Incorporating Indirect Reciprocity into Reinforcement Learning for Multi-Robot Collaboration in Mixed-Motive Stochastic Games

1st Yuxin Geng
School of Mathematical Sciences,
Beijing University of Posts and Telecommunications,
Beijing 100876, China
yuxin.evol@gmail.com

2nd Xingru Chen
School of Artificial Intelligence
Beihang University
Beijing 100191, China
xingrucz@gmail.com

***Abstract*—Multi-robot systems are increasingly deployed in scenarios where agents must balance self-interest with collective welfare. The standard Centralized Training with Decentralized Execution (CTDE) framework typically assumes full cooperation between agents. While it is effective in small teams, CTDE scales poorly and fails to capture the mixed-motive nature of many real-world scenarios, such as autonomous driving. In evolutionary game theory (EGT), indirect reciprocity (IR) has been proven to be an effective mechanism to foster large-scale decentralized cooperation in biological and social systems. The basic idea of IR is that cooperative individuals accumulate a positive reputation, and agents tend to cooperate with those who have a high reputation. We propose a novel multi-agent reinforcement learning (MARL) framework that integrates a reputation-based IR model into independent reinforcement learning (RL) agents, where the reputation evaluation can be either rule-based or powered by large language models (LLMs). To validate our approach, we consider a Sequential Snowdrift Game (SSG), where robots must decide whether to incur a personal cost to clear snow piles for a collective benefit. Our preliminary experiments with a straightforward rule-based reputation system show that independent Proximal Policy Optimization (PPO) agents fail to cooperate in such an environment. In contrast, introducing IR fosters the emergence of cooperation. Our findings demonstrate the potential of IR as a scalable, self-organizing coordination mechanism for multi-robot systems in mixed-motive scenarios, and pave the way for developing more socially intelligent robotic systems for real-world deployment.**

***Index Terms*—multi-agent reinforcement learning, mixed-motive stochastic games, indirect reciprocity, cooperation**

I. Introduction

Achieving complex tasks in multi-robot systems hinges on the coordination and cooperation of autonomous agents [1]. A popular paradigm in multi-agent reinforcement learning (MARL) is Centralized Training with Decentralized Execution (CTDE), where agents learn in a centralized manner but operate independently during execution [2]. While CTDE is effective for small teams of robots, it struggles to scale as the number of robots grows [3]. More importantly, the core assumption of full cooperation does not hold up in many real-world applications. These scenarios are typically mixed-motive, in the sense that each robot must balance its self-interest with the group's welfare [4]. For instance, in autonomous driving, each vehicle aims to minimize its own travel time while simultaneously benefiting from the overall smoothness of traffic flow [5]. These limitations motivate the pursuit of decentralized learning approaches in which cooperation emerges among self-interested robots without relying on a central controller or an engineered team reward.

Originating from Darwin's natural selection ideas, evolutionary game theory (EGT) provides a powerful theoretical framework to study large-scale cooperative behaviors [6], [7]. Within this framework, several key mechanisms have been identified that foster the emergence of cooperation in social dilemmas across natural and social systems, including direct reciprocity, indirect reciprocity (IR), and network reciprocity [8]. For direct reciprocity, agents cooperate based on repeated interactions with the same partners, originating from the simple idea of "if I cooperate now, you may cooperate later. On the other hand, IR functions on a broader, community-wide scale. Within IR, agents' cooperative behavior is based on the reputation of the other agents. Helping someone builds your own positive standing, which in turn makes it more likely that other members of the group will help you in the future [9]. Paired with social learning, the reputation-based dynamic has proven to be a remarkably effective and scalable mechanism for sustaining high levels of cooperation [10].

Inspired by the game-theoretic principles, recent MARL research has explored various mechanisms to foster cooperation in mixed-motive settings. A common approach is gifting, which allows agents to directly transfer rewards to one another [11], [12], enabling cooperative behavior to be incentivized. Other methods have focused on embedding psychological drivers like inequality-aversion into agents or leveraging the technique of self-play [13], [14]. More directly related to our work, a few studies have begun to integrate IR into MARL systems. For example, Smit et al. incorporate the reputation into the policy space of independent Q-learners and find that compared to social learning, only a narrower set of norms could lead to fair cooperation [15].

Corresponding author: Xingru Chen

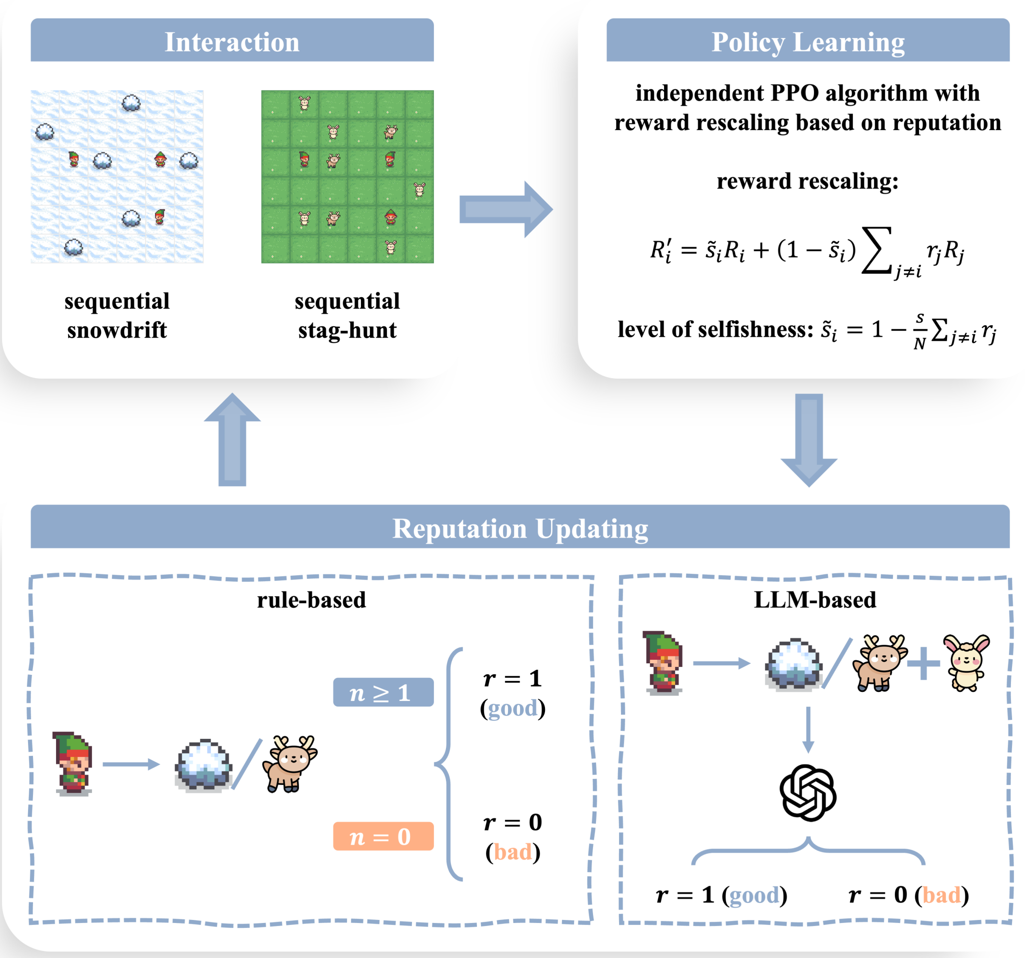
Interaction
sequential snowdrift
sequential stag-hunt
Policy Learning
independent PPO algorithm with reward rescaling based on reputation
reward rescaling:
$R_i' = \tilde{s}_i R_i + (1 - \tilde{s}_i) \sum_{j \neq i} r_j R_j$
level of selfishness: $\tilde{s}_i = 1 - \frac{s}{N} \sum_{j \neq i} r_j$
Reputation Updating
rule-based
$n \geq 1$
$r = 1$ (good)
$n = 0$
$r = 0$ (bad)
LLM-based
$r = 1$ (good)
$r = 0$ (bad)

Fig. 1. Algorithmic framework of the proposed IR-enhanced PPO

In this work, we introduce a novel framework that integrates IR directly into the reinforcement learning process for multi-robot systems. We treat reputation as a dynamic social layer that governs rescaling each agent's reward. The underlying social norm that determines what behaviors are considered "good" or "bad" can range from simple rule-based evaluation to more sophisticated LLM-based evaluation. Such a reputation system creates a positive feedback loop that allows early cooperative acts to be amplified through the reputation mechanism, thereby stabilizing cooperation. We validate our theoretical approach in the Sequential Snowdrift Game (SSG), in which robots must face the trade-off between the individual cost of clearing snow piles and the collective benefit of improving the navigability of the area. Although cooperation collapses sharply at the beginning of the training, with the IR mechanism, independent PPO agents can gradually restore to full cooperation. In essence, our framework reshapes agents' behavior to foster robust collaboration and offers a path toward more scalable and adaptable multi-robot systems.

II. Model and Method

A. Game Environment

The game environment is defined as an N-robot mixed-motive stochastic game. We consider a Sequential Snowdrift (SSG) game, in which there are a total of K randomly placed snow piles that the robots need to clear in the $H \times W$ grid world. The clearing of a snow pile provides a benefit b to all robots, but incurs a personal cost c (e.g., energy). The robots can choose to clear a snow pile by moving to the same spot of the snow pile. This setting naturally captures a social dilemma. On one hand, any individual robot is incentivized to free-ride by letting others incur the cost while still sharing the benefit. On the other hand, the episode reward is 0 for all robots if no snow pile is cleared. Therefore, the central challenge for the agents is overcoming the selfish incentive to wait for others to act and learning to collaborate.

As an alternative example, we use a Sequential Stag-Hunt (SSH) game as the testbed of our framework. In this environment, agents can hunt prey of hares and stags. A single agent can successfully catch a hare for a small amount of reward, while a stag yields a larger reward but requires the coordinated effort of at least two agents, with the reward being split equally among the catchers. This setup presents a difficult choice between pursuing the safe strategy of hunting hares individually and risking failure by attempting to cooperate to hunt a stag. For independent agents, mastering the coordination for a stag hunt is already difficult, and the constant availability of hares as an easy fallback further hinders the learning process towards cooperation.

B. Independent PPO with IR

Our method is built on the Independent PPO algorithm, with the key modification being the reward rescaling mechanism driven by each agent's reputation score [16]. Each robot maintains a reputation score that reflects its contributions to the shared tasks. We explore two approaches to evaluate the reputation of the agents:

1) Rule-based evaluation: For straightforward scenarios like the SSG, we can use a predefined rule to evaluate the reputation of the agents. Such a rule can be as simple as an agent earns a good reputation for the next training batch (or the next episode) if its cooperative actions (e.g., the number of snow piles it clears) exceed a predefined threshold.
2) LLM-based evaluation: For complex environments where ”cooperation” is nuanced and hard-coded rules are too brittle, we can alternatively use an LLM-based evaluator. In this mode, batched interaction data is structured and passed to an LLM, which assesses each agent's degree of cooperation and assigns a reputation score accordingly.

We adjust each agent's motivation using a reputation-based reward rescaling formula, which allows an agent to dynamically shift between selfish and altruistic behavior based on its reputation. Specifically, the rescaled reward R' for agent i is computed as

$$R'_i(\boldsymbol{r}) = \tilde{s}_i(\boldsymbol{r})R_i + (1 - \tilde{s}_i(\boldsymbol{r}))\sum_{j\neq i} r_j R_j. \tag{1}$$

Here, R_i is the original reward received, and $\boldsymbol{r}$ is the vector of all agents' reputation, with each r_j being binary (0 for bad and 1 for good). The weighting term $\tilde{s}_i(\boldsymbol{r})$ represents the agent's dynamic level of selfishness, which is calculated as

$$\tilde{s}_i(\boldsymbol{r}) = 1 - \frac{s}{N}\sum_{j\neq i} r_j. \tag{2}$$

The hyperparameter s controls the agents' baseline level of selfishness. The formula generalizes the reward exchange mechanism proposed by Willis et al. (2025), in the sense that our model reduces to in the special case where all agents are considered to have good reputations [17].

III. Experiments

We conduct preliminary experiments in SSG to evaluate the effectiveness of the proposed IR mechanism. The environment was configured with 5 snowpile, 4 PPO agents with baseline selfishness set to $s = 1$. We compared our IR-enhanced agents against a standard Independent PPO baseline. The mean episodic reward (solid line) along with the standard deviation (shaded area) across 5 runs is shown in Figure 2.

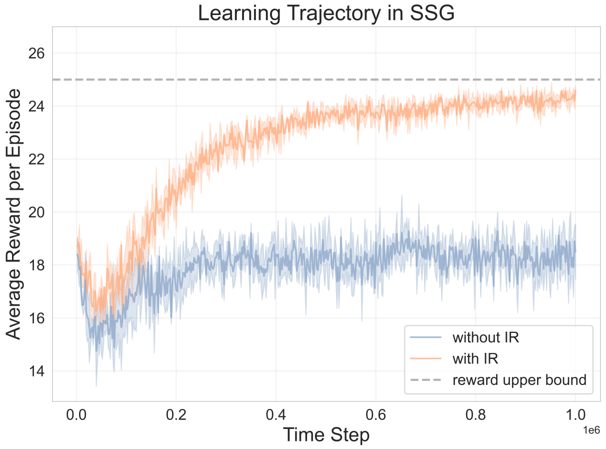

Fig. 2. Learning curve of the independent PPO with and without IR in SSG

Without IR, PPO agents fail to cooperate, with average rewards remaining below the random baseline. By contrast, IR acts as a bootstrap catalyst for collaboration, as occasional snow pile clearing by chance builds reputations, prompting other agents to align with high-reputation peers, which in turn reinforces cooperative behavior. This positive feedback loop steadily raises both reputation and cooperation levels and drives the average reward close to the theoretical maximum when all snow piles are cleared. Notably, the average reward initially declines, as agents discover that deferring snow clearing to others yields short-term gains. While independent PPO alone struggles to recover from this collapse, the IR mechanism facilitates the gradual build-up of positive reputations, thereby enabling agents to quickly reach and sustain full cooperation.

IV. Conclusion and Discussion

We proposed an IR-enhanced MARL framework that dynamically updates the cooperative relationships among agents through reputation. From this perspective, theoretical advances in cooperative RL can be naturally embedded into our framework as promising directions for future research. On the other hand, since our current work considers only the simplest form of social norm, it is worth exploring which types of norms most effectively foster collaboration. In summary, our study provides a promising paradigm for bridging EGT and MARL, offering both theoretical insights and practical tools for building scalable and socially intelligent robotic systems.

Acknowledgment

This work is supported by the Beijing Natural Science Foundation (grant no. 1244045).

References

[1] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, “Swarm robotics: a review from the swarm engineering perspective,” *Swarm Intelligence*, vol. 7, no. 1, pp. 1–41, 2013.

[2] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.

[3] K. Zhang, Z. Yang, and T. Başar, “Multi-agent reinforcement learning: A selective overview of theories and algorithms,” *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[4] 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.

[5] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus, “Social behavior for autonomous vehicles,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 50, pp. 24 972–24 978, 2019.

[6] J. M. Smith, “Evolution and the theory of games,” in *Did Darwin get it right? Essays on games, sex and evolution*. Springer, 1982, pp. 202–215.

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

[8] ——, “Five rules for the evolution of cooperation,” *science*, vol. 314, no. 5805, pp. 1560–1563, 2006.

[9] M. A. Nowak and K. Sigmund, “Evolution of indirect reciprocity by image scoring,” *Nature*, vol. 393, no. 6685, pp. 573–577, 1998.

[10] ——, “Evolution of indirect reciprocity,” *Nature*, vol. 437, no. 7063, pp. 1291–1298, 2005.

[11] A. Lupu and D. Precup, “Gifting in multi-agent reinforcement learning,” in *Proceedings of the 19th International Conference on autonomous agents and multiagent systems*, 2020, pp. 789–797.

[12] F. Kong, Y. Huang, S.-C. Zhu, S. Qi, and X. Feng, “Learning to balance altruism and self-interest based on empathy in mixed-motive games,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 135 819–135 842, 2024.

[13] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez-Guzman, A. García Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster *et al.*, “Inequity aversion improves cooperation in intertemporal social dilemmas,” *Advances in neural information processing systems*, vol. 31, 2018.

[14] N. Anastassacos, J. García, S. Hailes, and M. Musolesi, “Cooperation and reputation dynamics with reinforcement learning,” *arXiv preprint arXiv:2102.07523*, 2021.

[15] M. Smit and F. P. Santos, “Learning fair cooperation in mixed-motive games with indirect reciprocity,” *arXiv preprint arXiv:2408.04549*, 2024.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

[17] R. Willis, Y. Du, J. Z. Leibo, and M. Luck, “Quantifying the self-interest level of markov social dilemmas,” *arXiv preprint arXiv:2501.16138*, 2025.