Figure 1: (a) Overall group achievement with and without wealth inequalities w.r.t. the risk factor r. (b) Learned strategies of rich and poor players with respect to the game's risk r. (c) Overall group achievement for a population with wealth inequalities with respect to the group size N. In all panels, shaded areas represent the standard deviation over 5 runs.

makes it harder for poor players to capture the purpose of cooperating. This is also visible when looking at how players modified their policies with respect to the varying risk (see Figure 1b). While the red curve of the rich remains steep, i.e. they are reactive and understand that they need to cooperate more when the risk of a disaster is larger, the blue curve seems more stagnant. The group size here plays another important role. For $N = 6$, the probability of sampling a purely poor group is around 25%. Learning in such groups requires coordination (5 out of 6 poor players need to cooperate to reach the threshold). This can further impede effective learning of cooperation. Again, when evaluating the performance of the population, 25% of the groups will be purely poor. With an average cooperation of 41% for $r = 0.3$ for example, only 4.5% of these groups will reach the threshold. This alone drops overall population performance by 24%.

We repeated the simulations with a linear-utility function to confirm that the properties found extend beyond log-utility functions. The log-utility slightly increases cooperation (which is expected because players are more risk-averse), yet, overall, the relative impact of wealth inequality remains the same.

6.2 Effect of the group size

Group size seems to have a considerable effect on overall achievement under evolutionary dynamics [15, 21, 36, 52]. For this reason and because of the previously found results in Section 6.1, we choose to evaluate the impact of the group size on a population's performance. Again we train our heterogeneous population P_1 to play the game of Section 3 with varying group sizes N. We find that under RL dynamics and with wealth inequalities, the overall achievement of the population does not vary strongly for groups of 2, 4, 6, 8 and 10 players. The results are plotted in Figure 1c.

We investigate how the combination of wealth inequalities and RL can diminish the effect of the group size. We notice that opposite and canceling dynamics seem to emerge from group size variation with some increasing and others decreasing overall achievement. First, as group size increases, for rich players, the impact of a contribution decreases from 250% of the target threshold when $N = 2$ to 50% of the threshold in groups with $N = 10$. More coordination is needed and learning and achieving the threshold becomes more difficult even for rich players. Purely poor groups also have smaller chances of reaching the threshold when N increases. While 2 poor agents need to cooperate to reach the threshold for $N = 2$, a coordination of 8 cooperators is necessary for $N = 10$. This phenomena decreases overall achievement. However, for larger group sizes, purely poor groups that are the least performing groups and usually cause a large drop in the population's performance, become much less common to sample and this, on the other hand, increases overall achievement. With the two forces pushing in opposite directions, performance seems to be little affected by group size under wealth inequality. Future experiments will be key to understand the relative weight of these competing group size effects.

6.3 Effectiveness of RL in achieving best class-based Nash performance

Finally, we investigate the effectiveness or optimality of the learned strategies. We plot in Figure 2 the overall performance of our population P_1 found in Section 6.1 and that of a population following the payoff maximizing class-based Nash strategy found in Section 4.

We can see that learned strategies achieve lower performance rates than the class-based Nash ones for $r \geq 0.3$ but that for $r = 0.1$, learned policies seem to over-perform. It is worth mentioning that for $r = 0.1$ and with the chosen values of $c = 0.1$ and $p = 0.7$, the dilemma is broken because of the relative high cost of cooperation compared to the cost of failure (mathematically, the condition $c < rp$ from Section 3 is not satisfied). Cooperation can become more costly than failure. Indeed, when we compared the expected payoffs of the two populations, the under-performing class-based Nash one incurred smaller costs than the more performing P_1. It is unwise for a population to cooperate under these conditions.

We specify that this strong drop in optimal Nash performance is not observed for populations without wealth inequalities. There, cooperation costs can be better distributed over the population and performance rates more easily increased. As a numerical example, if all agents cooperate with a probability of 62%, a population without

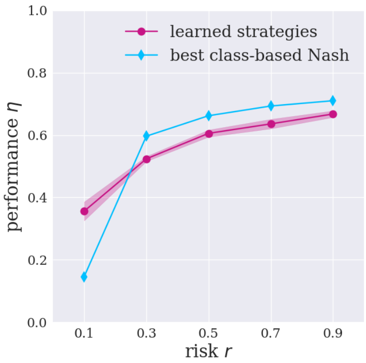

Figure 2: Performance (i.e., fraction of groups achieving the threshold) of populations with wealth inequalities following learned vs. best or payoff maximizing class-based Nash strategies. Besides maximizing performance as shown in this plot, we confirm that, the best class-based Nash also maximizes social welfare.

inequalities achieves the target 85% of the time compared to 63% of the time for a population like P_1. The combination of smaller cooperation costs and higher achieved performance when in a population without inequalities makes cooperating more desirable.

6.4 Intra and Inter-class fairness with RL

When computing the best (payoff maximizing) class-based Nash strategies, we assumed total fairness within the class of rich/poor players in the sense that cooperation costs were equally distributed and all followed the same policy. However, the learned strategies with RL present variations within each wealth class and not all players of the same wealth class follow the same policy. To increase intra-class fairness, we re-ran our study with synchronous training to see if a less randomized way of learning can decrease variations within a class. However, in all our experiments, we find similar results for the average and standard deviations of the learned strategies under both synchronous and asynchronous training.

Regarding inter-class fairness, we compare the average cooperation of rich vs. poor players. In our class-based Nash points, rich players cooperated with probability $\pi_R = 1$ whenever there was a dilemma ($r \geq 0.3$) while the poor never cooperated more than 40%. The responsibility of target achievement was heavily focused on a small class of rich players. Contrarily, under reinforcement learning, while the rich still cooperate more than the poor, the gap between the two classes is smaller. In all experiments posing a social dilemma, the average cooperation rate of rich players varied between 60 and 80% and that of the poor between 40 and 50%.

7 CONCLUSION AND DISCUSSION

We have studied how in a RL setting, wealth inequalities affect the overall achievement of a threshold target and how under these inequalities, parameters like risk and group size impact performance. We found that wealth inequalities lower overall achievement rates of a population. As for the group size, the results differed from what was obtained under EGT, and group size barely affected performance. We confirmed that RL agents have trouble converging to the payoff maximizing class-based Nash Equilibrium points and always incur unnecessary costs either by over-performing in low risk games or under-performing for higher risk values.

In our work, we introduced heterogeneity as an initial wealth inequality between agents. Yet heterogeneity can take several other forms. The risk factor r can vary between players. For example, the risk that a virus presents on people varies with their age, health etc. Similarly, the impact p of a disaster can vary. The consequences of ocean level rising are different in different geographical areas. Heterogeneity can also be introduced in the cost of cooperation c, or even emerge from uncertainty [9] or interaction structure [42]. Heterogeneous values of p and c will generate heterogeneous payoff matrices. But not only p and c can influence the payoff matrix. The utility function shortly mentioned in Section 3 can be different for different agents. Finally, the learning algorithm or other learning parameters (e.g. the learning rate or players' initial preferences for available actions) may differ in a population. Such heterogeneities are omnipresent in the real world and it would be interesting to incorporate them in future studies since many complex dynamics can emerge that we might be overlooking when assuming symmetry. Furthermore, future works shall study different levels of heterogeneity, for example, by testing different combinations of rich/poor distributions and inequality (i.e., different z_R/z_P and w_R/w_P).

Moreover, we point to the fact that RL, although technically allowing for the emergence of intra-class fairness, shows relatively high variance in the learned strategies within a wealth class. We have tested synchronous learning but, in that setting, we found no improvements in intra-class fairness. We encourage research for finding means of increasing fairness within classes, in populations adapting with RL. On the other hand, RL can increase inter-class fairness compared to the recommended class-based Nash strategies. This increase in fairness may result in a non optimal distribution of cooperation between rich and poor, lowering performance and incurring additional costs on the population. This should be kept in mind in future works where solutions with a trade-off between cost optimization and fairness might be desirable.

Lastly, we advocate studying the dynamics of the transition points where a regular game changes to a dilemma. Optimal performance with wealth inequality dropped severely when the dilemma was broken. The transition phase (here $0.1 < r < 0.3$) is interesting to investigate.

ACKNOWLEDGMENTS

This work was partially supported by FCT-Portugal (UIDB/50021/2020, PTDC/MAT-APL/6804/2020, and PTDC/CCI-INF/7366/2020). This work has also received funding from the European Union's H2020 program (grant 76595). F.P.S. acknowledges support from the James S. McDonnell Foundation Postdoctoral Fellowship Award.

REFERENCES

[1] Robert Axelrod. 1980. Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution* 24, 1 (1980), 3–25.
[2] Robert Axelrod and William Donald Hamilton. 1981. The evolution of cooperation. *Science* 211, 4489 (1981), 1390–1396.
[3] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Intell. Res.* 53 (2015), 659–697.

[4] Tilman Börgers and Rajiv Sarin. 1997. Learning through reinforcement and replicator dynamics. *J. Econ. Theory* 77, 1 (1997), 1–14.

[5] Kenneth S Chan, Stuart Mestelman, and R Andrew Muller. 2008. Voluntary provision of public goods. *Handbook of Exp. Econ. Res.* 1 (2008), 831–835.

[6] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998, 746-752 (1998), 2.

[7] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. 2009. The complexity of computing a Nash equilibrium. *SIAM J. Comput.* 39, 1 (2009), 195–259.

[8] Steven De Jong, Simon Uyttendaele, and Karl Tuyls. 2008. Learning to reach agreement in a continuous ultimatum game. *J. Artif. Intell. Res.* 33 (2008), 551–574.

[9] Elias Fernández Domingos, Jelena Grujić, Juan C Burguillo, Georg Kirchsteiger, Francisco C Santos, and Tom Lenaerts. 2020. Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization. *iScience* 23, 12 (2020), 101752.

[10] Elias Fernández Domingos, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts. 2021. Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning. *submitted* (2021).

[11] Ido Erev and Alvin E Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* (1998), 848–881.

[12] Ido Erev and Alvin E Roth. 2014. Maximization, learning, and economic behavior. *Proc Natl Acad Sci USA* 111 (2014), 10818–25.

[13] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with opponent-learning awareness. In *Proc. of the 17th Int. Conference on Autonomous Agents and MultiAgent Systems*. IFAAMS, 122–130.

[14] Drew Fudenberg, Fudenberg Drew, David K Levine, and David K Levine. 1998. *The theory of learning in games*. Vol. 2. MIT press.

[15] Christoph Hauert, Miranda Holmes, and Michael Doebeli. 2006. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proc. Roy. Soc. Lond. B* 273, 1600 (2006), 2565–2571.

[16] Oliver P Hauser, Christian Hilbe, Krishnendu Chatterjee, and Martin A Nowak. 2019. Social dilemmas among unequals. *Nature* 572, 7770 (2019), 524–527.

[17] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In *Int. Conf. on Machine Learning*. 1804–1813.

[18] Alexis Jacq, Julien Perolat, Matthieu Geist, and Olivier Pietquin. 2019. Foolproof Cooperative Learning. *ArXiv:1906.09831* (2019).

[19] D Katz, J Baptista, SP Azen, and MC Pike. 1978. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* (1978), 469–474.

[20] Peter Kollock. 1998. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology* 24, 1 (1998), 183–214.

[21] Shun Kurokawa and Yasuo Ihara. 2009. Emergence of cooperation in public goods games. *Proc. Roy. Soc. Lond. B* 276, 1660 (2009), 1379–1384.

[22] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proc. 16th Conf. on Autonomous Agents and MultiAgent Systems*. IFAAMS, 464–473.

[23] Carlton E Lemke and Joseph T Howson, Jr. 1964. Equilibrium points of bimatrix games. *J. Soc. Ind. Appl. Math.* 12, 2 (1964), 413–423.

[24] Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, and Thore Graepel. 2019. Emergent coordination through competition. *ArXiv:1902.07151* (2019).

[25] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*. 6379–6390.

[26] Michael W Macy and Andreas Flache. 2002. Learning dynamics in social dilemmas. *Proc Natl Acad Sci USA* 99, suppl 3 (2002), 7229–7236.

[27] Naoki Masuda and Mitsuhiro Nakamura. 2011. Numerical analysis of a reinforcement learning model with the dynamic aspiration level in the iterated Prisoner's dilemma. *Journal of Theoretical Biology* 278, 1 (2011), 55–62.

[28] Naoki Masuda and Hisashi Ohtsuki. 2009. A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. *Bull. Math. Biol.* 71, 8 (2009), 1818–1850.

[29] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. 2012. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *Knowl. Eng. Rev.* 27, 1 (2012), 1–31.

[30] Manfred Milinski, Ralf D Sommerfeld, Hans-Jürgen Krambeck, Floyd A Reed, and Jochem Marotzke. 2008. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proc Natl Acad Sci USA* 105, 7 (2008), 2291–2294.

[31] Pragnesh Jay Modi, Wei-Min Shen, Milind Tambe, and Makoto Yokoo. 2005. ADOPT: Asynchronous distributed constraint optimization with quality guarantees. *Artificial Intelligence* 161, 1-2 (2005), 149–180.

[32] Allen Newell and Paul S Rosenbloom. 1981. Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and Their Acquisition* 1, 1981 (1981), 1–55.

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

[34] Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. 2012. Game theory and multi-agent reinforcement learning. In *Reinforcement Learning*. Springer, 441–470.

[35] Shayegan Omidshafiei, Karl Tuyls, Wojciech M Czarnecki, Francisco C Santos, Mark Rowland, Jerome Connor, Daniel Hennes, Paul Muller, Julien Perolat, Bart De Vylder, et al. 2020. Navigating the Landscape of Multiplayer Games. *Nature Communications* 11, 5603 (2020).

[36] Jorge Peña and Georg Nöldeke. 2018. Group size effects in social evolution. *J Theor Biol* 457 (2018), 211–220.

[37] Ole Peters and Murray Gell-Mann. 2016. Evaluating gambles using dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26, 2 (2016), 023103.

[38] Luis G Quintas. 1989. A note on polymatrix games. *Int. J. Game Theory* 18, 3 (1989), 261–272.

[39] B Myerson Roger. 1991. Game theory: analysis of conflict. *The President and Fellows of Harvard College, USA* (1991).

[40] Alvin E Roth and Ido Erev. 1995. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games Econ. Behav.* 8, 1 (1995), 164–212.

[41] Francisco C Santos and Jorge M Pacheco. 2011. Risk of collective failure provides an escape from the tragedy of the commons. *Proc Natl Acad Sci USA* 108, 26 (2011), 10421–10425.

[42] Francisco C Santos, Marta D Santos, and Jorge M Pacheco. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454, 7201 (2008), 213.

[43] Fernando P Santos, Samuel F Mascarenhas, Francisco C Santos, Filipa Correia, Samuel Gomes, and Ana Paiva. 2019. Outcome-based Partner Selection in Collective Risk Dilemmas. In *Proc. 18th Int. Conference on Autonomous Agents and MultiAgent Systems*. IFAAMS, 1556–1564.

[44] Fernando P Santos, Jorge M Pacheco, Francisco C Santos, and Simon A Levin. 2021. Dynamics of informal risk sharing in collective index insurance. *Nature Sustainability* (2021), 1–7.

[45] Fernando P Santos, Francisco C Santos, Francisco S Melo, Ana Paiva, and Jorge M Pacheco. 2016. Dynamics of fairness in groups of autonomous learning agents. In *Int. Conf. Autonomous Agents and Multiagent Syst.* Springer, 107–126.

[46] Fernando P Santos, Francisco C Santos, Francisco S Melo, Ana Paiva, and Jorge M Pacheco. 2016. Multiplayer Ultimatum Game in Populations of Autonomous Agents. In *Adaptive and Learning Agents Workshop (ALA 2016), Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS 2016)*.

[47] Fernando P Santos, Francisco C Santos, Ana Paiva, and Jorge M Pacheco. 2015. Evolutionary dynamics of group fairness. *J Theor Biol* 378 (2015), 96–102.

[48] Yoav Shoham, Rob Powers, and Trond Grenager. 2007. If multi-agent learning is the answer, what is the question? *Artificial Intelligence* 171, 7 (2007), 365–377.

[49] Karl Sigmund. 2010. *The calculus of selfishness*. Princeton University Press.

[50] Brian Skyrms. 2010. *Signals: Evolution, learning, and information*. Oxford University Press.

[51] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[52] Attila Szolnoki and Matjaž Perc. 2011. Group-size effects on the evolution of cooperation in the spatial public goods game. *Phys Rev E* 84, 4 (2011), 047102.

[53] Alessandro Tavoni, Astrid Dannenberg, Giorgos Kallis, and Andreas L"oschel. 2011. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proc Natl Acad Sci USA* 108, 29 (2011), 11825–11829.

[54] Edward L Thorndike. 1898. Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review* 2, 4 (1898), i.

[55] Karl Tuyls and Ann Nowé. 2005. Evolutionary Game Theory and Multi-Agent Reinforcement Learning. *Knowl. Eng. Rev.* 20, 1 (March 2005), 63–90.

[56] Karl Tuyls and Simon Parsons. 2007. What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence* 171, 7 (2007), 406–416.

[57] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. 2003. A selection-mutation model for q-learning in multi-agent systems. In *Proc. 2nd Int. Joint Conference on Autonomous Agents and multiagent systems*. 693–700.

[58] Sven Van Segbroeck, Steven De Jong, Ann Nowé, Francisco C Santos, and Tom Lenaerts. 2010. Learning to coordinate in complex networks. *Adapt. Behav.* 18, 5 (2010), 416–427.

[59] Vitor V Vasconcelos, Francisco C Santos, and Jorge M Pacheco. 2013. A bottom-up institutional approach to cooperative governance of risky commons. *Nat. Clim. Change* 3, 9 (2013), 797–801.

[60] Vítor V Vasconcelos, Francisco C Santos, Jorge M Pacheco, and Simon A Levin. 2014. Climate policies under wealth inequality. *Proc Natl Acad Sci USA* 111, 6 (2014), 2212–2216.

[61] Jing Wang, Feng Fu, and Long Wang. 2010. Effects of heterogeneous wealth distribution on public cooperation with collective risk. *Phys Rev E* 82, 1 (2010), 016102.