Dynamic Reward Sharing to Enhance Learning in the Context of Multiagent Teams

Extended Abstract

Kyle Tilbury*
University of Waterloo
Waterloo, Canada
ktilbury@uwaterloo.ca

David Radke*
Chicago Blackhawks
Chicago, USA
dradke@blackhawks.com

ABSTRACT

In multiagent environments with individual learning agents, social structure, defined through shared rewards, has been shown to significantly impact how agents learn. However, defining reward-sharing parameters within a social structure that best support learning remains a challenging, domain-dependent problem. We address this challenge with a decentralized framework inspired by meta-reinforcement learning where independent reinforcement learning (RL) agents dynamically learn reward-sharing hyperparameters using a secondary RL policy. Agents' secondary RL policies shape the reward function and guide the learning process for their primary behavioral policies acting within a multiagent RL (MARL) environment. We show that our process enhances individual learning and population-level outcomes for overall reward and equality compared to agents without this secondary reward function shaping policy. Furthermore, we show that our framework learns highly effective heterogeneous reward-sharing parameters.

KEYWORDS

Multiagent Reinforcement Learning, Meta-Reinforcement Learning, Coordination

ACM Reference Format:
Kyle Tilbury* and David Radke*. 2025. Dynamic Reward Sharing to Enhance Learning in the Context of Multiagent Teams: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

The study of cooperation is a crucial component for future artificial intelligence (AI) systems and is often explored in reinforcement learning (RL) and multiagent systems [3, 4]. Individual agents that learn to cooperate can enhance their capabilities beyond those of a single agent; however, agents often encounter mixed motive scenarios where RL is known to develop non-cooperative behavior [6]. Fully cooperative systems have typically been used as a benchmark with which to compare coordination abilities; however, recent work has highlighted limitations of fully cooperative systems while emphasizing some benefits of mixed incentives for learning efficient behavior [5, 10, 11].

These studies typically assume some underlying structure to the population in the form of teams or groups [8, 9] and mixed incentives are explored using exhaustive or heuristic search [7]. Mixed incentives can be implemented by either defining some population structure and modifying agents' social dependencies within this structure (through sharing reward among groups) or defining inter-agent social dependencies among groups and changing the underlying population structure. We present an agent architecture that learns to navigate the space of inter-agent dependencies using RL, autonomously adapting within an environment with underlying social structure among other learning agents. We assume populations of individual learning agents with some defined social structure through *teams*, sub-groups of agents that may have some degree of common interest where $T_n \in \mathcal{T}_i$ represents a team that any agent i belongs to. Agents must belong to at least one team; however, our architecture allows agents to modify how much reward they choose to keep to themselves, share with any team, or share with the entire system, expanding the *credo* model [9].

We focus on the *prescriptive* agenda of multiagent RL (MARL), studying the behaviors and performance of agents *during* learning [1]. Our agent architecture utilizes two internal RL policies, behavioral and reward-sharing, that operate at different timescales. This interaction is similar to problems in meta-RL, where the behavioral policy represents the "inner-loop" and the reward-sharing policy represents the "outer-loop" [2, 12] to learn hyperparameters of the reward function that best support the overall learning objective. With this framework, we show how populations autonomously converge to heterogeneous reward-sharing schemes while generating 34.2% more reward than a population of agents that do not update their reward sharing behavior. We also show that these learned reward-sharing schemes are highly effective in terms of global reward and equality when used for new agent populations, overcoming the need for exhaustive or heuristic search.

2 METHODOLOGY

We propose a decentralized framework that dynamically updates reward-sharing parameters in MARL environments using meta-RL. Agents learn to update their reward-sharing configurations while they learn to behave in the environment. Each agent maintains a reward-sharing configuration, called *credo* [9], $\mathbf{cr}_i = \langle \psi_i, \phi_i^{T_n}, \omega_i \rangle$. Credo defines a reward-sharing configuration of how much reward an agent shares among each group, where ψ is the credo parameter

*These authors contributed equally to this work.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

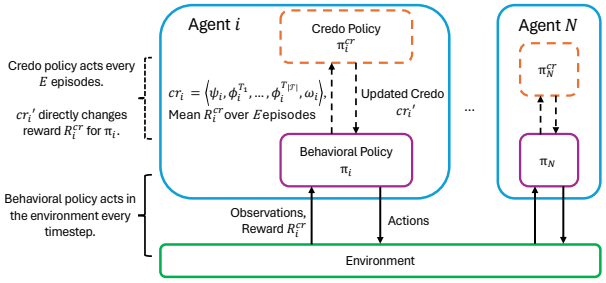
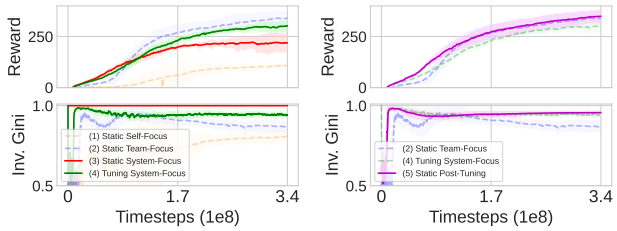


Figure 1: Overview of our agent architecture.



(a) Online Tuning trial. **(b) Post-tuning trial.**

Figure 2: Mean population reward (top) and Inverse Gini index (bottom) across tuning and post-tuning experiments.

for i's individual reward IR_i, $\phi_i^{T_n}$ is the credo parameter for a team-based reward $TR_i^{T_n}$ from team T_n, and ω_i is the credo parameter for the system-based reward of the entire system SR. The credo parameters within $\mathbf{cr}_i$ always sum to one.

Figure 1 shows our proposed agent architecture, in which we define agents as *tuning* credo agents. Each agent i is composed of two internal policies, $i = \langle \pi_i, \pi_i^{\mathbf{cr}} \rangle$. Agent i's *low-level behavioral policy* (purple solid box), denoted π_i, operates within a multiagent environment, observing states and taking actions in the context of other learning agents. Agent i's *high-level credo policy* (orange dashed box), denoted $\pi_i^{\mathbf{cr}}$, operates at a longer timescale in the space of the behavioral policy's reward function, a meta-environment, by observing how the behavioral policy shares reward among different groups in the population and taking actions to update the credo parameters of the agent's reward function.

Since $\pi_i^{\mathbf{cr}}$ updates i's own reward-sharing parameters, different agents can develop heterogeneous parameters for the same group. Allowing for these heterogeneous reward-sharing distributions between agents, we define the team-based reward channel for any team T_n as: $TR_i^{T_n} = \sum_{j \in T_n} \phi_j^{T_n} R_j(S, A_j, S)$, and define the system-based reward channel as: $SR_i = \sum_{j \in \mathcal{N}} \omega_j R_j(S, A_j, S)$. Let IR_i represent the agent's normal individual exogenous reward they receive from the environment for observing the state and taking individual actions. Given our definitions of IR_i, $TR_i^{T_n}$, and SR_i, we define each agent's credo-based reward function $R_i^{\mathbf{cr}}$ for their behavioral policy, π_i, to be:

$$R_i^{\mathbf{cr}} = \psi_i \cdot IR_i + \sum_{T_n \in \mathcal{T}_i} \frac{\phi_i^{T_n}}{\sum_{j \in T_n} \phi_j^{T_n}} \cdot TR_i^{T_n} + \frac{\omega_i}{\sum_{j \in \mathcal{N}} \omega_j} \cdot SR_i.$$

Note that the team and system-based reward channels depend on both the reward and credo parameters of all agents on that respective team or in the system. Furthermore, our allocation of team and system-based reward among several agents based on their ratio of credo parameters maintains the budget balance principle. Agents' credo policies, $\pi_i^{\mathbf{cr}}$, learn using RL to optimize the mean credo-based reward of their behavioral policy over E episodes.

3 RESULTS

We empirically evaluate our framework and compare with various static (i.e., non-tuning) populations using the Cleanup Gridworld Game (Cleanup) [14] as the environment for the behavioral policies. Cleanup presents a social dilemma where agents have the short term incentive to act selfishly but gain higher global rewards through cooperation. Agents' low-level behavior policies, π_i, learn using PPO [13, 16]. Agents' high-level credo policies, $\pi_i^{\mathbf{cr}}$, modifies how much reward they allocate to themselves, their teams, or the overall system, thereby influencing their long-term behavior in the low-level environment. The credo policy takes actions in the meta-environment that we define to be a discretized space of possible credo parameters. The credo policy for each agent is implemented with Q-Learning with ϵ-greedy exploration ($\epsilon = 20\%$) [15]. Experiments consist of 3.4×10^8 environmental timesteps (episodes of 1,000 timesteps). We define three disjoint teams of two agents each from a population of six agents. In all tuning experiments, the credo policy takes actions every $E = 96$ episodes.

Figures 2a shows mean population reward (top) and Inverse Gini index (bottom) for various experiments with 95% confidence intervals. The green solid line represents agents utilizing our meta-RL architecture that are first instantiated as fully cooperative agents (i.e., *system-focused*; all agents share rewards). We contrast these results with *static* self-focused agents (yellow dashed; individual RL), *static* system-focused agents (red solid; fully cooperative), and *static* team-focused agents (blue dashed; shares reward with team T_n, current best result) that do not update their reward functions. Our results show that credo-tuning agents update their joint behavior to achieve 34.2% more reward than the system-focused setting they are instantiated with (red line). Furthermore, credo-tuning populations develop noticeably more reward equality than the static team-focused setting that achieves slightly more population reward.

In our next experiment, we instantiate new behavioral policies with fixed heterogeneous credo parameters learned during the previous credo-tuning experiment. Figure 2b shows the results of this population (purple solid) contrasted with the previous tuning experiment (green dashed) and static team-focused (blue dashed) populations. Utilizing these learned credo parameters in these new agents achieves the highest observed mean population reward in Cleanup, slightly higher than previously best-observed setting (static team-focused), while achieving significantly higher reward equality.

Takeaways. Our approach demonstrates the viability of agents autonomously adapting in MARL environments, improving cooperation and learning outcomes without manual tuning of reward-sharing schemes within a social structure. These agents simultaneously explore the space of reward-sharing parameters and develop heterogeneous schemes which would be infeasible using previous methods.

REFERENCES

[1] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. 2024. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press. https://www.marl-book.com

[2] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. 2023. A Survey of Meta-Reinforcement Learning. *arXiv preprint arXiv:2301.08028* (2023).

[3] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: Machines Must Learn to Find Common Ground. *Nature* 593 (2021), 33–36.

[4] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).

[5] Ishan Durugkar, Elad Liebman, and Peter Stone. 2020. Balancing Individual Preferences and Shared Objectives in Multiagent Reinforcement Learning. *Good Systems-Published Research* (2020).

[6] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 464–473.

[7] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duèñez-Guzmán, Edward Hughes, and Joel Z Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. *AAMAS* (2020).

[8] David Radke, Kate Larson, and Tim Brecht. 2022. Exploring the Benefits of Teams in Multiagent Learning. In *IJCAI*.

[9] David Radke, Kate Larson, and Tim Brecht. 2023. The Importance of Credo in Multiagent Learning. *Proceedings of the 22nd International Conference on Autonomous Agents and MultiAgent Systems* (2023).

[10] David Radke, Kate Larson, Tim Brecht, and Kyle Tilbury. 2023. Towards a Better Understanding of Learning with Multiagent Teams. *IJCAI* (2023).

[11] Stefan Roesch, Stefanos Leonardos, and Yali Du. 2024. The Selfishness Level of Social Dilemmas. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2441–2443.

[12] Jürgen Schmidhuber. 1987. *Evolutionary Principles in Self-Referential Learning, or On Learning How to Learn: The Meta-Meta-... Hook*. Ph.D. Dissertation. Technische Universität München.

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* (2017).

[14] Eugene Vinitsky, Natasha Jaques, Joel Leibo, Antonio Castenada, and Edward Hughes. 2019. An Open Source Implementation of Sequential Social Dilemma Games. https://github.com/eugenevinitsky/sequential_social_dilemma_games/issues/182. GitHub repository.

[15] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8 (1992), 279–292.

[16] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.