# Fairness in Cooperative Multi-agent Multi-objective Reinforcement Learning using the Expected Scalarized Return

Extended Abstract

Farès Chouaki
LIP6, Sorbonne Université, CNRS
F-75005 Paris, France
fares.chouaki@lip6.fr

Aurélie Beynier
LIP6, Sorbonne Université, CNRS
F-75005 Paris, France
aurelie.beynier@lip6.fr

Nicolas Maudet
LIP6, Sorbonne Université, CNRS
F-75005 Paris, France
nicolas.maudet@lip6.fr

Paolo Viappiani
LAMSADE, CNRS, Université Paris Dauphine - PSL
Paris-75016, France
paolo.viappiani@lamsade.dauphine.fr

## ABSTRACT

Fairness is essential for deploying artificial decision-making agents in the real world. Existing work in sequential decision-making ensures fairness among agents or objectives but struggles with real-world problems that are both multi-agent and multi-objective. Furthermore, research integrating fairness into Multi-Objective Reinforcement Learning (MORL) is focused on ensuring fairness over the objectives only on the average of several executions of a policy, which is achived by optimizing the policy's scalarized expected return (SER). To achieve fairness over objectives during each execution the expected scalarized return (ESR) of a policy needs to be optimized instead. This paper presents an argument on the necessity of using ESR in the context of fair multi-objective decision-making and proposes the first mono-policy algorithm able to learn efficient decentralized policies while ensuring fairness across objectives under ESR.

## KEYWORDS

Multi-agent learning, Reinforcement learning, Fairness, Multi-objective learning

**ACM Reference Format:**
Farès Chouaki, Aurélie Beynier, Nicolas Maudet, and Paolo Viappiani. 2025. Fairness in Cooperative Multi-agent Multi-objective Reinforcement Learning using the Expected Scalarized Return: 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

Solving real-world sequential decision-making problems often requires balancing multiple conflicting objectives while coordinating several agents. However, existing reinforcement learning based solutions typically simplify the problem to a single-agent or single-objective setting, making them inadequate for multi-agent, multi-objective scenarios. Furthermore, to be deployable, solutions must not only address these complexities but also guarantee ethical values such as fairness at every execution, not just on average across multiple runs. For instance consider the problem where a fleet of agents need to deliver a limited amount of resources to households of different types (see Figure 1). The goal is to learn decentralized policies that maximize delivery efficiency while ensuring that each type of household receives the same proportion of the available resources during each execution. To address this gap, this paper explains why optimizing the Expected Scalarized Return (ESR) of agents can be better suited for fair multi-objective decision-making, and introduces a novel algorithm for the multi-agent case. Evaluation results demonstrate that the algorithm balances efficiency and fairness during each policy execution.



*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).

## 2 BACKGROUND AND NOTATIONS

The MO-DEC-POMDP framework models multi-objective cooperative multi-agent reinforcement learning problems. It extends the DEC-POMDP model [6] and is a special case of the MO-POSG framework [10].Supported by [5] we use the Nash Social Welfare (NSW) function to balance fairness and efficiency. This scalarization function can be applied in two ways within MORL [9]:

- **Scalarized Expected Return (SER):** Computes the policy value by first averaging rewards over time and then applying the scalarization function. Thus, the value $V_u^\pi$ of a policy $\pi$ under this criterion is given by: $V_u^\pi = u\left(\mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^t r_i \mid \pi, s_0\right]\right)$.
- **Expected Scalarized Return (ESR):** Applies the scalarization function to individual returns before averaging. The value $V_u^\pi$ of a policy $\pi$ under ESR is therefore given by: $V_u^\pi = \mathbb{E}\left[u\left(\sum_{i=0}^{\infty}\gamma^t r_i\right) \mid \pi, s_0\right]$.

When the scalarization function is non-linear, these criteria yield different optimal policies [3, 9, 10].

## 3 RELATED WORK

Our approach solves Multi-Objective Multi-Agent Reinforcement Learning (MOMARL) problems using a utility-based approach. This section reviews existing solutions in the field.

Roijers et al. [8] proposed the Expected Utility Policy Gradient (EUPG) algorithm, which learns an optimal policy under ESR with a non-linear utility function. EUPG extends the policy gradient method by conditioning the policy on both the accrued return and

the state and including the accrued return in the computation of the loss. EUPG was extended by [7] using an actor-critic method that learns a multivariate distribution over returns. Extensions of the Monte Carlo tree search algorithm were proposed by [2] to learn policies under ESR. Hu et al. [4] introduced the first multipolicy algorithm for cooperative MOMARL under SER. Current MOMARL algorithms cannot solve cooperative tasks with known utilities, and MORL solutions overlook ESR when considering fairness. This paper addresses these limitations by proposing a decentralized algorithm that ensures fairness and efficiency under ESR.

## 4 FAIRNESS WITH EXPECTED SCALARIZED RETURN

This section demonstrates the need to differentiate between policies ensuring fairness under SER and ESR. We show that algorithms optimized for SER can lead to unfair policies, and demonstrate how ESR optimization addresses this issue.

**Example 1.** Consider the MO-MDP [3] with initial state $s_0$, actions $a_0$, $a_1$, and $a_2$, and terminal states $t_1$, $t_2$, $t_3$. The transition function is given by Table 1 and the reward function by Table 2 with $\alpha, \epsilon \in \mathbb{R}_+^*$ and $\alpha > \epsilon$. We compare deterministic policies $\pi_1$ and $\pi_2$, where $\pi_1(s_0) = \pi_2(s_0) = a_0$, $\pi_1(s_1) = a_1$, and $\pi_2(s_1) = a_2$. Under SER, $\pi_2$ is preferred ($\pi_2 \succ \pi_1$), but under ESR, $\pi_1$ is preferred ($\pi_1 \succ \pi_2$).

We argue that $\pi_1$ is the fairest policy in the example above. It

| $s$ | $a$ | $s'$ | $Pr$ |
|---|---|---|---|
| $s_0$ | $a_0$ | $s_1$ | 1 |
| $s_1$ | $a_1$ | $t_1$ | 1 |
| $s_1$ | $a_2$ | $t_2$ | 0.5 |
| $s_1$ | $a_2$ | $t_3$ | 0.5 |

**Table 1: Transition function of the MO-MDP of example 1**

| $s$ | $a$ | $s'$ | $r$ |
|---|---|---|---|
| $s_0$ | $a_0$ | $s_1$ | $(0, \alpha)$ |
| $s_1$ | $a_1$ | $t_1$ | $(\alpha, 0)$ |
| $s_1$ | $a_2$ | $t_2$ | $(0, \alpha)$ |
| $s_1$ | $a_2$ | $t_3$ | $(2\alpha, \epsilon - \alpha)$ |

**Table 2: Reward function of the MO-MDP of example 1**

achieves fairer returns during a single execution and remains fair on average across multiple executions, while $\pi_2$ only achieves fairness on average. We also note that a fair policy optimized for SER can be riskier[2] and the importance of conditioning the agent by the past returns for ESR optimization. Moreover, in the MO-MDP example, the returns from action $a_1$ are Pareto-dominated by $a_2$, but using NSW under ESR prefers $\pi_1$. This shows that NSW does not satisfy Pareto optimality under ESR, unlike under SER [5]. We argue that first-order stochastic dominance and its extension proposed by [1] are more suitable dominance properties for ESR optimization.

**Example 2.** Consider the multi-objective matrix game in Table 3. We compare policy $\pi_1$ that always selects joint action $(b, b)$, with policy $\pi_2$, selecting joint actions $(a, b)$ and $(b, a)$ with equal probability. Under SER, $\pi_2 \succ \pi_1$, while under ESR, $\pi_1 \succ \pi_2$. This difference shows that, in both stochastic and deterministic settings, distinguishing between ESR and SER is crucial when the agents' policies are stochastic and scalarization function is non-linear.

## 5 DECENTRALIZED EXPECTED UTILITY POLICY GRADIENT(DEC-EUPG)

We propose a decentralized policy-gradient algorithm to learn fair distributed policies under ESR. Inspired by [8], each agent applies

| | $a$ | $b$ |
|---|---|---|
| $a$ | $(1, 1)$ | $(0, 11)$ |
| $b$ | $(11, 0)$ | $(5, 5)$ |

**Table 3: MO matrix game considered in Example 2**

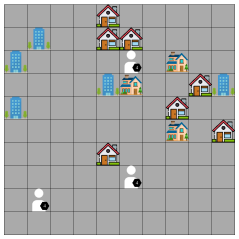

**Figure 1: Environment with 3 agents and 3 objectives**

EUPG independently, treating other agents as part of the environment. The policy of agent $i$ $\pi_i^{\theta_i}$ is conditioned on its observation history $h_t^i$ and the global accumulated return $\mathbf{G}_t^- = \sum_{k=0}^{t} \gamma^k \mathbf{R}_k$. Policies update after each episode using the update rule given by Equation 1,

$$\theta_i \leftarrow \theta_i + \alpha \gamma^t u(\mathbf{G}_t^- + \mathbf{G}_t^+) \nabla_{\theta_i} \ln \pi_{\theta_i}(a_t^i | h_t^i, \mathbf{G}_t^-) \quad (1)$$

where $a_t^i$ is the agent's action and $\mathbf{G}_t^+$ its future reward. However, a key limitation of this approach is reliance on global accumulated returns, often unavailable during execution. To address this, we propose a variant where, instead, the agent's policy is conditionned by its local accumulated return $\mathbf{G}_t^{i-} = \sum_{k=0}^{t} \gamma^k \mathbf{R}_k^i$.

We evaluate our approach on the delivery task described in Section 1. This task allows to test the validity and scalability of the proposed algorithm. Since no existing algorithm solves such a task, we compare our method to a centralized agent controlling all agents and a decentralized mono-agent mono-objective approach which takes a problem with $n$ objectives and $n$ agents and decomposes into $n$ mono-objective mono-agent problems by assigning to each agent a unique objective to optimize. To ensure a fair comparison between our approach and these baselines, the EUPG algorithm was used to train the centralized baseline and the policy gradient algorithm was used to train the decomposition baseline.

Results show that Dec-EUPG solves cooperative MOMARL problems while ensuring fairness. The local-reward variant proved to be more efficient than the global-reward variant while achieving the same fairness. It makes our algorithms more practical and suitable for deployment in real-world applications eventhough this approach is only feasible when global rewards can be transformed into local ones. We argue that the proposed solutions are most suited for applications where fairness needs to be guaranteed at each execution such as ethical applications of MORL.

## 6 CONCLUSION

This paper highlights a flaw in existing fair MORL algorithms: current approaches ensure fairness only over multiple policy executions but fail to guarantee fairness across objectives for singlular executions of the policy. To address this, we proposed Dec-EUPG, an algorithm for solving cooperative multi-agent multi-objective tasks while balancing fairness and efficiency. Experiments on delivery tasks show it matches centralized fairness while learning decentralized policies and outperforming decomposition-based approaches. However, challenges remain, including sample efficiency, credit assignment, and deployment when local rewards are infeasible.

## REFERENCES


[1] Conor Hayes, Timothy Verstraeten, Diederik Roijers, Enda Howley, and Patrick Mannion. 2022. Expected scalarised returns dominance: a new solution concept for multi-objective decision making. *Neural Computing and Applications* (07 2022). https://doi.org/10.1007/s00521-022-07334-x

[2] Conor F. Hayes, Mathieu Reymond, Diederik M. Roijers, Enda Howley, and Patrick Mannion. 2023. Monte Carlo tree search algorithms for risk-aware and multi-objective reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 37, 2 (April 2023), 37. https://doi.org/10.1007/s10458-022-09596-0

[3] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2022. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems* 36, 1 (April 2022). https://doi.org/10.1007/s10458-022-09552-y

[4] Tianmeng Hu, Biao Luo, Chunhua Yang, and Tingwen Huang. 2023. MO-MIX: Multi-Objective Multi-Agent Cooperative Decision-Making With Deep Reinforcement Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 12098–12112. https://doi.org/10.1109/TPAMI.2023.3283537

[5] Debmalya Mandal and Jiarui Gan. 2022. Socially fair reinforcement learning. *arXiv preprint arXiv:2208.12584* (2022).

[6] Frans A Oliehoek, Christopher Amato, et al. 2016. *A concise introduction to decentralized POMDPs*. Vol. 1. Springer.

[7] Mathieu Reymond, Conor Hayes, Denis Steckelmacher, Diederik Roijers, and Ann Nowe. 2023. Actor-critic multi-objective reinforcement learning for non-linear utility functions. *Autonomous Agents and Multi-Agent Systems* 37 (04 2023). https://doi.org/10.1007/s10458-023-09604-x

[8] Diederik M. Roijers, Denis Steckelmacher, and Ann Nowé. 2020. Multi-objective reinforcement learning for the expected utility of the return. 2018 Adaptive Learning Agents, ALA 2018 - Co-located Workshop at the Federated AI Meeting, FAIM 2018 ; Conference date: 14-07-2018 Through 15-07-2018.

[9] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. 2013. A Survey of Multi-Objective Sequential Decision-Making. *Journal of Artificial Intelligence Research* 48 (Oct. 2013), 67–113. https://doi.org/10.1613/jair.3987

[10] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2019. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34, 1 (Dec. 2019). https://doi.org/10.1007/s10458-019-09433-x