# Improved Cooperation by Balance Exploration and Exploitation in Intertemporal Social Dilemma Tasks

Abstract: When individual behavior has rational characteristics, it may lead to irrational total benefits for the group. Humans and animals with many social traits tend to evolve a social trait of cooperation to meet this challenge. Therefore, cooperation between individuals is of great significance for social organisms to adapt to the changes of their natural environment. Based on multi-agent reinforcement learning, we propose a learning rate defined by the difference between the cumulative benefits of individual stages and the target benefits. This learning rate can adjust the strategies of agents by switching between exploration and exploitation according to the environmental benefits, so as to form an individual strategy with relatively better overall benefits for the group. The results show that this strategy has a relatively better overall strategy in the intertemporal social dilemma problem, and its strategy does not need to directly acquire the information of other individuals in the group. In particular, the heterogeneity of individual internal needs can help individuals better balance exploration and exploitation, so as to promote group cooperation.

## 1 Introduction

The results of group behavior of animals or humans are not only affected by the environment, but also affected by individual behavior strategies within the group. For example, the migration of animals or humans will be affected by the environmental resources they are in, and animals or humans will tend to migrate to places with abundant environmental resources. At the same time, the benefits obtained by individuals in the group from the environment are also affected by the strategies of other individuals in the group. When the strategy of all individuals in a group is to move to an area where resources are abundant, the gains of the individuals in that area will gradually decrease. The problem that individual rationality leads to group irrationality is the so-called intertemporal social dilemma problem [1]. In order to obtain the optimal collective reward in the situation of intertemporal social distress, the individual in the group needs to be able to make a trade-off between individual short-term gain and group long-term gain. However, it is not clear how individuals' strategies balance their short-term gains against the group's long-term gains.

Multi-agent reinforcement learning can simulate the behavior strategies of multiple agents in dynamic and variable environments. At present, many studies have found that multi-agent reinforcement learning can simulate how groups form cooperation, so as to obtain the optimal solution in various social dilemmas. These models often use internal rewards to make intelligent bodies rational strategies for the group. Intrinsic rewards include aversion to inequality [2], prosociality [3], and reputation [4]. These internal rewards discourage agents from adopting strategies that are detrimental to the group's overall benefits. For example, an agent abhors inequality, and when it finds that its own benefits are far greater than those of other agents, it discourages itself from pursuing a strategy that maximizes its individual benefits. In order to form internal rewards, these models all assume that an individual in a group can directly obtain information from other peers, thus forming an individual's prosocial attributes. However, it is difficult to provide a reasonable explanation for the formation of cooperation mechanisms in groups that can only indirectly obtain information from a small number of partners, such as fish and ants.

Based on Ericcharnov's marginal value theory, this study proposes that agents can simply adjust the learning rate to balance exploration and exploitation, so as to form cooperation in intertemporary social hardship tasks, thus obtaining higher total group benefits. In the reinforcement learning model, exploration is represented by agents choosing the actions that currently fail to obtain the optimal reward in order to avoid the local optimal solution, while exploitation refers to agents selecting the actions that currently receive the optimal reward. In order to make a trade-off between exploration and exploitation, we proposed a learning rate defined by the difference between an individual's cumulative reward and the target reward based on deep Q learning. This learning rate can adjust the strategies of agents by switching between exploration and exploitation according to the environmental benefits, so as to form an individual strategy with relatively better overall benefits for the group.

## 2 Related research

In the face of social dilemmas, how agents gradually form cooperative behavior has always been an important research problem in social science, economics and psychology. By constructing a strategy game in which two participants interact, Komorita and Parks [5] et al. (1995) found that by setting the benefits of two strategies of "worry" and "greed", the cooperative behavior of two participants can be formed. Evolutionary dynamics models have found that cooperation is promoted by a tit-for - tat strategy (Axelrod 1984 [6]), by cooperating with those who directly help one (Nowak 2006 [7]), or by punishing others in order to obtain sufficient rewards (Fehr and Gachter 2002 [8]). Although the above studies have given the possible factors of forming group cooperation, they have not given the specific strategies of individuals.

With the remarkable achievements of reinforcement learning in solving games such as Go [9] and multi-agent cooperative games [10, 11], many researchers have begun to use the multi-agent reinforcement learning model to study the mechanism of how groups form cooperation [12, 13]. By setting the decision task of the agent and the strategy parameters needed by the agent to form cooperation, the model is used to explain the possible strategy parameters of the animal or human when forming cooperation behavior. Sequeira et al. [14] proposed that agents form social attributes by exploring intrinsic motivations. Foerster et al. [15] (2017) made agents form cooperation in the multi-round Prisoner's Dilemma game by modeling the learning results of other individuals. Peysakhovich et al. [16] (2018) found that when agents pay more attention to other individual benefits, they can form prosocial strategies in Stag Hunt Games. Hughes [12] et al. (2018) integrate the aversion to inequality into the internal reward of the agent, so as to adjust the strategy when its own benefit is much larger than that of other individuals in the group or its own benefit is much smaller than that of other individuals in the group, thus forming cooperation. Jaques [17] et al. (2018) converted the influence of individual actions on the group into internal rewards to form cooperation among agents in social dilemma. Wang [18] et al. (2019) proposed evolutionary deep reinforcement learning, which defined past and future rewards of other individuals as internal rewards of agents to evolve cooperative strategies. Khadka et al. [19] (2019) designed a method for learning multiple strategies with shared playback buffers and dynamically selecting the best learner to evolve cooperation between multiple agents. Badjatiya [13] et al. (2020) proposed to design a status-quo loss function to make agents follow the Status Quo as far as possible, so as to evolve cooperative behavior in a social dilemma environment. McKee et al. [20] (2020) sample their rewards from groups with heterogeneous characteristics, allowing agents to acquire prosocial attributes. Danassis [21] et al. (2021) found that agents can improve their cooperative behavior by integrating common signals (such as periodic numbers such as time and date) during learning. A common feature of the above models is that agents need to directly obtain relevant information of all other agents in order to form cooperative behaviors. These models do not provide a reasonable explanation for the formation of cooperation mechanisms in groups that can only indirectly obtain information from a small part of the group, such as fish and ants.

Due to the dynamic changes and uncertainties of the environmental state, the agents either use the existing experience for exploitation or take the risk of not being able to get better rewards for exploration in the hope of getting better strategies. Therefore, exploration and exploitation have always been important research topics in reinforcement learning. In the early days of solving the multi-armed gambling machine problem, Epsilon-greedy [22], Upper confidence bounds [23] and Boltzmann exploration [24] can be used in exploration and exploitation to get the best overall benefit. However, in the real environment, due to the sparse nature of reward signals and the abnormal noise in the environment state, the above simple exploration strategy cannot obtain good overall reward.

A more general method is to design an internal reward function to form the internal motivation of the agent [25], so as to guide the agent to explore through such as curiosity. Curiosity includes discovering new states, or improving the accuracy of agent's estimation of environmental changes [22], etc. This kind of exploration strategy based on internal reward may have problems such as slow convergence speed and non-stationary exploration reward, which makes it difficult to form a fixed exploration strategy. Therefore, the memory-based exploration strategy [26] and the resampling Q-value exploration strategy [26] were developed to avoid the shortcomings of the exploration strategy based on internal rewards. However, the above single agent exploration strategy is not necessarily suitable for multi-agent cooperative exploration.

In the case of multi-agent exploration, it is not only necessary to encourage agents to explore new states and deal with the problem of sparse reward signals, but also to cooperate with the actions of agents to form cooperation to explore the environment. Agogino and Tumer [27] define a method for evaluating the effectiveness of a reward function for multiple agents in a smaller-scale state space. Jaques [17] et al. defined an intrinsic reward function for multi-agent reinforcement learning, which encourages agents to take actions that have the greatest impact on the behaviors of other agents, so as to obtain cooperative exploration strategies. Mahajan et al. [28] introduced a mechanism for implementing 'commitment' exploration, allowing agents to explore common strategies for temporary expansion. Wang et al. [18] defined impact-based rewards that encourage agents to visit areas where their behavior affects the transformation and reward of other agents. Recently, Iqbal and Sha (2021) [29] proposed a kind of exploration method based on internal reward. The main feature of this method is that it can coordinate the exploration strategies among agents and enable agents to better obtain overall benefits. The above multi-agent exploration strategy still requires the agent to obtain the information of other agents by lipolysis. Further research is needed for the exploration strategy under the condition that only a small part of other agents can be obtained, or even the information of other agents is not required.

## 3 Multi-agent reinforcement learning and decision making tasks

### 3.1 Multi-agent reinforcement learning

We define the multi-agent reinforcement learning model as a quad, which includes the state set , the state transfer function , the action set and the reward , i.e. . There are agents in the environment, and the state that each agent can perceive is ，it means that the can observe dimensions of the state. That is, the agent can only partially observe its state. Each agent in the environment interacts with the environment through its actions *An*, and the actions of the agent will cause changes in the state of the environment. The change is described by the state transition function: . That is, the actions of all agents in the environment work together to change the state of the environment from to another state .

Each agent learns the strategy according to its observation . After the agent executes the action , it will get the reward , and evaluate the result of the action through the reward. The goal of the agent is to learn an optimization strategy in order to obtain the greatest long-term benefits. The definition of long-term benefits of an agent as：

Formula 1.

Among them, is a discount factor between 0 and 1. For simplicity, . For agent , in order to obtain the maximum expected reward, the function can be updated according to the following function [30].

Formula 2.

### 3.2 Learning rate based on target benefit

We consider to define the learning rate through the staged cumulative benefits and target benefits of the agent. The learning rate reflects the impact of changes in the environment on the agent's strategy. In order to achieve this goal, we define the difference between the stage cumulative reward and the target reward as the learning rate：

Formula 3.

is a constant, and the size is set to 0.001. The phase cumulative reward is the cumulative reward value of the agent in time , which reflects the indirect influence of other agents on the individual's reward in a certain period of time. The target reward is a fixed value, and each agent has a target reward, which reflects the degree of satisfaction of the agent. When the target profit is large, it means that the agent needs to obtain more accumulated profit to be satisfied. If the cumulative benefit of the agent's stage is less than the target benefit, it indicates that the goal of the agent has not been reached, and it shows more exploration. When the accumulated reward of the agent's stage is close to the target reward, it indicates that the agent's strategy has reached its expectations, and it shows more exploitation.

According to the above definition of learning rate, when the environment is in a stable state, the agent's strategy gradually converges, and its learning rate is at a low level. When there is a sudden change in the environmental state, the agent's strategy must be able to adapt to this change quickly, and the learning rate during this period is at a relatively high level.

### 3.3 Decision task

According to the resource collection task of Hughes [12] et al. (2018), we designed a similar intertemporal social dilemma task. The task environment contains two resource areas with different values, namely the apple area and the garbage area. The size of the environmental map is units, garbage is distributed in the upper half of the environment, and apples are distributed in the lower half of the environment. Garbage appears in its area with probability , and the amount of garbage in the environment is recorded as . Apples appear with probability in the area where they are located, and the number of apples in the environment is recorded as . Apple’s growth rate is negatively correlated with the amount of garbage:

Formula 4.

Where is the maximum growth rate of apples, and is half of the garbage area in the map.

There are a number of agents distributed in the environment, and agents move in the environment to either reap rewards at their location or clean up the garbage at their location. The reward of the agent harvesting apples is recorded as , and the reward of cleaning garbage is recorded as . The purpose of the agent is to obtain the most collective reward. In this task, each agent can only perceive the information in the limited field of view around it, and the field of view of the agent is denoted as .

The dilemma of this decision task is as follows, the growth of apples and junk affects each other. Since the reward of apples is greater than the reward of garbage, the agent's individual strategy will tend to pick apples rather than clean up the garbage. However, the decrease in the number of apples will lead to an increase in the amount of garbage, thereby suppressing the probability of apples appearing. Therefore, for the agent group, it is necessary for some agents to clean up the garbage and some agents to collect apples in order to obtain more collective rewards as a whole.

Since the growth of garbage and apples is not balanced, the multi-agent cleans up all the garbage in the field of vision when cleaning up the garbage . When collecting apples, only the apples at the current location are collected. Therefore, the actual reward obtained by the agent for cleaning up garbage is . The actual reward for collecting apples is . Set in the decision task, which represents the collective reward of the agent in the time span .

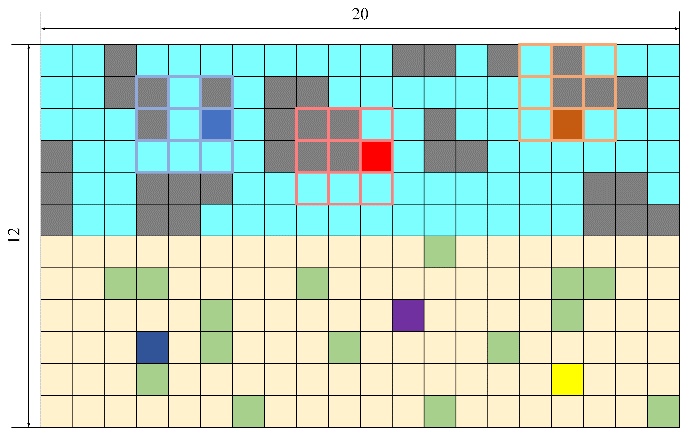


Figure1. Game map

In the simulation, agents are randomly placed in the environment, and their respective learning functions are shown in formula 1. We map the location of each cell in the map to the interval [-120,120]. Where [0,120] represents the cell in the garbage area, and [-1,-120] represents the cell in the apple area. represents the initial position of each agent in the environment, and . For each agent, each episode includes 100 trials, and each group of experiments includes 300 rounds of episodes.

### 3.4 Homogeneous and heterogeneous group attributes

According to the value method of , the agent group is divided into heterogeneity and homogeneity. Heterogeneity means that takes a random value within a given range, and this heterogeneity reflects the diversity of individual agents. When the target reward of the agent satisfies , it is called the high target rewarder. When the target reward of the agent satisfies , it is called the low target rewarder. Agents in these two cases have homogeneous group attributes. The parameter settings related to the group attributes in the simulation are shown in the table 1.

Table 1. Group attribute parameter settings

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| Heterogeneous |  |  |  | 5 | 10 | 5 |  |
| Homogeneous |  |  |
| 0 |  |

## 4 Result

We use the stage cumulative reward and the target reward to define the dynamic learning rate . It is verified that the learning rate can form group cooperation by balancing exploration and exploitation in inter-period social dilemma tasks, so as to obtain relatively good overall benefits.

As shown in Figure 2, we compare the benefits of groups performing inter-period social dilemma tasks under a fixed learning rate and a dynamic learning rate. The group collective reward under the dynamic learning rate can converge to between 2200 and 2500. However, the collective reward of a group with a fixed learning rate (=0.001) can only converge to between 1300 and 1600. The agent only uses random strategies, and the collective reward converges between 700 and 1000.

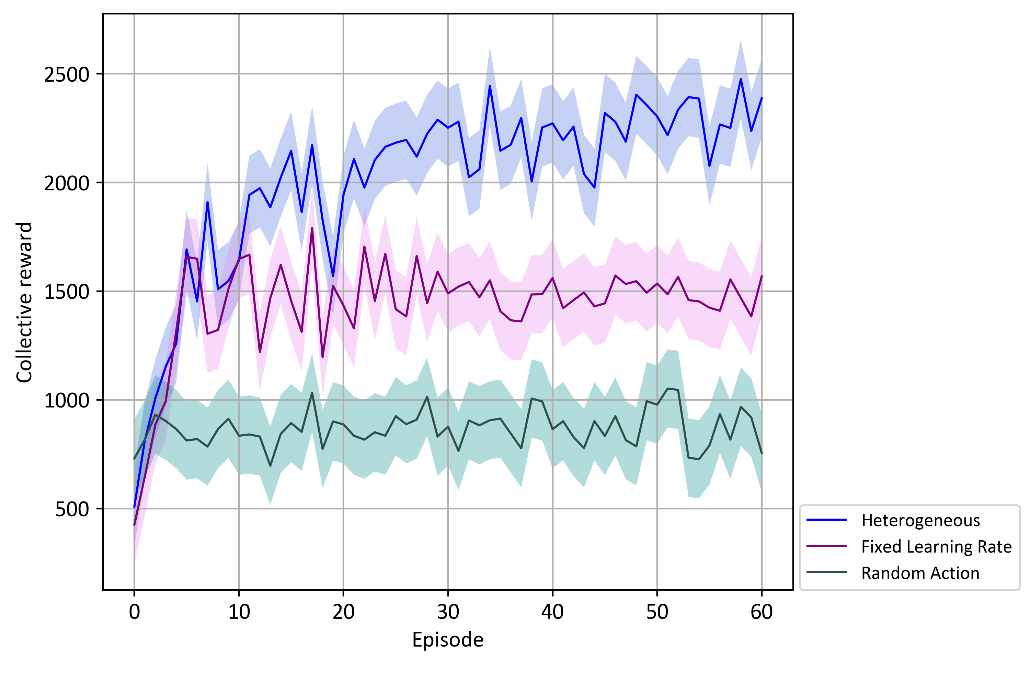


Figure 2. Comparison of benefits between dynamic learning rate and fixed learning rate

We set different distributions of target rewards to verify the role of target revenue in weighing exploration and exploitation. According to the distribution of target rewards, we defined three group attributes: Heterogeneous, Homogeneous High and Homogeneous Low. As shown in Figure 3, the collective reward of the Heterogeneous group in these three groups is higher than the collective reward of the two Homogeneous groups. The collective reward of the Homogeneous Low group is the lowest, even lower than the collective reward of the random strategy.

In order to show the changes of exploration and exploitation of each agent in the environment more clearly, we draw the active position of each agent in the environment in Figure 4. Heterogeneous groups conduct Exploitation according to their own target rewards to form a division of labor (Figure 4A). This kind of division allows some agents (agents with low goal rewards) to collect garbage in the garbage area, and some agents (agents with high goal rewards) collect apples in the apple growth area. It is this distribution that leads to the highest collective reward for the Heterogeneous group. When the target reward of each agent in the group is high, they have been using exploration in the environment to obtain high-yield apples (Figure 4B). And when the target reward of each agent in the group is low, their reward in the apple area is greater than the individual's own expectations, and the model will inhibit low target rewarders from greedily collecting apples. They will update their strategy when they explore the area that matches their own target reward (junk area) (Figure 4C). The group that chooses actions completely at random has been carrying out a goalless exploration (Figure 4D).

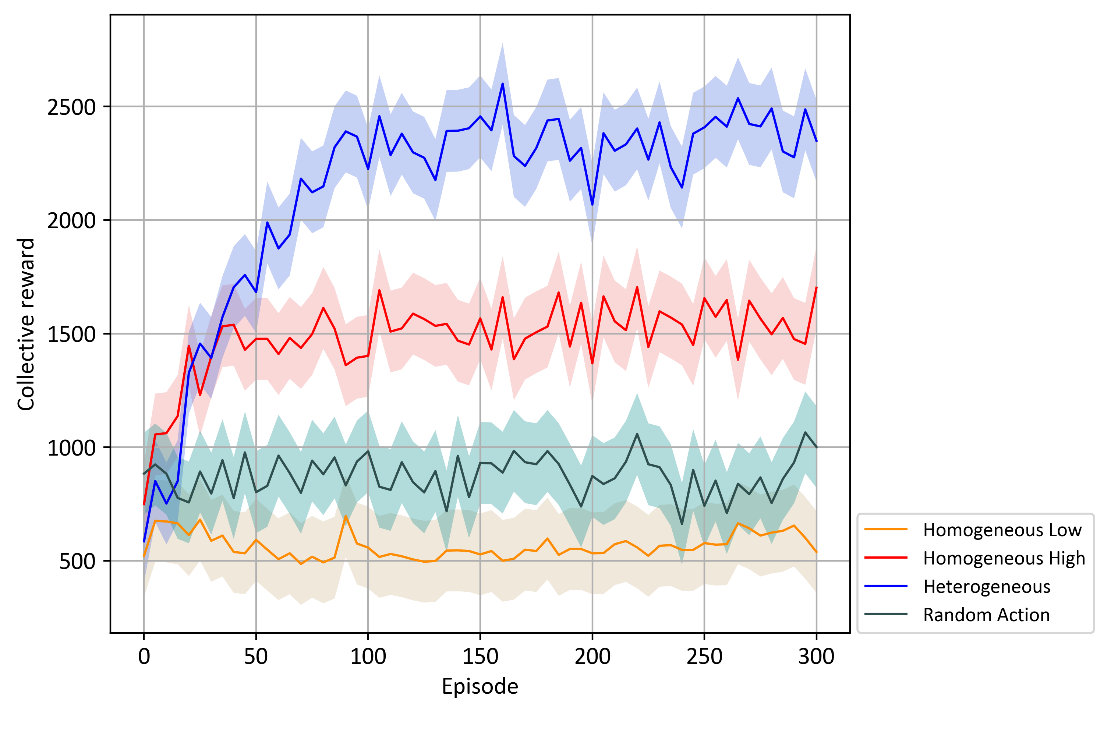


Figure 3. Comparison of collective rewards between heterogeneous groups and homogeneous groups

1. Heterogeneous
2. Random Action
3. Homogeneous Low



1. Homogeneous High

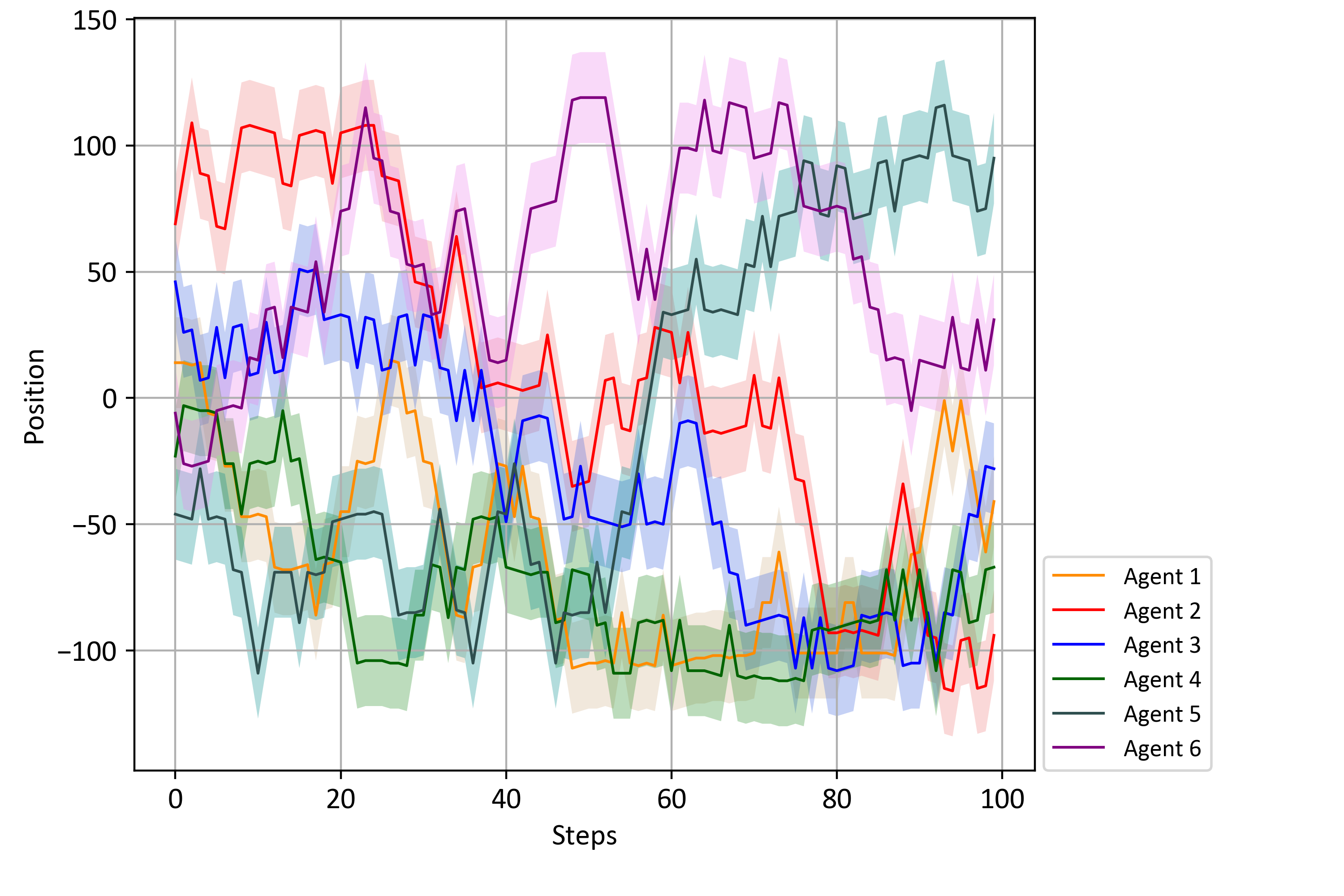
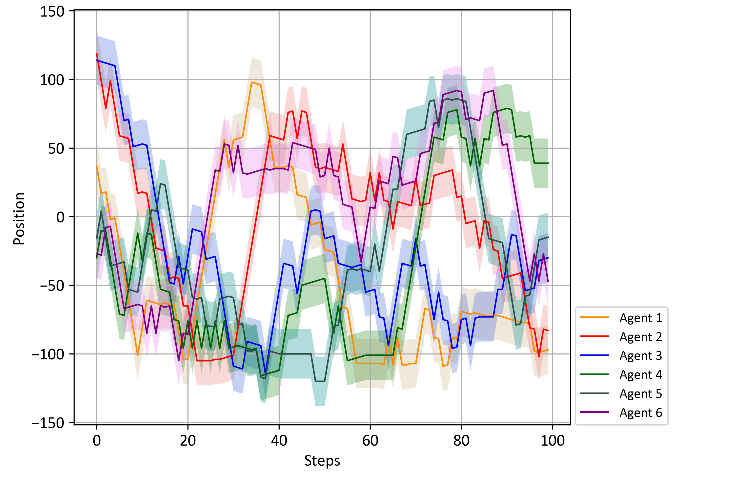


Figure 4. Comparison of the activity position of agents in a heterogeneous group and a homogeneous group

1. Homogeneous High
2. Random Action
3. Heterogeneous
4. Homogeneous Low

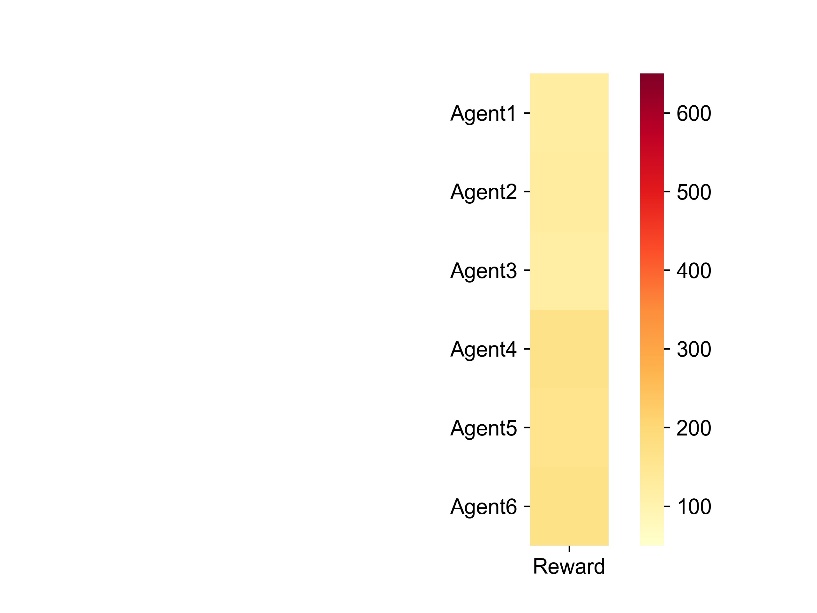
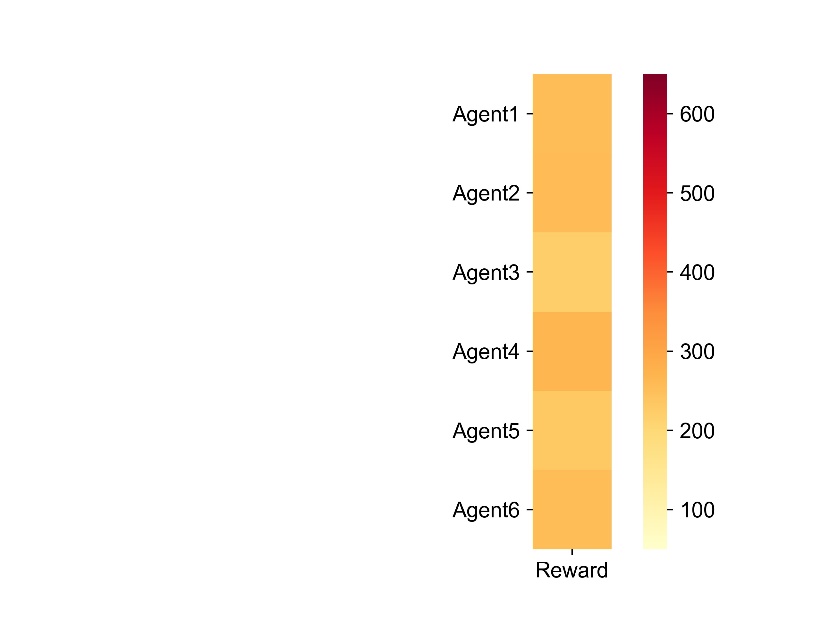
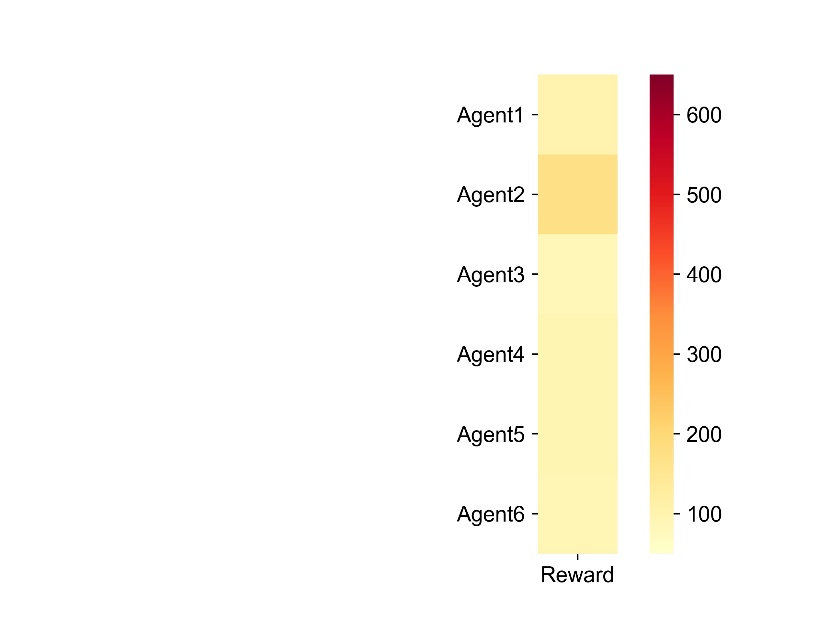
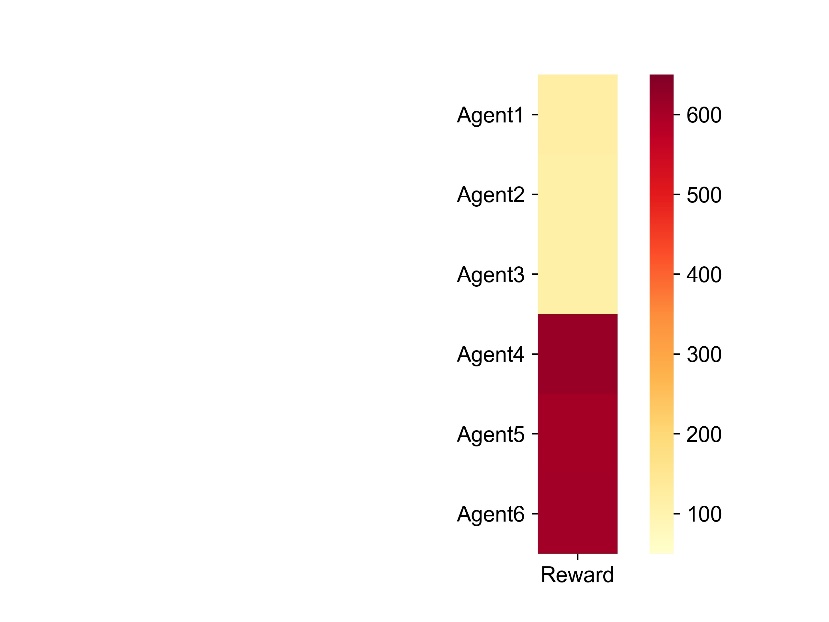


Figure 5. Comparison of personal rewards of agents in different groups

As shown in Figure 5, the total reward between each agent in different groups is compared in order to compare whether the cooperative behavior causes the gap between the rich and the poor. The results show that compared with the total reward difference within the homogeneous group, the reward difference between the heterogeneous groups is the largest. This shows that although the heterogeneity of the group promotes the cooperative behavior between agents, it can also cause the gap between the rich and the poor within the group to increase.

We use the Heterogeneous group as the experimental object separately, and control the target reward of each agent unchanged to explore the influence of the cumulative length of different stages on the group cooperation behavior. As shown in Figure 6, with the continuous increase of , the collective reward of the group continues to decrease; when the value of is too small, the collective reward of the agent is also small. When the value of is large, it is difficult for the agent to perceive changes in the environment, thereby reducing the possibility of its use of exploration. When the value of is small, the agent only pays attention to the current reward, and it is also difficult to perceive changes in the environment, which makes the agent frequently adopt exploration.

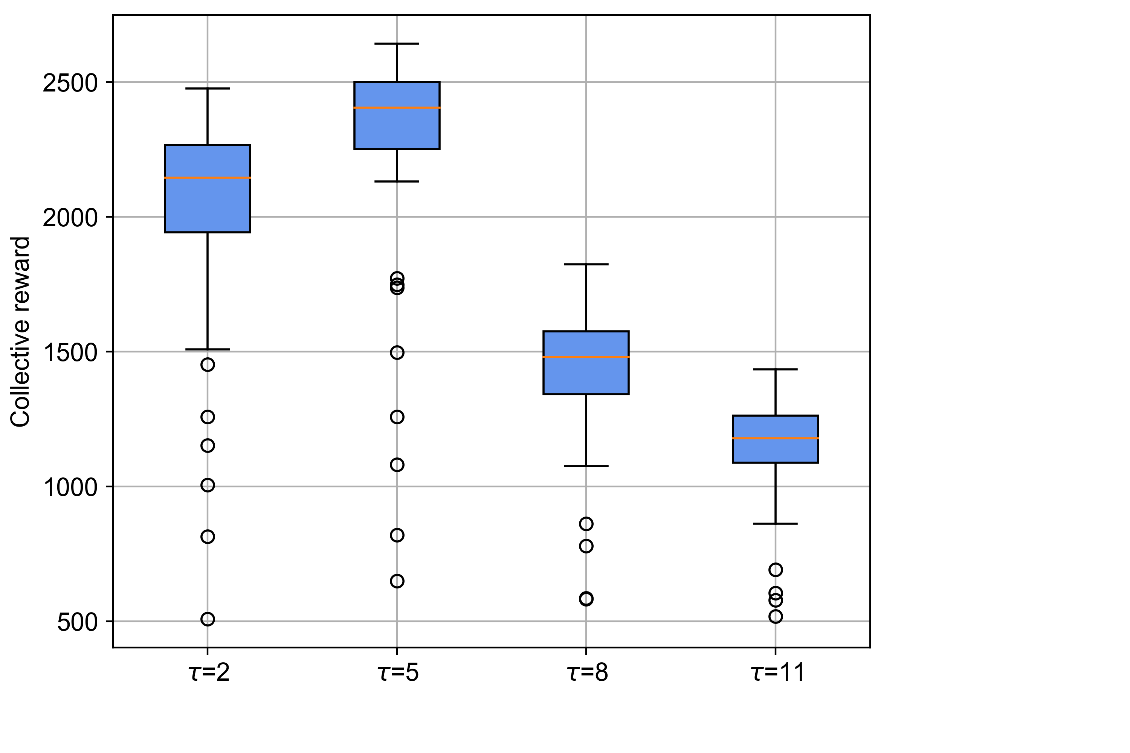


Figure 6. Collective rewards of cumulative length τ in different stages

## 5 Discussion

边际价值理论提出人对物品的欲望会随其不断被满足而递减。基于边际价值理论我们提出了一个通过平衡利用-探索，从而在跨期社会困境中多智能间形成合作的方法。面对跨期社会困境问题，智能体只需要计算在一段时期内的收益，而不需要从其他智能体获取额外的信息，就可以形成类似合作的行为。我们的结果表明，智能体间的异质性，如各个智能体的目标收益的异质性，是形成合作行为的关键。这个结论与McAvoy等（2020）和McKee等（2020）[31]最近的研究结果类似，他们发现智能体连接数量的异质性和社会偏好的异质性，能促进智能体社会行为的形成。我们的研究结果进一步揭示，其它类型的异质性，如本文设定的各个智能体目标需求的异质性，也能促进智能体间合作行为的形成。在此基础上，我们推断也许还存在其它可以促进智能体形成合作行为的异质性参数，这一点值得后续更为深入的理论与实验研究。

我们研究的一个主要不足之处在于，设计的智能体学习算法并没有考虑智能体间直接的互动和交互。智能体间通过一个时间段内集体奖励来产生相互的影响。对某个智能体而言，如果群体内的其它智能体行为都是非社会性，那么这个智能体一个时间段内集体奖励就会变少，这会引起该智能体根据其自身的目标收益来调整利用和探索策略。也就是说，其它智能体的非社会性行为间接引起了该智能体策略的变化。相似地，如果群体内的其它智能体行为都是亲社会性，他们的行为也会间接引起了该智能体利用和探索策略的变化。我们将在后续研究考虑，在算法中增加智能体间交互的参数。

需要特别指出，我们只是在跨期社会困境中验证了智能体通过调整利用与探索策略来形成合作，对于其它类型的社会困境问题能否得到同样的结论还需要进一步的验证。

## References

[1] E. Hughes *et al.*, "Inequity aversion improves cooperation in intertemporal social dilemmas," *arXiv preprint arXiv:1803.08884,* 2018.

[2] D. Engelmann and M. Strobel, "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments," *American economic review,* vol. 94, no. 4, pp. 857-869, 2004.

[3] A. Peysakhovich and A. Lerer, "Prosocial learning agents solve generalized stag hunts better than selfish ones," *arXiv preprint arXiv:1709.02865,* 2017.

[4] J. A. Cuesta, C. Gracia-Lázaro, A. Ferrer, Y. Moreno, and A. Sánchez, "Reputation drives cooperative behaviour and network formation in human groups," *Scientific reports,* vol. 5, no. 1, pp. 1-6, 2015.

[5] Komorita, Samuel, S., Parks, Craig, and D., "Interpersonal relations: Mixed-motive interaction," *Annual Review of Psychology,* 1995.

[6] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Quarterly Review of Biology,* vol. 79, no. 2, pp. 135-160, 1981.

[7] M. Nowak, "Five Rules for the Evolution of Cooperation," *Science,* vol. 314, no. 5805, pp. 1560-1563, 2006.

[8] S. Gaechter and E. Fehr, "Altruistic Punishment in Humans," *Nature,* 1997.

[9] D. Silver *et al.*, "Mastering the game of go without human knowledge," *nature,* vol. 550, no. 7676, pp. 354-359, 2017.

[10] N. Jaques *et al.*, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International Conference on Machine Learning*, 2019: PMLR, pp. 3040-3049.

[11] Y. Du, L. Han, M. Fang, J. Liu, T. Dai, and D. Tao, "Liir: Learning individual intrinsic reward in multi-agent reinforcement learning," *Advances in Neural Information Processing Systems,* vol. 32, pp. 4403-4414, 2019.

[12] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, and T. G Ra Epel, "Inequity aversion improves cooperation in intertemporal social dilemmas," in *The 32nd Conference on Neural Information Processing Systems (NIPS)*, 2018.

[13] P. Badjatiya, M. Sarkar, A. Sinha, S. Singh, and B. Krishnamurthy, "Inducing Cooperation in Multi-Agent Games Through Status-Quo Loss," 2020.

[14] P. Sequeira, F. S. Melo, P. Rui, and A. Paiva, "Emerging social awareness: Exploring intrinsic motivation in multiagent learning," in *Development and Learning (ICDL), 2011 IEEE International Conference on*, 2011.

[15] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual Multi-Agent Policy Gradients," 2017.

[16] A. Peysakhovich and A. Lerer, "Prosocial learning agents solve generalized Stag Hunts better than selfish ones," 2017.

[17] N. Jaques *et al.*, "Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning," 2018.

[18] J. X. Wang, E. Hughes, C. Fernando, W. M. Czarnecki, and J. Z. Leibo, "Evolving intrinsic motivations for altruistic behavior," 2018.

[19] S. Khadka, S. Majumdar, T. Nassar, Z. Dwiel, and K. Tumer, "Collaborative Evolutionary Reinforcement Learning," 2019.

[20] K. R. McKee, I. Gemp, B. McWilliams, E. A. Duéñez-Guzmán, E. Hughes, and J. Z. Leibo, "Social diversity and social preferences in mixed-motive reinforcement learning," *arXiv preprint arXiv:2002.02325,* 2020.

[21] P. Danassis, Z. D. Erden, and B. Faltings, "Improved Cooperation by Exploiting a Common Signal," *arXiv preprint arXiv:2102.02304,* 2021.

[22] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-agent reinforcement learning in sequential social dilemmas," *arXiv preprint arXiv:1702.03037,* 2017.

[23] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research,* vol. 3, no. Nov, pp. 397-422, 2002.

[24] M. Edman and N. Dhir, "Boltzmann Exploration Expectation-Maximisation," *arXiv preprint arXiv:1912.08869,* 2019.

[25] A. Zgonnikov and I. Lubashevsky, "Intrinsic motivation and learning dynamics," Citeseer, 2013.

[26] L. Weng, "Exploration strategies in deep reinforcement learning," *Online:* [*https://lilianweng*](https://lilianweng)*. github. io/lil-log/2020/06/07/exploration-strategies-in-deep-reinforcement-learning. html,* 2020.

[27] A. K. Agogino and K. Tumer, "Analyzing and visualizing multiagent rewards in dynamic and stochastic domains," *Autonomous Agents and Multi-Agent Systems,* vol. 17, no. 2, pp. 320-338, 2008.

[28] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson, "MAVEN: Multi-Agent Variational Exploration," 2019.

[29] S. Iqbal and F. Sha, "Actor-Attention-Critic for Multi-Agent Reinforcement Learning," 2018.

[30] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning,* vol. 8, no. 3-4, pp. 279-292, 1992.

[31] A. Mcavoy, B. Allen, and M. A. Nowak, "Social goods dilemmas in heterogeneous societies," *Nature Human Behaviour,* vol. 4, no. 8, pp. 1-13, 2020.