Pareto-Guided Exploration for Multi-Objective Multiagent Learning

Extended Abstract

Gaurav Dixit
Oregon State University
Corvallis, USA
dixitg@oregonstate.edu

Kagan Tumer
Oregon State University
Corvallis, USA
kagan.tumer@oregonstate.edu

ABSTRACT

Cooperative multi-objective multiagent settings require discovering teams that realize diverse trade-offs across conflicting objectives. In multi-objective reinforcement learning, policies are typically optimized via scalarization of expected returns, while Pareto-based methods approximate sets of non-dominated solutions in objective space. In cooperative settings, however, these perspectives are typically treated in isolation. We introduce **Multiagent Pareto-Led Exploration (MAPLE)**, a framework that couples preference-aligned actor-critic learning with Pareto dominance-based selection over expected team returns. MAPLE trains policies under multiple preference vectors while preserving and reusing agents based on their contribution to non-dominated teams through a shared archive. Empirical results in cooperative continuous-control benchmarks demonstrate improved Pareto front coverage and greater policy composability across trade-offs. These findings suggest that coupling scalarized learning with Pareto-level selection provides a principled mechanism for multi-objective multiagent learning.

KEYWORDS

Multiagent Learning; Reinforcement Learning; Evolutionary Algorithms; Multi-Objective Multiagent Optimization

ACM Reference Format:
Gaurav Dixit and Kagan Tumer. 2026. Pareto-Guided Exploration for Multi-Objective Multiagent Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/AAWR9928

1 INTRODUCTION

Cooperative multiagent systems often operate under multiple, conflicting objectives [2, 12, 17]. In such multi-objective multiagent systems (MOMAS), success requires discovering sets of teams that realize diverse trade-offs, rather than a single optimal policy [13]. Moreover, team composition and objective priorities may change after training, demanding policies that are reusable and composable across preferences [1, 4].

In multi-objective reinforcement learning, policies are typically optimized by maximizing a scalarization of expected vector returns, $u_\lambda(\mathbf{J}) = \lambda^\top \mathbb{E}[\mathbf{J}]$, under a preference vector λ [14]. While this approach efficiently specializes policies to individual trade-offs, it does not by itself ensure recovery of a well-covered Pareto set of team outcomes [8]. Conversely, population-based methods can approximate Pareto fronts via dominance-based selection over expected returns, but often lack gradient-based mechanisms for improving individual policies through shared experience [3, 7]. As a result, scalarization-based learning and Pareto-level selection are rarely tightly coupled in cooperative settings. This separation is limiting in MOMAS [16]. Scalarized learning alone may discard policies that are suboptimal under a particular preference yet essential for forming Pareto-efficient teams. Pareto selection alone may preserve diverse teams but provide weak guidance for improving individual agents [5, 6].

We introduce **Multiagent Pareto-Led Exploration (MAPLE)**, a framework that explicitly couples these two mechanisms. MAPLE trains preference-aligned anchor policies via decentralized actor-critic updates while evaluating and preserving policies based on non-dominated team outcomes. A shared archive links learning and selection: policies are retained for their contribution to Pareto-efficient teams, and this external multi-objective pressure shapes subsequent exploration and optimization. Empirical results demonstrate improved Pareto coverage and greater policy composability relative to scalarization-only baselines, suggesting that integrating scalarized learning with Pareto dominance-based selection offers a principled mechanism for cooperative MOMAS.

2 BACKGROUND

We consider a cooperative multi-objective decentralized partially observable Markov decision process (MO-Dec-POMDP) in which a team T of agents interacts in a shared environment and produces an episode-level vector return $\mathbf{J}(T) \in \mathbb{R}^d$ [13]. This corresponds to the team-reward / individual-utility setting in the MOMA taxonomy: agents share a vector-valued team outcome but evaluate it through individual preference vectors. In multi-objective reinforcement learning, a common formulation is scalarised expected return (SER), where utility under preference $\lambda \in \Delta^d$ is given by $u_\lambda(T) = \lambda^\top \mathbb{E}[\mathbf{J}(T)]$ [15]. SER is appropriate when policies are executed repeatedly and performance is determined by expected outcomes. Policies trained under different λ specialize to different trade-offs. Alternatively, teams can be evaluated directly in objective space using Pareto dominance: T dominates T' if $\mathbb{E}[\mathbf{J}(T)]$ is no worse in all objectives and strictly better in at least one. The set of non-dominated teams approximates the Pareto frontier, representing the best achievable trade-offs. Recent population-based approaches maintain sets of specialized policies across preference

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/AAWR9928

vectors, such as Malthusian Reinforcement Learning (MRL) and its multi-objective variants [10, 11]. Extensions like MO-AIM incorporate Pareto-based team selection, while gradient-based methods such as MOMAPPO optimize scalarized objectives [5, 8]; however, these approaches typically emphasize either decentralized scalarized learning or Pareto-level selection without consideration for both perspectives.

3 METHOD: MULTIAGENT PARETO-LED EXPLORATION

MAPLE couples scalarization-based policy optimization with Pareto dominance-based selection over team outcomes. The framework maintains a set of K anchor policies, each associated with a preference vector λ_k. Each anchor is trained using Twin Delayed Deep Deterministic Policy Gradients (TD3), optimizing the scalarised expected return $\lambda_k^\top \mathbb{E}[\mathbf{J}]$ [9]. This produces policies specialized to distinct trade-offs while leveraging off-policy experience.

To promote structured exploration around each trade-off, MAPLE applies a Cross-Entropy Method (CEM) update around each anchor. Policy parameter variants are sampled from anchor-centered distributions and evaluated as part of candidate teams. Teams are formed by combining anchors, sampled variants, and archived policies, and are evaluated to produce expected vector returns $\mathbb{E}[\mathbf{J}(T)]$.

MAPLE then performs non-dominated sorting (NSGA-II) over team returns to identify Pareto-efficient teams [3]. Policies that participate in non-dominated teams are inserted into a shared archive. The archive serves both as a reservoir for future team composition and as a mechanism for refreshing anchor centers toward high-performing archived policies. Thus, TD3 drives preference-aligned improvement under fixed scalarizations, CEM encourages local behavioral diversity, and Pareto selection preserves policies based on their contribution to frontier teams. The shared archive links these processes, allowing external multi-objective pressure to shape both exploration and subsequent learning. This ensures that policy improvement is guided by both: the scalarized returns and by the contribution to the evolving Pareto frontier.

4 EXPERIMENTS

Environment. We evaluate MAPLE on *Assurance ItemGathering*, a continuous multiagent benchmark derived from MO-ItemGathering. Agents navigate a shared two-dimensional environment to collect two resource types corresponding to distinct objectives. Nectar yields high reward but requires coordinated collection by multiple agents, while sap can be collected individually for lower reward. This induces a classic assurance-style trade-off between risky coordination and guaranteed individual gain, producing a team-level vector return $\mathbf{J}(T) \in \mathbb{R}^2$.

Metrics. We report Hypervolume (HV) and Sparsity of the recovered Pareto frontier over expected team returns. HV measures the dominated objective-space volume, while Sparsity captures the density of solutions along the frontier.

Results. Table 1 shows that MAPLE achieves higher hypervolume and lower sparsity than scalarization-only baselines, indicating broader coverage and denser approximation of the Pareto frontier. MAPLE consistently recovers both extreme and balanced trade-offs, whereas scalarization-based training under-represents intermediate regions.

Table 1: Pareto front metrics for the Assurance ItemGathering environment (mean ± std over 10 seeds). MAPLE achieves the highest HV and lowest sparsity.

Method	HV (↑)	Sparsity (↓)
Assurance ItemGathering		
MAPLE	**0.63 ± 0.042**	**0.018 ± 0.01**
MOMAPPO	0.51 ± 0.063	0.01 ± 0.067
MO-AIM	0.61 ± 0.044	0.021 ± 0.033
MRL	0.44 ± 0.083	0.017 ± 0.024
TD3 + NSGA-II	0.55 ± 0.02	0.053 ± 0.08

We additionally evaluate MAPLE on *Allelopathy*, a cooperative benchmark derived from Malthusian Reinforcement Learning featuring metabolic specialization and indirect coordination. Results show similar improvements in Pareto coverage. Finally, varying the number of anchors K reveals a trade-off between coverage and convergence: moderate values provide strong frontier approximation, while too few anchors limit exploration and excessive anchors dilute optimization pressure.

These results demonstrate that coupling scalarized learning with Pareto-level selection improves frontier quality and policy composability across distinct cooperative multi-objective regimes.

5 DISCUSSION

MAPLE demonstrates that coupling scalarized policy learning with Pareto dominance-based team selection yields measurable gains in frontier coverage and policy composability. Rather than treating scalarization and Pareto search as separate paradigms, MAPLE integrates them through a shared archive that preserves agents based on their contribution to Pareto-efficient teams. This design shifts the unit of selection from isolated policy performance under a fixed preference to compositional value within multi-objective teams. Importantly, MAPLE does not rely on explicit diversity-based rewards or novelty objectives. Instead, diversity emerges from selection pressure at the team level: agents are retained because they enable frontier trade-offs, not because they are behaviorally distinct in isolation. This suggests that Pareto-led selection can act as a coordination prior, implicitly encouraging complementary specialization without sacrificing learning efficiency.

More broadly, MAPLE points toward a general design principle for cooperative MOMAS: combine preference-aligned reinforcement learning for efficient specialization with population-level Pareto selection to preserve trade-off coverage. Future work may extend this framework through adaptive preference placement, shared preference-conditioned critics, or credit assignment mechanisms that more precisely quantify individual contributions to team-level Pareto improvements. Such extensions could further strengthen the bridge between decentralized learning and multi-objective coordination.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant No. NSF IIS-2112633.

REFERENCES

[1] Adrian K Agogino and Kagan Tumer. 2004. Unifying temporal and structural credit assignment problems. In *Autonomous agents and multi-agent systems conference*.

[2] Wolfram Burgard, Mark Moors, Cyrill Stachniss, and Frank E Schneider. 2005. Coordinated multi-robot exploration. *IEEE Transactions on robotics* 21, 3 (2005), 376–386.

[3] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.

[4] Gaurav Dixit and Kagan Tumer. 2023. Learning inter-agent synergies in asymmetric multiagent systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1569–1577.

[5] Gaurav Dixit and Kagan Tumer. 2023. Learning synergies for multi-objective optimization in asymmetric multiagent systems. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 447–455.

[6] Gaurav Dixit and Kagan Tumer. 2024. Informed Diversity Search for Learning in Asymmetric Multiagent Systems. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 313–321.

[7] Gaurav Dixit and Kagan Tumer. 2024. Objective-Informed Diversity for Multi-Objective Multiagent Coordination.. In *ECAI*. 3660–3667.

[8] Florian Felten. 2024. Multi-Objective Reinforcement Learning. (2024).

[9] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*. PMLR, 1587–1596.

[10] Joel Z. Leibo, Julien Perolat, Edward Hughes, Steven Wheelwright, Adam H. Marblestone, Edgar Duéñez Guzmán, Peter Sunehag, Iain Dunning, and Thore Graepel. 2019. Malthusian Reinforcement Learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1099–1107.

[11] Thomas Pierrot, Guillaume Richard, Karim Beguir, and Antoine Cully. 2022. Multi-objective quality diversity optimization. In *Proceedings of the genetic and evolutionary computation conference*. 139–147.

[12] Roxana Rădulescu, Manon Legrand, Kyriakos Efthymiadis, Diederik M Roijers, and Ann Nowé. 2018. Deep multi-agent reinforcement learning in a homogeneous open population. In *Benelux Conference on Artificial Intelligence*. Springer, 90–105.

[13] Roxana Rădulescu, Patrick Mannion, Diederik M Roijers, and Ann Nowé. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34, 1 (2020), 10.

[14] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research* 21, 178 (2020), 1–51.

[15] Diederik M Roijers, Shimon Whiteson, Ronald Brachman, and Peter Stone. 2017. *Multi-objective decision making*. Springer.

[16] Peter Vamplew, Benjamin J Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M Roijers, Conor F Hayes, Fredrik Heintz, Patrick Mannion, Pieter JK Libin, et al. 2022. Scalar reward is not enough: A response to silver, singh, precup and sutton (2021). *Autonomous Agents and Multi-Agent Systems* 36, 2 (2022), 41.

[17] Logan Yliniemi, Adrian K Agogino, and Kagan Tumer. 2014. Multirobot coordination for space exploration. *AI Magazine* 35, 4 (2014), 61–74.