Reaching Cooperation using Emerging Empathy and Counter-empathy

Jize Chen
Harbin Institute of Technology
Harbin, China
cjz74007@163.com

Changhong Wang
Harbin Institute of Technology
Harbin, China
cwang@hit.edu.cn

ABSTRACT

According to social neuropsychology, the cooperative behavior is largely influenced by empathy, which is deemed essential of emotional system and has wide impact on social interaction. In the work reported here, we believe that the emergence of empathy and counter-empathy is closely related to creatures' inertial impression on intragroup coexistence and competition. Based on this assumption, we establish a unified model of empathy and counter-empathy in light of Hebb's rule. We also present Adaptive Empathetic Learner (AEL), a training method for agents to enable affective utility evaluation and learning procedure in multi-agent system. In AEL, the empathy model is integrated into the adversarial bandit setting in order to achieve a high degree of versatility. Our algorithm is first verified in the survival game, which is designed to simulate the primitive hunting environment. In this game, empathy and cooperation emerge among agents with different power. In another test about Iterated Prisoners' Dilemma, cooperation was reached even between an AEL agent and a rational one. Moreover, when confronted with hostile, the AEL agent showed sufficient goodwill and vigilantly protected its safe payoffs. In the Ultimatum Game, it's worth mentioning that absolute fairness could be achieved on account of the self-adaptation of empathy and counter-empathy.

CCS CONCEPTS

• **Theory of computation → Multi-agent learning**; *Convergence and learning in games*; *Computational advertising theory*; • **Computing methodologies → Cooperation and coordination**; *Cognitive science*; *Multi-agent systems*;

KEYWORDS

cooperation; empathy and counter-empathy; multi-agent system; adversarial bandit

ACM Reference Format:
Jize Chen and Changhong Wang. 2019. Reaching Cooperation using Emerging Empathy and Counter-empathy. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 8 pages.

1 INTRODUCTION

It's a great ideal to make the machine more human-like on both actions and thoughts. Yet while the research goes deeper, wide concerns arise about machine ethics [8, 13]. Excessive of imitation, especially the general principle that maximizing the self-profits, may lead the machine into immorality with high probability. In order to make machines interact safely and properly in multi-agent environment or human-agent environment, we should unearth some well-meaning moral factors and introduce them into agents' inner attributes. Thus they can regulate their behavior and adopt more positive strategies to handle complex situations [11, 12].

In previous studies, multi-agent learning algorithms that collect and process overmuch observation information, such as Nash Q-learning [16], or that observe ones' own information in the independent mode, such as [5, 18, 26, 27], are mostly confined to the setting that maximizing one's own profits. In this case, algorithms incarnate rationality and conform to the wish that converging to the Nash equilibrium. But even so, the equilibrium result may not be the optimal distribution, or even any pareto-optimal states [4]. This pervasive problem can be found in lots of circumstances such as Prisoners' Dilemma and the Ultimatum Game.

We seek to find a more general and fundamental method for agents to promote proper interaction in different situations, especially containing a dilemma. Some previous studies tried to design new learning structure. Works in [30] proposed a novel idea of multi-agent learning algorithm that seeks the best-response strategy instead of finding an equilibrium solution. However, the selection of the best-response strategies is difficult to implement. [3] introduced the Conditional Joint Action Learner (CJAL) which learns strategies that converge to a Pareto-Optimal outcome, while the apply is restricted in situations with no Pareto improvement. Most other works focused on modeling emotions or affective functions such as guilt and forgiveness, or by modeling social fairness to achieve prosocial behavior [19, 24, 31]. But as far as our best knowledge, those models are not able to be generalized to a more general case.

Inspired by the research of social neuropsychology, we introduce empathy mechanism into the process of agents' learning. Psychology studies have shown that empathy is an ability that deeply influences human emotions and social interactions. Through empathy, individuals can experience similar emotional feelings to others, generate sympathy for the situation of others, and further generate motivation to alleviate the suffering of others. This in turn induces individual prosocial behavior and suppresses offensive behavior [10]. By contrast, counter-empathy is individuals' emotional response opposite to the emotion of the observed person. Research suggests that counter-empathy is associated with envy and gloating [15]. It seems to be antisocial, nevertheless, counter-empathy plays a essential role in the evolution of organisms and can increase the competitive consciousness of individuals. Empathy and counter-empathy play a part in almost every moment of individuals' growth

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

and even the evolution of species. Therefore, it will be extremely meaningful to have agents with this two basic features.

We first analyzed how empathy and counter-empathy emerge in the interaction and gave an attempt to explain the incentive mechanism under some specific environments. Then, inspired by Hebb's rule, we established a general unified model of empathy and counter-empathy in a natural way. Finally, combined with extended structure of adversarial bandits, we designed Adaptive Empathetic Learner (AEL) which can adjust the degree of empathy adaptively according to the environment. In the survival game and Iterated Prisoners' Dilemma, AEL showed increased cooperation. Cooperative behavior, especially absolute fairness was also detected when agents were transferred into the Ultimatum Game.

2 BACKGROUND

2.1 Extended structure of adversarial bandits

The adversarial bandit problem is closely related to the problem of l earning to play an unknown N-person finite game, where the same game is played repeatedly by N agents [1]. In the traditional $K-$armed adversarial bandits, there is an arbitrary sequence of reward vectors $v = (r_1, \cdots, r_n)$ where $r_t \in [0,1]^K$ for each $t \in [n]$. In each round, the agent chooses an action $A_t \in [K]$ and observes the reward $X_t = r_{tA_t}$. The agent's value is reflected in the maximizing of total reward $S_n = \sum_{t=1}^{n} X_t$ (basic value), and the forming of complete value function is then directly guided by the basic rule and the assessment (reward) from the environment.

In this setting, driven by the basic value, agents will make a decision that most benefits to their own long-term external reward and psychological factors are rarely considered. Although the traditional adversarial bandits assumes an external critic (to provide the reward signal), this actually happens inside the brain of real-world organisms [22]. What we expect is to endow the feedback of environment with more psychological guidance.

Thus, a extended structure of adversarial bandits was proposed. In this structure, feedback of actual stimulus for the training of value function can be divided into two parts. As Fig.1 shows, the first part gets the sensation from the external environment while the next part maps the sensation to the psychological reward according to the value model.

2.2 Conditioned reflex and Hebb's rule

Conditioned reflex refers to a learning procedure in which a temporary neural connection is established between an external stimulus and an organism's response. It was first studied in detail by Ivan Pavlov through experiments with dogs [23]. In his experiment, each time before the dog is fed, bell will ring. As time passes, the dog will make a link between the ring-tone and the food in its nervous system.

Inspired by this experiment, Hebb proposed rules for changes in the strength of neuronal connections [14], which can be expressed as:

(1) If two neurons activate simultaneously, the synaptic link between them strengthens.
(2) If two neurons activate asynchronously, the synaptic link between them weakens.

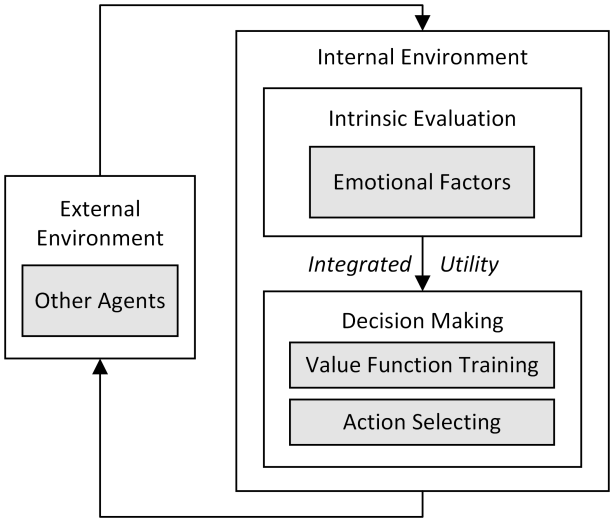

Figure 1: Schematic representation of dynamic interaction between external environment and internal environment based on [22].

This type of learning procedure is a common phenomena in nature, which is generally called positive feedback. In some researches on this aspect, positive feedback is regarded as the reason for the development and evolution of things [21]. This also means that it is highly probable that similar evolutionary patterns will emerge in different hierarchies (living or non-living). We will discuss this further in the next section.

3 A GENERAL MODEL OF EMERGING EMPATHY AND COUNTER-EMPATHY

3.1 How does empathy emerge?

Considering to the surviving way of human-beings in some primitive conditions, we hypothesize that the generation of empathy is closely related to individuals' strong dependence on group. Taking the hunting activity as an example, because of the limited ability of individuals, the odds of success when hunting alone is much smaller than that of group. Therefore, driven by the inherent rational interests, individuals tend to choose hunting in group. This will prompt the group to present a phenomenon of coexistence, which means one success leads to another and so does failure. Long-term symbiotic stimuli will contribute to the emergence of empathy, which in turn lead to mutual help and other well-meaning behavior.

This process is also evident in other group-living animals, such as the "more eyes" phenomenon existed in birds [25] and monkeys [7]. In these groups, the individual's alarm for danger is shared with others. Thus, although less individuals' time is spent on vigilance, danger can be more quickly alerted to all [17].

In the view of that above consideration, we believe that empathy is the internal solidification of coexistence to enhance this kind of positive correlations [20]. The emerging empathy can further consolidate individuals' dependence on group, making individuals more trusting, the community more stable. What's more, the generalized empathy towards similar creature is also beneficial to the

expansion and replacement of the community, which is a virtuous circle for the symbiotic community in the early stage.

On the other hand, if the community continues to expand, the individual difference will lead to an imbalance in the allocation of internal resources, which in turn presents a competitive situation. In this case, the individual's income is negatively correlated with others', which is mapped into counter-empathy in the interior, further enhancing the individual's sense of competition.

3.2 How to describe empathy?

Based on the assumption above, we can infer that the situation affects the trend of income, and the relationship of income directly determine the property of empathy. Individuals tend to follow individuals with positive correlations and reject ones with negative correlations.

Moreover, comparing the emergence of empathy with the Hebb's rule mentioned in background, we can get the approximate corresponding relationship as shown in Fig.2. The activity of cells can be compared to the state of gaining payoffs and the demand of simultaneous activity can be translated to the meaning that dynamic payoffs of two agents should have similarity. From this, we can deduct that degree of empathy should increase if payoffs of the two agents is correlated and similar, and decrease on the contrary.

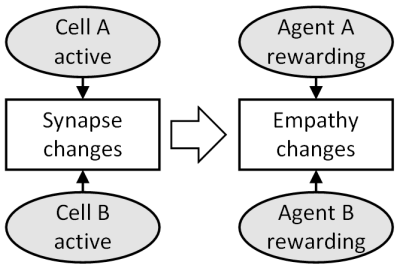

Figure 2: Analogy of neuronal link changes and empathy changes.

Note that the Hebb's rule mentioned here is not to emphasize the inevitable causal relationship between synapse and empathy, but as an inspiration to illustrate the possibility of similar dynamic characteristics in different structural levels. Thus, we give a simple definition of empathy according to the payoffs agents receive.

Definition 1: For an environment with N agents, $i, j \in N$, if agent i gets a reward vector $X_{i,t} = (r_{i,t-w_e}, \cdots, r_{i,t})$ and agent j gets $X_{j,t} = (r_{j,t-w_e}, \cdots, r_{j,t})$, where $r \in [0,1]$,$w_e \in \mathbb{N}$, the empathy between agent i and j is

$$\lambda_{i,j,t} = 1 - \frac{2}{\sqrt{w_e + 1}} \|X_{i,t} - X_{j,t}\| \tag{1}$$

In our model, w_e stands for the length of the memory window, which means empathy will have the property of memory if w_e increases. The range of empathy between agent i and j is $[-1, 1]$. As $\lambda_{i,j,t}$ changes, we can get empathy when $\lambda_{i,j,t}$ is positive, and counter-empathy when negative, and two extreme modes– full antagonism and full cooperation will be achieved when $\lambda_{i,j,t} = -1$ *or* 1. The empathy model mainly considers the positive correlation between the similarity of income and empathy. It seems restrictive that this model only gives a metric for the co-occurrence of similar rewards. However, some potential capabilities can be recognized if we analyze this model under different situations (see farther below). Using our model does not mean blind cooperation, it also depends on the environment settings and opponent types.

3.3 How to learn with empathy?

We don't oppose the nature of selfishness, but think that selfishness is a driving force for learning. As shown in Fig.1, this rational learning procedure does not contradict the existence of empathy. Thus, we can get Adaptive Empathetic Learner (AEL) by introducing the empathy and counter-empathy into the extended adversary bandit structure. To this end, we need to modify the external reward using empathetic utility evaluation first.

Definition 2: For agent $i \in N$, the empathetic utility is defined by

$$E_{i,t} = \Lambda_{i,t} Y_t^T \tag{2}$$

where

$$\Lambda_{i,t} = \left(\lambda_{i,1,t}, \lambda_{i,2,t}, \cdots, \lambda_{i,N,t}\right) \tag{3}$$

$$Y_t = \left(r_{1,t}, r_{2,t}, \cdots, r_{N,t}\right) \in [0,1]^N \tag{4}$$

The empathetic utility is part of integrated utility arisen by empathy. Under conditions with a certain w_e, if one has positive empathy, the rewards (external feelings) of the others will be added on the one's utility. When others' rewards fluctuate greatly, agent's own experience also receives a direct positive correlation effect, but it will fade if there is a lack of sustained reinforcement. This is consistent with the feature of empathy–experiencing similar emotional feelings. On this basis, the agent will further emerge sympathy and generate motivation to alleviate the suffering of others. We will see this in section 4.1 and 4.3 (based on different games, linked with equality and fairness). And on the contrary, negative empathy (counter-empathy) will lead a difference in rewards between the two agents, which represents envy and gloating to some extent.

Note that if we want to introduce empathy into the learning procedure, the agents must have abilities to know others' reward. So here we give an assumption to limit the conditions that can be applied.

Assumption 1: In the $N-$agents reward-observable certain environment (ROCE), for arbitrary agent $i \in N$, its reward information $r_{i,t}$ could be observed by any agent $j \in N$, and the reward function of specific action combination $r_t(a_{1,t}, \cdots, a_{N,t}), a \in [K]$ is totally certain with no random distribution.

Then in ROCE, the learning objective for empathetic agent i is to maximize the total empathetic utility $S_{i,n} = \sum_{t=1}^{n} E_{i,n}$, which is a indeterminate quantity that depends on all actions of agents in the environment. For convenience, we can introduce the regret as most bandit algorithms do

$$R_{i,n} = n * E_{i,n}^* - \sum_{t=1}^{n} E_{i,n} \tag{5}$$

where n is the training horizon, $E_{i,n}^* = \max_{t, t \le n} E_{i,t}$. $R_{i,n}$ can transform the maximization problem into the minimization problem, so as to facilitate the analysis of the algorithm. Based on this,

the objective is to find a policy with sublinear regret such that

$$\lim_{n\to\infty} \frac{R_{i,n}}{n} = 0 \tag{6}$$

Next, data should be normalized before the training of learners. For the condition we defined in $definition1$ and $definition2$, our method makes agent i do random actions for n_r times at the beginning, and in the meantime record the maximum reward $r^*_{i,t}$ and the minimum reward $\tilde{r}_{i,t}$, then calculate the normalization for the actual reward $\hat{r}_{i,t}$ by

$$r_{i,t} = \frac{\hat{r}_{i,t} - \tilde{r}_{i,t}}{r^*_{i,t} - \tilde{r}_{i,t}} \tag{7}$$

We also record the empathetic utility $E_{i,t}(a)$ for each action a based on Eq.(1), Eq.(2) and Eq.(7). After the random state, we select action in a greedy mode to maximize the adaptive empathetic evaluation which is defined as

Definition 3: For agent i, the adaptive empathetic evaluation of action a is

$$A_{i,t}(a) \leftarrow \hat{E}_{i,t}(a) + \frac{\left(\beta^{m_{i,t}(a)} \max_t E_{i,t}(a) - \hat{E}_{i,t}(a)\right)}{\sqrt{T_{i,t}(a) - T_{i,t_c}(a) + 1}} \tag{8}$$

where

$$\hat{E}_{i,t}(a) = \begin{cases} \alpha \hat{E}_{i,t-1}(a) + (1-\alpha) E_{i,t}(a), & a = a_{i,t} \\ \hat{E}_{i,t-1}(a), & else \end{cases} \tag{9}$$

where α and β are decay factors, $T_{i,t}(a)$ stands for the times choosing action a in round t, and m is the times that actions can't remain receiving max utility, t_c is the time to clear the counts of action. There always exists a decay possibility for clearing the counts, which makes it always possible to try $a = \arg\max_a \beta^{m_{i,t}(a)} E_{i,t}(a)$ again. The latter two parameters will be further showed in the algorithm.

We maintain the greedy mode for at least n_g times and test the stability of utility by the dynamic changing rate.

Definition 4: For agent i in $N-$agents environment, with a reward vector $X_{i,t} = (r_{i,t-w_c}, \cdots, r_{i,t})$, where $r \in [0,1]$,$w_c \in \mathbb{N}$, the dynamic changing rate can be calculated by

$$c_{i,t} = \left(X^T_{i,t} X_{i,t}\right)^{-1} X^T_{i,t} y_i \tag{10}$$

where

$$y_i = (1, 2, \cdots, w_c + 1)^T \tag{11}$$

From the n_gth time, if the changing rate $c_{i,t} = 0$, the training will maintain greedy mode, otherwise, entering the random mode again.

Thus, the strategy for selecting action can be described as

$$a_{i,t} = \begin{cases} random, & t \le n_r, t_r < t \le t_r + n_r \\ \arg\max_a A_{i,t-1}(a), & else \end{cases} \tag{12}$$

where t_r is the time when training goes back to random mode again.

To this step, a completed structure of AEL can be carried out. The details of algorithm AEL for agent i is showed in Algorithm 1.

It can be easily proved that if $\beta = 1$ and the action for maximum empathetic utility of each agent is consistent and unique, the agents trained with AEL will get sublinear regret. As showed in Fig.3, while $n \to \infty$, agents with different n_r and n_g will have possibility to meet in greedy mode. Because of the mechanism that it also always possible to choose action of max utility in greedy mode, AELs will converge to action of max utility with probability to 1.

In our algorithm, the minimum length of greed stage n_g is restricted by four aspects. First, in order to form a certain consensus among individuals, we need a large n_g to ensure that different individuals can be greedy at the same time with a high probability (see Fig.3). Secondly, in the face of non-empathic objects, we need to maintain a period of time in the greedy stage to fully express the goodwill and then influence the decision-making trend of others. Thirdly, in the environment with multiple maxima (e.g. situations in the UG), the game with n_g times in the greedy stage can ensure the decay for the non-consensus maxima of individuals. Fourthly, the excessive greed stage n_g may lead to the insufficient exploration of the individual when facing the strategy change agent. Therefore, n_g's design, like the traditional exploration and exploitation problems, has a natural contradiction. At this stage, we still design ng based on experience.

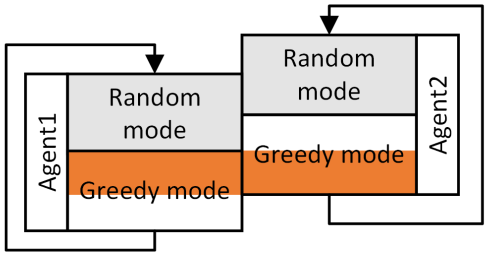

Figure 3: The interaction sketch of two AEL agents when their greedy modes have a period of time (the orange streak) activating simultaneously.

4 EXPERIMENTAL RESULTS

In this section, we present several experimental results of different games. Each game is designed specifically to test the performance of our algorithm.

4.1 Survival Game

Survival game is a kind of positive-sum stochastic game. We design this game to simulate the original hunting environment. In this game, multiple agents can choose to hunt independently or in teams. Generally, for a rational agent, it can evolve into a cooperative state in such positive-sum game, which is beneficial to its interests. Therefore, we can infer that rational agents will tend to team up to hunt in the survival game. Then what will be brought about by introducing empathy on this basis?

For this purpose, we slightly complicated the game settings, as follows:

(1) Agent types:
- $I \in \{I_1, I_2\}$ – stands for the strong agent;
- $I \in \{I_3, I_4\}$ – stands for the weak agent.

(2) Agent actions:
- $A_i \in \{1, 2, 3, 4\}$;
- $A_i = i$ – stands for hunting alone;
- $A_i = j, j \neq i$ – stands for teaming with agent j (only if $A_i = j$ and $A_j = i$, team could be formed successfully).

Algorithm 1: Adaptive Empathy Learner for agent i

Input: Initialize n,N,K,w_e,w_c,α,β,n_r,n_g,$\epsilon = 1$,$Flag1 = 0$,$Flag2 = 0$,$m_{i,0}(a) = 0$
Output: Output $a_{i,t}, r_{i,t}$

```
while not at end of training do
    while t < n do
        Do random ation for n_r times;
        Record r* and r̃ to do normalization;
        Update E*(a) for all a ∈ [K];
        Flag1 = 0;
        while t < n and Flag1 = 0 do
            Choose action to maximize Eq.(8);
            Record r* and r̃ to do normalization;
            Update E*(a) for all a ∈ [K];
            Set T_{i,t_c}(a) = T_{i,t}(a) for all a ∈ [K] with decay
              probability ϵ− = 1/n;
            if E_{i,t−1}(a_{i,t−1}) = max_a E*(a) then
                if E_{i,t}(a_{i,t}) < max_a E*(a) then
                    m_{i,t}(a) + +;
                end
            end
            Update Eq.8 for all a ∈ [K];
            Flag2 + +;
            if Flag2 >= n_g then
                Calculate the c_{i,t};
                if c_{i,t} > 0 then
                    Flag1 = 1
                end
            end
        end
    end
end
```

(3) Reward:

- $R = 1$ – if weak agent hunts alone;
- $R = 2$ – if 2 weak agents hunt together;
- $R = 3$ – if strong agent hunts alone;
- $R = 4$ – if the weak and the strong hunt together;
- $R = 6$ – if 2 strong agents hunt together.

We designedly introduce the diversity of agents on power and the setting that one team can accommodate up to two members, which make it possible to form two different teams. The experiment was simulated to test AEL's performance and with QL (learner with Q-learning method) as controls. The parameter setting of AEL is ($N = 4$, $K = 4$, $w_e = 1$, $w_c = 2$, $\alpha = 0.9$, $\beta = 0.99$, $n_r = 250$, $n_g = 250$). The results were collected after 10 episodes with 5000 steps training and we recorded the average value changes of these 10 episodes.

Fig.4 depicts the choices and payoffs among two strong agents AEL1, AEL2 and two weak agents AEL3, AEL4, contrasted by agents with Q-learning strategy. From the trend of payoffs, we can see that agents with AEL strategy finally formed teams that combine the strong and the weak while agents with Q-learning strategy

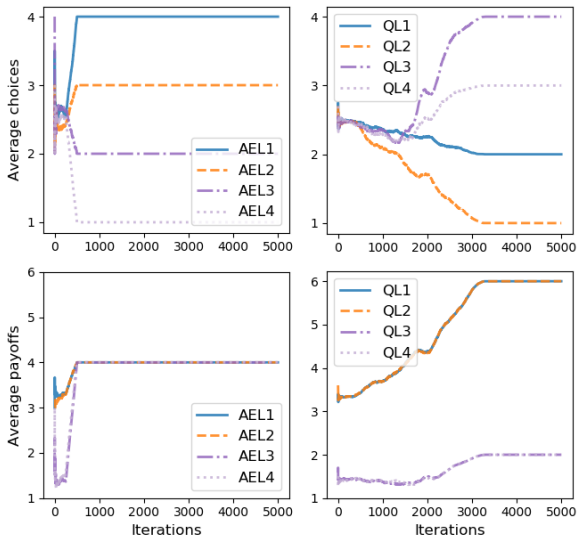

Figure 4: The average choices and payoffs of participants in the survival game, contrasted by four AELs and four QLs. AELs showed more cooperation between the strong and the weak.

formed teams with no such mixture. It is positive that agents with different power can collaborate with each other to reduce the gap between rich and poor. Besides, we can regard this kind of tendency as a primary drive of helping. Comprehensively, AELs with no-empathy initial state, generated large empathy as a result of long-term coexistence and then formed a team that contributes to the overall interests. And because of the emerging empathy across different teams, AELs may have enough motivation to help ones in another team.

The result is consistent with our previous assumption that the generalized empathy towards similar creature is beneficial to the stability of the community, which is a virtuous circle for the symbiotic community in the early stage.

4.2 Iterated Prisoner's Dilemma

The Prisoner Dilemma (PD) game is a paradigmatic example of game theory using to explain why it is difficult to maintain cooperation even when the cooperation is beneficial to both agents [2]. Iterated Prisoner's Dilemma (IPD) is the iterated form of Prisoner's Dilemma (PD). In IPD, each agent has access to the information of previous rounds and the symmetric interaction between agents could be computed by the canonical payoffs matrix defined as follows

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 3,3 & 0,5 \\ 5,0 & 1,1 \end{bmatrix} \end{array}$$

For rational agents, strategy of IPD converge to pure-strategy Nash equilibrium, more precisely $defect(1)$ even when the $cooperation(0)$ is beneficial to both agents. One simple reason is that players lacks trust and emotional concern on their partners.

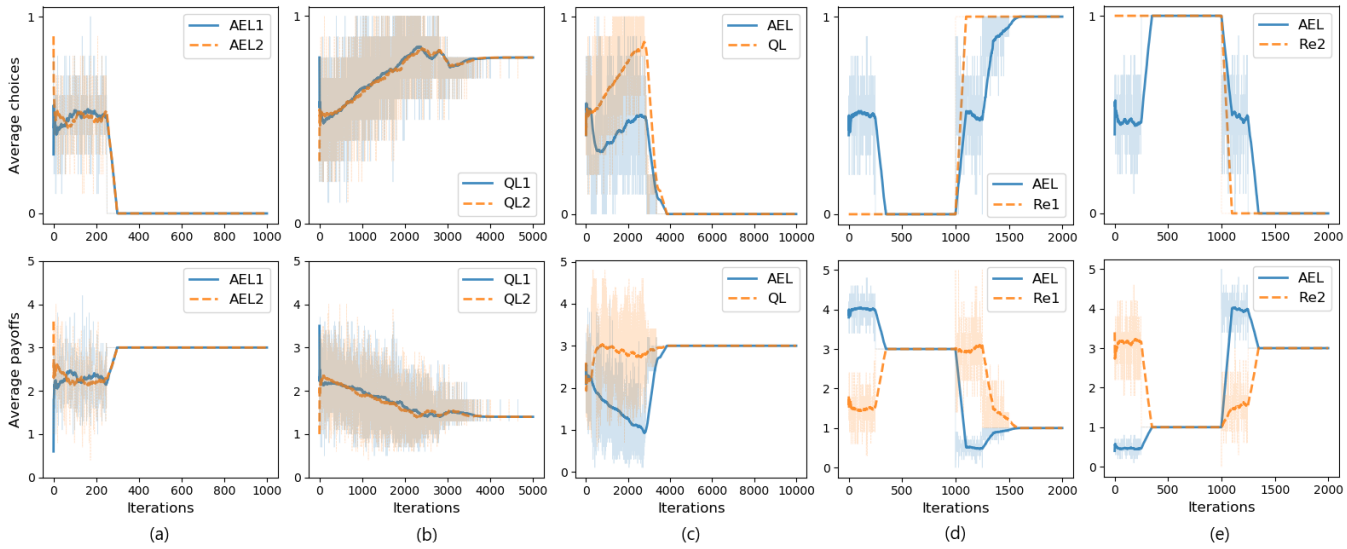

Figure 5: (a) The average choices and payoffs of two AELs playing in IPD. Both choices converge to *cooperate* with few iterations. (b) The average choices and payoffs of two QLs playing in IPD. The QLs tend to *defect* with a large possibility. (c) The average choices and payoffs of AEL and QL playing in IPD. QL's choice converge to *cooperate* after a long time play with AEL. (d) (e) The average choices and payoffs when AEL facing with deteriorated and ameliorated agents respectively in IPD. AEL can dynamically adjust choice if opponents change their strategies in process of interaction.

Thus, we simulated this experiment to test the tendency of agents' behavior, especially cooperation, after adaptive empathy is introduced in the internal model. For a more comprehensive comparison, we simulated the IPD games with the setting that AEL agent was paired with agents under different strategies separately and with two QLs pair as controls. The agent types considered follow TABLE 1.

Table 1: Agent types considered in IPD

Agent Type	Agent Characteristics	
	Strategy	***Sketch***
Empathetic	AEL	act to minimize of Eq.(5)
Rational	QL	act to maximize reward
Deteriorated	Re1	Reverse from *cooperate* to *defect*
Ameliorated	Re2	Reverse from *defect* to *cooperate*

The parameter setting of AEL is ($N = 2$, $K = 2$, $w_e = 1$, $w_c = 2$, $\alpha = 0.9$, $\beta = 0.99$, $n_r = 250$, $n_g = 250$). The results were collected after 10 episodes of training and we recorded the average values of the 10 episodes.

Data obtained in previous study using a simplified empathy model indicated that increased empathy affected agents' decisions [29]. In our works, AEL took into account a more general way of emotional transmission, involving adaptive empathy and counter-empathy. In our test, results depicted in Fig.5(a) showed that AEL could tend to *cooperate* rapidly when facing with AEL. This trend is consistent with the intention of self-compatibility – cooperation should be promoted if both agents use AEL strategy. And by comparing Fig.5(b) and Fig.5(c), we can see that, by adjusting choices properly during the interaction, AEL could affect the decisions of rational agent QL to some extent. And furthermore, AEL could show sufficient goodwill in the first place and once opponent was in antagonistic state, AEL was always equipped to protect the safe payoffs. This kind of interaction changed the mean payoffs of QL and make the QL tend to choose *cooperate*.

To do further research, we designed a comparative test between AEL and agents which changed strategy in the process of interaction. The relevant results were depicted in Fig.5(d) and Fig.5(e) in which Re1 stands for the deteriorated agent who reverse action from always *cooperate* to always *defect*, while Re2 stands for the ameliorated agent who reverse action from always *defect* to always *cooperate*. We can see that AEL can dynamically adjust strategy if opponents change their strategies in progress of interaction. Note that the adjusting time is different in the tests with Re1 and Re2. It took more time for AEL to adjust its behavior when opponent changed to be *defect*. This is consistent with AEL's characteristics that promoting proper competition in antagonistic environment and agonistic behavior is suppressed as much as possible.

4.3 Ultimatum Game

The Ultimatum Game (UG) is a classic non-zero game with two participants. It is the golden standard of examine fairness in biology and behavioral economics [6, 9]. In this game, fixed resources are split divided up by a proposer and a responder. The proposer presents a scheme for allocating resources to the responder, and if the responder agrees with the scheme, the resources are allocated as agreed. However, if the responder refused it, the two participants get nothing.

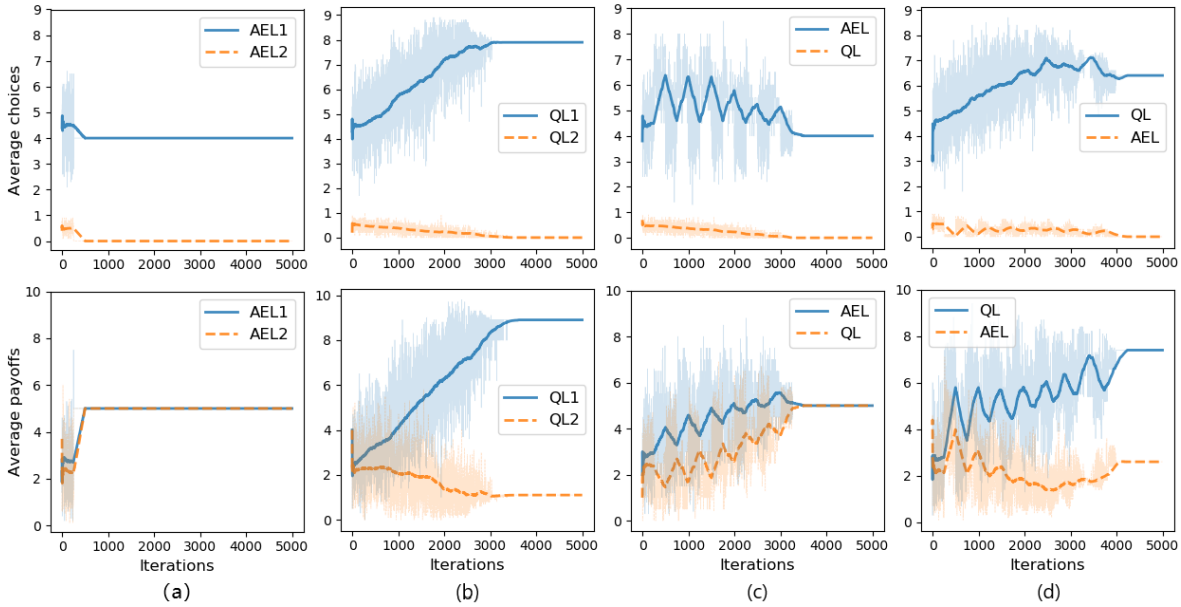

Figure 6: (a) The average choices and payoffs of two AELs playing in UG. The consensus of uniform distribution was reached in few interactions. (b) The average choices and payoffs of two QLs playing in UG. The proposer tend to give the responder as less as it can and a extremely unequal consensus was reached with a high probability. (c) The average choices and payoffs recorded when AEL acted as a proposer and QL as a responder in UG. AEL provided half resources to QL at last. (d) The average choices and payoffs recorded when QL acted as a proposer and AEL as a responder in UG. After a long term of gaming, AEL got a better result than the responder of (b).

A noteworthy feature of this game is the unequal interaction between the agents. The proposer is proactive and dominant on allocating the resources, while the responder's option is confined to acceptance or rejection. For the proposer, the rational strategy suggested by classical game theory is to offer the smallest possible positive share to the responder [28]. In this experiment, we are concerned about how the interactive behavior change if both proposer and responder learn with AEL.

The canonical payoffs matrix of the Ultimatum Game is defined as

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} p(r) & \dots & p(r_p) \quad \dots \quad p(0) \end{array} \\ \begin{bmatrix} r,0 & r_p, r-r_p & 0,r \\ 0,0 & 0,0 & 0,0 \end{bmatrix} \end{array}$$

where r stands for the total resources, $p(*)$ is the proposal that $*$ for the proposer itself and $(rki - *)$ for its partner, 0, 1 stand for partner's action *accept* or *reject*. In this article, the total resources r is 10 and the tick size (minimum change of proposal) is 1.

In marked contrast to the previous work on the Ultimatum Game, we simulated this experiment with AEL strategies and QL strategies respectively. The parameter setting of AEL is ($N = 2, K = 2, w_e = 1$, $w_c = 2, \alpha = 0.9, \beta = 0.99, n_r = 250, n_g = 250$). The results were also collected after 10 times of 5000 iterations training and we recorded the average values of the 10 times. As Fig.6(a)–(b) showed, game with both AEL strategies converged to a absolute fairness successfully while game with both QL strategies converged to the Nash equilibrium with a high probability. And according to Fig.6(c), when simulating UG using AEL as proposer and QL as responder, we found AEL provided half resources to QL at last. Conversely, if setting QL as proposer and AEL as responder, we found that, after a long term of gaming, AEL got a better result than the responder of Fig.6(b). This further validates the robustness of the algorithm.

The traditional improved methods directly introduced the relative difference as the optimization target of the proposer, so that the resources can be distributed actively and equitably. Different from the previous algorithms, AEL implements a more natural way of confrontation and achieves a fair distribution through continuous gaming. The key to proper confrontation is the existence of counter-empathy, when responder's relative income is less than the other side, the negative decision can be made firmly. Responder would rather has no income than accept less income, thus forcing the proposer to yield. We consider such result as a intrinsic moral improvement because the total resources is a constant and there is actually no Pareto improvement or Kaldor–Hicks improvement in Ultimatum Game.

5 CONCLUSIONS

Inspired by social neuroscience, our learning method AEL, for the first time, modeled empathy and counter-empathy in a unified way and illustrated that empathy and counter-empathy can act as a fundamental affective drive underlying interaction including cooperation and proper competition. This provides novel methods and insights in promoting complex interaction in multi-agent systems. It can also be used as artificial subjects in psychology and behavioral economics simulations and experiments.

Previous works inspired by psychology has already showed that introducing affective functions such as guilt and forgiveness could enhance cooperation. However, as for an autonomous multi-agent system as complex as human society, proper competition other than cooperation is also imperative to guarantee smooth functioning of the system. By introducing adaptive empathy and counter-empathy as a state sharing function, similar emotional states are induced among individuals, enabling each individual to feel what others feel. In this case, when conflicted interest occurs as in the prisoner dilemma game, the empathetic individual is able to take its opponent into consideration and sacrifice short-term reward for cooperation. Besides, the adjustment mechanism of empathy and counter-empathy introduced in this paper, makes it possible for agents to complete if the environment is competitive, which assures safe profits of each individual and secures the society's stable function as a whole.

There are also some limitations in this paper. The assumption that reward could be observed fully is a critical obstacle for the use of the empathetic utility in actual systems. We believe that a feasible method is to refer to the improvement of MDP by POMDP. Specifically, the empathetic utility model can be expressed as the probability distribution on the action set, by which the uncertainty of reward can be introduced. Then, according to the bayesian method, the posterior probability can be updated through the partial observation data. We will explore this direction in the future work.

In summary, AEL was designed by introducing emerging empathy and counter-empathy into the learning procedure, which is meaningful for agents to survive in social networks. Taken together, AEL has the following performance:

- With extended adversarial bandit structure, AEL can learn in the multi-agent environment.
- Cooperation could be reached among AELs, and in some situations, cooperation is possible even between AEL and rational agent.
- AEL can fully express its goodwill on the premise of protecting its safe payoffs, and won't take the initiative to make acts that harm the interests of the others.
- When the situation is antagonistic, AEL can adjust the degree of empathy and counter-empathy adaptively, and then change the coping strategy, cooperate or compete.
- AEL can elicit intrinsic moral improvement, especially absolute fairness if the state has no Pareto improvement or Kaldor–Hicks improvement.

REFERENCES

[1] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32, 1 (2002), 48–77.

[2] Robert Axelrod et al. 1987. The evolution of strategies in the iterated prisoner's dilemma. *The dynamics of norms* (1987), 1–16.

[3] Dipyaman Banerjee and Sandip Sen. 2007. Reaching pareto-optimality in prisoner's dilemma using conditional joint action learning. *Autonomous Agents and Multi-Agent Systems* 15, 1 (2007), 91–108.

[4] Tamer Basar and Geert Jan Olsder. 1999. *Dynamic noncooperative game theory*. Vol. 23. Siam.

[5] Michael Bowling and Manuela Veloso. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence* 136, 2 (2002), 215–250.

[6] Sarah F Brosnan and Frans BM de Waal. 2014. Evolution of responses to (un) fairness. *Science* 346, 6207 (2014), 1251776.

[7] Dorothy L Cheney and Robert M Seyfarth. 1985. Vervet monkey alarm calls: Manipulation through shared information? *Behaviour* 94, 1-2 (1985), 150–166.

[8] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[9] Stephane Debove, Nicolas Baumard, and Jean Baptiste Andre. 2016. Models of the evolution of fairness in the ultimatum game: a review and classification. *Evolution and Human Behavior* 37, 3 (2016), 245–254.

[10] F. U. Di, Q. I. Yanyan, W. U. Haiyan, and Xun Liu. 2017. Integrative neurocognitive mechanism of empathy and counter-empathy. *Chinese Science Bulletin* 62, 22 (2017), 2500–2508.

[11] Dietrich Dörner and Ulrike Starker. 2004. Should Successful Agents Have Emotions? The Role of Emotions in Problem Solving.. In *ICCM*. 344–345.

[12] Jean-Marc Fellous. 2004. From human emotions to robot emotions. *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence* (2004), 39–46.

[13] Judy Goldsmith and Emanuelle Burton. 2017. Why teaching ethics to AI practitioners is important. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

[14] Donald O. Hebb. 1988. *The organization of behavior*. MIT Press.

[15] Sarah E. Hill and David M. Buss. 2006. Envy and Positional Bias in the Evolutionary Psychology of Management. *Managerial & Decision Economics* 27, 2-3 (2006), 131–143.

[16] Hu, Junling, Wellman, and P Michael. 2004. Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research* 4, 4 (2004), 1039–1069.

[17] RE Kenward. 1978. Hawks and doves: factors affecting success and selection in goshawk attacks on woodpigeons. *The Journal of Animal Ecology* (1978), 449–460.

[18] S. Lakshmivarahan and Kumpati S. Narendra. 1981. Learning Algorithms for Two-Person Zero-Sum Stochastic Games with Incomplete Information. *Mathematics of Operations Research* 6, 3 (1981), 379–386.

[19] L. A. Martinezvaquero, T. A. Han, L. M. Pereira, and T. Lenaerts. 2015. Apology and forgiveness evolve to resolve failures in cooperative agreements. In *Benelux Conference on Artificial Intelligence*. 10639.

[20] William McDougall. 2015. *An introduction to social psychology*. Psychology Press.

[21] Melanie Mitchell and Zoltán Toroczkai. 2010. Complexity: A Guided Tour. *Physics Today* 63, 2 (2010), 47–48.

[22] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. 2017. Emotion in reinforcement learning agents and robots: a survey. *Machine Learning* 5 (2017), 1–38.

[23] I. P Pavlov. 2010. Conditioned reflex: An investigation of the physiological activity of the cerebral cortex. *Ann Neurosci* 8, 17 (2010), 136–141.

[24] Luís Moniz Pereira, Francisco C Santos, Tom Lenaerts, et al. 2013. Why is it so hard to say sorry? evolution of apology with commitments in the iterated Prisoner's Dilemma. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 177–183.

[25] GVN Powell. 1974. Experimental analysis of the social value of flocking by starlings (Sturnus vulgaris) in relation to predation and foraging. *Animal Behaviour* 22, 2 (1974), 501–505.

[26] Mandayam AL Thathachar and Pidaparty S Sastry. 2011. *Networks of learning automata: Techniques for online stochastic optimization*. Springer Science & Business Media.

[27] Karl Tuyls, Pieter Jan Hoen, and Bram Vanschoenwinkel. 2006. An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games. *Autonomous Agents and Multi-Agent Systems* 12, 1 (2006), 115–153.

[28] Mascha Van't Wout, René S Kahn, Alan G Sanfey, and André Aleman. 2006. Affective state and decision-making in the ultimatum game. *Experimental brain research* 169, 4 (2006), 564–568.

[29] Rodrigo Ventura. 2010. Emotions and empathy: A bridge between nature and society? *International Journal of Machine Consciousness* 2, 02 (2010), 343–361.

[30] Michael Weinberg and Jeffrey S. Rosenschein. 2004. Best-Response Multiagent Learning in Non-Stationary Environments. In *International Joint Conference on Autonomous Agents and Multiagent Systems*. 506–513.

[31] Chao Yu, Minjie Zhang, and Fenghui Ren. 2013. Emotional Multiagent Reinforcement Learning in Social Dilemmas. In *International Conference on Principles and Practice of Multi-Agent Systems*. 372–387.