Towards Fair and Efficient Policy Learning in Cooperative Multi-Agent Reinforcement Learning

Extended Abstract

Umer Siddique
University of Texas San Antonio
San Antonio, USA
muhammadumer.siddique@my.utsa.edu

Peilang Li
University of Texas San Antonio
San Antonio, USA
peilang.li@my.utsa.edu

Yongcan Cao
University of Texas San Antonio
San Antonio, USA
yongcan.cao@utsa.edu

ABSTRACT

In this paper, we consider the problem of learning independent fair policies in cooperative multi-agent reinforcement learning (MARL). Our objective is to design multiple policies simultaneously that optimize a welfare function for fairness. To achieve this objective, we propose a novel Fairness-Aware multi-agent Proximal Policy Optimization (FAPPO) algorithm, which enables each agent to learn its policy independently while optimizing a welfare function. Unlike standard approaches that focus on maximizing performance metrics such as rewards, FAPPO focuses on fairness in an independent learning setting, where each agent estimates its local value function. Furthermore, when inter-agent communication is allowed, we introduce an attention-based FAPPO (AT-FAPPO), which incorporates a self-attention mechanism to facilitate communication and coordination among agents. This variant allows agents to share relevant information during training, leading to more fair outcomes. To demonstrate the effectiveness of the proposed methods, we perform experiments in various environments and show that our approach outperforms existing methods both in terms of efficiency and equity.

KEYWORDS

Multi-agent reinforcement learning; Fair optimization; Welfare functions

ACM Reference Format:
Umer Siddique, Peilang Li, and Yongcan Cao. 2025. Towards Fair and Efficient Policy Learning in Cooperative Multi-Agent Reinforcement Learning: 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

Recent advances in reinforcement learning (RL) and multi-agent RL (MARL) have significantly improved the abilities of adaptive artificial agents to cooperate and solve complex tasks, including autonomous vehicles [4, 11], traffic light control [15], data center control [17], and wireless networks [8]. Despite the diverse applications of these systems, their primary focus has been mainly on optimizing a single performance metric. However, this singular focus on performance optimization often neglects the consideration of fairness, particularly in scenarios where these systems impact multiple end-users. Hence, fairness becomes a key factor for the deployment and operation of such systems if we want the users to trust and use the systems.

Fairness is a multifaceted concept, often framed through lenses such as Pareto dominance, equity, symmetry, demographic parity, and proportionality. In this work, we adopt a definition of fairness grounded in distributive justice [7], emphasizing Pareto efficiency, equity, and symmetry (impartiality). This definition of fairness can be encoded into a welfare function that aggregates the utilities of users/agents and provides a principled evaluation and comparison of solutions based on fairness.

Several studies have focused on incorporating fairness into multi-agent systems [1, 3], and MARL [5, 6, 19, 20]. However, these methods either focus on static environments that do not require learning or require a specialized network architecture that learns individual policies before optimizing fairness via a centralized leader or team policy. In contrast, our proposed methods do not rely on a specialized network architecture or a hierarchical structure. Instead, we propose a fairness-aware multi-agent Proximal Policy Optimization (FAPPO), an extension to the independent PPO (IPPO) [16], that learns individual policies for all agents separately in the context of cooperative MARL and optimizes a welfare function to ensure equitable treatment for all agents. When inter-agent communication is available, we propose an attention-based variant of FAPPO (AT-FAPPO) by incorporating a self-attention mechanism [2, 14] for communication.

2 FAIRNESS FORMULATION

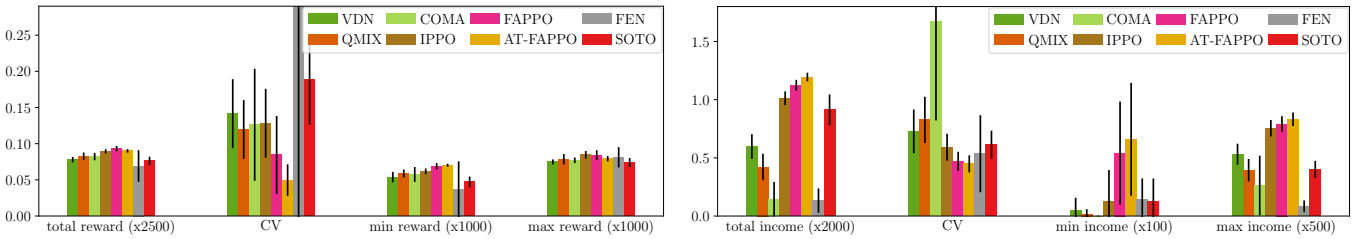
Following the prior work on fairness in RL [10, 12, 18], we define a fair solution as one that satisfies three properties: *Pareto-dominance*, *equity*, and *impartiality*. These properties ensure that a solution is Pareto-optimal, aligns with the *Pigou-Dalton principle*, and adheres to the "equal treatment of equals" principles. To make these properties operational, we employ the *generalized Gini welfare function* which satisfies these fairness properties and provides a principled approach for fair optimization:

$$\phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{\omega}_i \boldsymbol{x}_i^{\uparrow}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{\omega} \in \Delta_n$ is a fixed positive weight vector whose components are strictly decreasing (i.e., $\boldsymbol{\omega}_1 > \ldots > \boldsymbol{\omega}_n > 0$). Intuitively, by assigning larger weights to smaller utility values, this welfare function will yield larger scores when the utility distribution becomes more balanced.

This work is licensed under a Creative Commons Attribution International 4.0 License.

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).



(a) Random MDP. **(b) Matthew effect.**

Figure 1: Total rewards, CV, minimum, and maximum reward of MARL baselines, FEN, SOTO, and our proposed methods.

3 PROPOSED METHOD

We consider fully cooperative MARL tasks, where a set of agents cooperate to solve a given task. Our optimization objective can be formulated as

$$\max_{\boldsymbol{\pi_\theta}} \phi_{\boldsymbol{\omega}}(\boldsymbol{J}(\boldsymbol{\pi_\theta})), \tag{2}$$

where $\boldsymbol{\pi_\theta}$ represents the joint policy for all agents parameterized by $\boldsymbol{\theta}$, $\boldsymbol{J}(\boldsymbol{\pi_\theta}) = \mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$ denotes the joint expected discounted returns, and $\phi_{\boldsymbol{\omega}}$ is the welfare function. The objective is to maximize the welfare utility over the joint policy $\boldsymbol{\pi_\theta}$.

To solve the problem (2), we adapt the Independent Proximal Policy Optimization (IPPO) [16] to optimize the welfare function $\phi_{\boldsymbol{\omega}}$ and refer to it as FAPPO. In FAPPO, each agent learns its policy independently using PPO [9], relying solely on local observations. Since our method learns stochastic policies, we can optimize the welfare function $\phi_{\boldsymbol{\omega}}$ by computing gradients using a variant of the policy gradient theorem to update the policies as

$$\nabla_{\boldsymbol{\theta}} \phi_{\boldsymbol{\omega}}(\boldsymbol{J}(\boldsymbol{\pi_\theta})) = \nabla_{\boldsymbol{J}(\boldsymbol{\pi_\theta})} \phi_{\boldsymbol{\omega}}(\boldsymbol{J}(\boldsymbol{\pi_\theta}))^\top \cdot \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\pi_\theta}) = \boldsymbol{w}_\sigma^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\pi_{\boldsymbol{\theta}}),$$

where $\nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\pi_\theta})$ is a $n \times D$ matrix representing the joint policy gradient over the n agents, $\boldsymbol{w}_\sigma$ is a vector sorted based on the values of $\boldsymbol{J}(\boldsymbol{\pi_\theta})$, and D denotes the number of policy parameters.

Interestingly, in the independent learning setting, $\boldsymbol{J} = (J^1(\pi_\theta), \ldots J^n(\pi_\theta))$, where J^a is the utility of agent a. Thus, our optimization problem (2) can be expressed as

$$\max_{\boldsymbol{\pi_\theta}} \phi_{\boldsymbol{\omega}}(J^1(\pi_{\theta_1}), \ldots J^n(\pi_{\theta_n})),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is the policies parameters $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^n)$ respectively. Using the policy gradient theorem [13], the gradient of the utility function $J^a(\pi_\theta)$ for each agent a can be computed as,

$$\nabla_\theta J^a(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[A^a_{\pi_\theta}(z^a, u^a) \nabla_\theta \log \pi_\theta(u^a \mid z^a)\right], \tag{3}$$

where A^a is the advantage estimation for the agent a. As we are in an independent learning setting, we estimate the advantage for each agent as $\sum_t (\gamma\lambda)^{t-1}\delta^a_t$, where $\delta^a_t = r_t(z^a_t, u^a_t) + \gamma V_\theta(z^a_{t+1}) - V_\theta(z^a_t)$. We use the team reward $r_t(s_t, \boldsymbol{u}_t)$ as the per-time-step reward $r_t(z^a_t, u^a_t)$ of agent a for approximation. $V_\theta(z^a_t)$ denotes the value function associated with agent a with local observation z_t, and λ represents the temporal difference estimation of the advantage function. Finally, for each agent a, the clipping objective becomes

$$\mathbb{E}_{z^a_t \sim \rho\pi, u^a_t \sim \pi\theta(\cdot|z^a_t)}\left[\min(\rho_\theta A^a_{\pi_\theta}(u^a_t|z^a_t), \bar{\rho}_\theta A^a \pi_\theta(u^a_t|z^a_t))\right],$$

where $\rho_\theta = \dfrac{\pi_\theta(u^a_t|z^a_t)}{\pi_{\theta_{\text{old}}}(u^a_t|z^a_t)}$, $\bar{\rho}_\theta = \text{clip}(\rho_\theta, 1-\epsilon, 1+\epsilon)$, $\pi_{\theta_{\text{old}}}$ represents the policy generating the transitions.

Furthermore, when inter-agent communication is allowed, we introduce AT-FAPPO, which integrates a self-attention mechanism [14]. This mechanism enables agents to share information during learning, further improving fairness and coordination.

4 EXPERIMENTAL RESULTS

To validate the efficacy of our proposed methods, we performed experiments in two different environments. Our first environment is a Random MDP, a grid-world-based multi-agent environment where three agents navigate a 5×2 grid with a fully random transition function. Our second environment is Matthew effect [5] which contains 10 pac-man agents and 3 ghosts, where an agent's size and speed increase with the number of ghosts it consumes, leading to inherent imbalance.

For a comprehensive performance evaluation of our proposed methods, we compare our methods against state-of-the-art MARL baselines. Figure 1a shows the experimental results of MARL baselines and our proposed methods in a Random MDP environment. Interestingly both fair MARL baselines, FEN and SOTO, perform worse, which is likely because of the fact they are designed for environments where neighbors' information is necessary to learn fair optimal solutions. On the other hand, independent learning algorithms, including IPPO and our proposed methods, perform well with less hyperparameter tuning. Notably, FAPPO and AT-FAPPO outperform all baselines in terms of total rewards and *Coefficient of variation* (CV). A lower CV indicates reduced variability in agent rewards, as confirmed by the minimum reward metric, where only our proposed methods manage to significantly increase the minimum agent reward, thereby establishing a more balanced reward distribution among all agents.

Figure 1b depicts the experimental results for the Matthew effect environment. Once again, independent learning algorithms perform better than other baselines. FAPPO and AT-FAPPO maximize total income, showing they can achieve an efficient solution while simultaneously minimizing CV and maximizing the minimum agent income, thus resulting in a fair solution.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research under Grant N000142412405 and the Army Research Office under Grant W911NF2310363.

REFERENCES

[1] Haris Aziz, Anna Bogomolnaia, and Hervé Moulin. 2019. Fair mixing: the case of dichotomous preferences. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. Association for Computing Machinery, New York, NY, USA, 753–781.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). https://api.semanticscholar.org/CorpusID:11212020

[3] Aurélie Beynier, Yann Chevaleyre, Laurent Gourvès, Julien Lesca, Nicolas Maudet, and Anaëlle Wilczynski. 2018. Local Envy-Freeness in House Allocation Problems. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 292–300.

[4] Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. 2012. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics* 9, 1 (2012), 427–438.

[5] Jiechuan Jiang and Zongqing Lu. 2019. *Learning fairness in multi-agent systems*. Curran Associates Inc., Red Hook, NY, USA.

[6] Peizhong Ju, Arnob Ghosh, and Ness Shroff. 2023. Achieving Fairness in Multi-Agent MDP Using Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

[7] Hervi Moulin. 1988. *Axioms of Cooperative Decision Making*. Cambridge University Press.

[8] Navid Naderializadeh, Jaroslaw J Sydir, Meryem Simsek, and Hosein Nikopour. 2021. Resource management in wireless networks via multi-agent deep reinforcement learning. *IEEE Transactions on Wireless Communications* 20, 6 (2021), 3507–3523.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 http://arxiv.org/abs/1707.06347

[10] Umer Siddique, Abhinav Sinha, and Yongcan Cao. 2023. Fairness in Preference-based Reinforcement Learning. *arXiv preprint arXiv:2306.09995* (2023).

[11] Umer Siddique, Abhinav Sinha, and Yongcan Cao. 2024. On Deep Reinforcement Learning for Target Capture Autonomous Guidance. In *AIAA SCITECH 2024 Forum*. 0957.

[12] Umer Siddique, Paul Weng, and Matthieu Zimmer. 2020. Learning Fair Policies in Multi-Objective (Deep) Reinforcement Learning with Average and Discounted Rewards. In *International Conference on Machine Learning*.

[13] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in neural information processing systems*.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[15] Marco A Wiering et al. 2000. Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*. 1151–1158.

[16] C. S. D. Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge? *ArXiv* abs/2011.09533 (2020). https://api.semanticscholar.org/CorpusID:227054146

[17] Zhiyuan Yao, Zihan Ding, and Thomas Clausen. 2022. Multi-agent reinforcement learning for network load balancing in data center. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3594–3603.

[18] Guanbao Yu, Umer Siddique, and Paul Weng. 2023. Fair Deep Reinforcement Learning with Preferential Treatment. In *ECAI*.

[19] Chongjie Zhang and Julie A. Shah. 2014. Fairness in multi-agent sequential decision-making. In *Advances in neural information processing systems*.

[20] Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. 2021. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*.