Fairness in Preference-based Reinforcement Learning

Umer Siddique [1] **Abhinav Sinha** [1] **Yongcan Cao** [1]

Abstract

In this paper, we address the issue of fairness in preference-based reinforcement learning (PbRL) in the presence of multiple objectives. The main objective is to design control policies that can optimize multiple objectives while treating each objective fairly. Toward this objective, we design a new fairness-induced preference-based reinforcement learning or FPbRL. The main idea of FPbRL is to learn vector reward functions associated with multiple objectives via new *welfare-based* preferences rather than *reward-based* preference in PbRL, coupled with policy learning via maximizing a generalized Gini welfare function. Finally, we provide experiment studies on three different environments to show that the proposed FPbRL approach can achieve both efficiency and equity for learning effective and fair policies.

1. Introduction

The broad application of reinforcement learning (RL) faces a significant challenge, namely, the design of appropriate reward functions that align with specific mission objectives in given environments. To mitigate this challenge, preference-based RL (PbRL) (see, for example, (Christiano et al., 2017)) has emerged as a promising paradigm, leveraging human feedback to eliminate the need for manual reward function design. However, real-world missions often entail multiple objectives and the consideration of preferences among diverse users, necessitating a balanced approach. Existing PbRL methods primarily focus on maximizing a single performance metric, neglecting the crucial aspect of equity or fairness, e.g., (Stiennon et al., 2020; Wu et al., 2021; Lee et al., 2021). Consequently, the lack of fairness considerations poses a barrier to the widespread deployment of PbRL for systems affecting multiple end-users when it is critical to address fairness among these users.

To address this critical gap, the development of methods enabling fairness in PbRL becomes imperative. While recent advancements have explored fairness in RL, albeit not within the PbRL framework, notable contributions in, e.g., (Weng, 2019; Siddique et al., 2020; Fan et al., 2022), have employed welfare functions to ensure fairness in the single-agent RL setting. Furthermore, the work in (Zimmer et al., 2021) considered fairness in a multi-agent RL setting.

This paper proposes an approach that builds upon existing studies on fairness, focusing on a PbRL setting. In particular, rather than relying on known ground truth rewards, our method involves learning fair policies by incorporating fairness directly into the PbRL paradigm, thereby eliminating the need for hand-crafted reward functions. By doing so, we aim to address fairness in PbRL without compromising on its advantages.

Contributions. In this paper, we present a novel approach that addresses fairness in PbRL. Our proposed method introduces a novel technique to learn vector rewards associated with multiple objectives by leveraging welfare-based preferences rather than reward-based preferences in (Christiano et al., 2017). Hence, the proposed approach provides new insights and techniques to address fairness in PbRL. We validate the effectiveness of our approach through comprehensive experiments conducted in three real-world domains. The proposed approach is expected to provide solutions for RL problems when reward functions are absent, or it is too costly to design them.

2. Related Work

The concept of having equity and fairness, especially in real-world missions with multiple objectives and diverse users, is imperative. Such a concept has also been given careful consideration in many domains, including economics (Moulin, 2004b), political philosophy (Rawls, 2020), applied mathematics (Brams & Taylor, 1996) operations research (Bäuerle & Ott, 2011), and theoretical computer science (Ogryczak et al., 2014). Fairness considerations have been incorporated into classic continuous and combinatorial optimization problems in scenarios where the underlying model was assumed to be fully known, and learning might not be necessary (Nei-

[1]Unmanned Systems Lab, Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, 78249, USA. Correspondence to: Umer Siddique <muhammadumer.siddique@my.utsa.edu>.

The Many Facets of Preference Learning Workshop at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

dhardt et al., 2008; Ogryczak et al., 2013; Nguyen & Weng, 2017; Busa-Fekete et al., 2017; Agarwal et al., 2018). Such methods include linear programming and other model-based algorithms that consider the feedback effects and dynamic impacts in decision-making processes, allowing for the development of fair policies that adapt to changing circumstances. While such methods yielded satisfactory results, they cannot be directly used if the underlying model is unknown or too complex to be modeled.

The study of fairness in RL, especially within a model-free paradigm, has gained significant attention in recent years, with notable contributions shedding light on various aspects of this emerging field. Initial work by (Jabbari et al., 2017) laid the foundation by focusing on scalar rewards, paving the way for further advancements. Researchers have pursued diverse directions to incorporate fairness into RL frameworks. (Wen et al., 2021) explored fairness constraints as a means to reduce discrimination, while the work of (Jiang & Lu, 2019; Zimmer et al., 2021; Ju et al., 2023) delved into achieving fairness among agents. The work of Siddique et al. (2020) introduced a novel fair optimization problem within the context of multi-objective RL, enabling modifications to the existing deep RL algorithms to ensure fair solutions. Chen et al. (2021) extended the scope by incorporating fairness into actor-critic RL algorithms, optimizing general fairness utility functions for real-world network optimization problems. The work of Zimmer et al. (2021), on the other hand, focused on fairness in decentralized cooperative multi-agent settings, developing a framework involving self-oriented and team-oriented networks concurrently optimized using a policy gradient algorithm. Notably, the work in Ju et al. (2023) introduced online convex optimization methods as a means to learn fairness with respect to agents.

Despite the significant successes achieved in the field of deep RL, these methods heavily rely on the availability of known reward functions. However, in many real-world problems, the task of defining a reward function is often challenging and sometimes even infeasible. To address this limitation, PbRL has emerged as an active area of research (Christiano et al., 2017). Within PbRL, different settings have been explored, depending on whether the involvement of humans is direct or if simulated human preferences are derived from the ground truth rewards. In the context of PbRL, the standard approach typically revolves around maximizing a single criterion, such as a reward, which is inferred from the preferences (Stiennon et al., 2020; Lee et al., 2021; Wu et al., 2021). However, it is clear that focusing exclusively on maximizing rewards falls short of assuring fairness across various objectives. Our approach, which is consistent with the fundamental concepts of preference-based learning, digs into the investigation of learning fair policies in the context of PbRL.

3. Preliminaries

3.1. Preference-based RL (PbRL)

We consider a Markov Decision process without reward (MDP\R) augmented with preferences, which is a tuple of the form $(\mathcal{S}, \mathcal{A}, T, \rho, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of possible actions, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a state transition probability function specifying the probability $p(s' \mid s, a)$ of reaching state $s' \in \mathcal{S}$ after taking action a in state s, γ is a discount factor, and $\rho : \mathcal{S} \to [0, 1]$ specifies the initial state distribution. The learning agent interacts with the environment through rollout trajectories, where a length-k trajectory segment takes the form $(s_1, a_1, s_1, a_1, \ldots, s_k, a_k)$. A *policy* π is a function that maps states to actions, such that $\pi(a \mid s)$ is the probability of taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$.

PbRL is an approach to learning policies without rewards in which humans are asked to compare pairs of trajectories and give relative preferences between them (Christiano et al., 2017). More specifically, in PbRL, a human is asked to compare a pair of length-k trajectory segments $\sigma^1 = (s_1^1, a_1^1, s_1^2, a_1^2, \ldots, s_k^1, a_k^1)$ and $\sigma^2 = (s_1^2, a_1^2, s_2^2, a_2^2, \ldots, s_k^2, a_k^2)$, where $\sigma^1 \succ \sigma^2$ indicates that the user preferred σ^1 over σ^2. Owing to the unavailability of the reward function, many PbRL algorithms learn an estimated reward function model, $\hat{r}(\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The reward estimate $\hat{r}(\cdot, \cdot)$ can be viewed as an underlying latent factor explaining human preferences. In particular, it is often assumed that the human's probability of preferring a segment σ^1 over σ^2 is given by the Bradley-Terry model (Christiano et al., 2017),

$$P(\sigma^1 \succ \sigma^2 \mid \hat{r}) = \frac{e^{\hat{R}(\sigma^1)}}{e^{\hat{R}(\sigma^1)} + e^{\hat{R}(\sigma^2)}}, \tag{1}$$

where $\hat{R}(\sigma_i) := \sum_{t=1}^{k} \gamma^{t-1} \hat{r}(s_t^i, a_t^i)$ is the estimated total discounted reward of trajectory segment σ_i, and (s_t^i, a_t^i) is the t^{th} state-action pair in σ_i. One can minimize the cross-entropy loss between the Bradley-Terry preference predictions and true human preferences, given by (Christiano et al., 2017),

$$L(\hat{r}) = -\sum_{(\sigma^1, \sigma^2, \mu) \in S} \big(\mu(1) \log P[\sigma^1 \succ \sigma^2] + \mu(2) \log P[\sigma^2 \succ \sigma^1]\big), \tag{2}$$

where $\mu(i),\ i \in \{1, 2\}$ is an indicator such that $\mu(i) = 1$ when trajectory segment σ^i is preferred, whereas S is the dataset of labeled human preferences. By optimizing $L(\hat{r})$, an estimated reward function $\hat{r}(\cdot, \cdot)$ can be obtained to help explain human preferences.

3.2. Notion of Fairness

The fairness concept used in previous work such as (Speicher et al., 2018; Weng, 2019; Siddique et al., 2020; Zim-

mer et al., 2021) enforces three natural properties: *efficiency*, *equity*, and *impartiality*. The concept of efficiency, also referred to as *optimality*, implies that the solution should be optimal and Pareto dominant. Equity is often associated with the concept of distributive justice, as it pertains to the fairness of resource or opportunity distribution. This property ensures that a fair solution follows the Pigou-Dalton principle (Moulin, 2004a), which states that by transferring rewards from the more advantaged to the less advantaged users, the overall fairness of the solution can be improved. Impartiality or equality requires that all users be treated equally, without favoritism towards any particular user in terms of the solution's outcomes.

To operationalize this notion of fairness, the use of welfare functions is employed. These welfare functions aggregate the utilities of all users and provide a measure of the overall desirability of a solution for the entire group. While there exist various welfare functions, we only consider those that satisfy the three fairness properties discussed earlier. One welfare function that satisfies the aforementioned properties is the *generalized Gini welfare function* (Weymark, 1981), which is defined as follows:

$$\phi_{\boldsymbol{w}}(\boldsymbol{u}) = \sum_{i \in \mathcal{K}} \boldsymbol{w}_i \boldsymbol{u}_i^{\uparrow}\,, \tag{3}$$

where $\boldsymbol{u} \in \mathbb{R}^{\mathcal{K}}$ represents the utility vector of a size $\mathcal{K}$, $\boldsymbol{w} \in \mathbb{R}^{\mathcal{K}}$ is a fixed weight vector with positive components that strictly decrease (i.e., $\boldsymbol{w}_1 > \ldots > \boldsymbol{w}_K$), and $\boldsymbol{u}^{\uparrow}$ denotes the vector obtained by sorting the components of $\boldsymbol{u}$ in increasing order (i.e., $\boldsymbol{u}_1^{\uparrow} \leq \ldots \leq \boldsymbol{u}_K^{\uparrow}$). For consistency, bold variables represent vectors/matrices. In essence, this function computes the summation of the weight multiplied by the sorted utility for each objective. The weight vector is fixed, positive, and strictly decreasing. It is important to note that the strict decrease in weights is crucial to ensure a fair and Pareto optimal, as well as an equitable solution.

4. Approach

In order to account for the impact of an agent's actions on multiple objectives, i.e., users in the notion of fairness in Section 3.2, we extend previous RL formulations by redefining the estimated reward function as a vector function, denoted as $\hat{\boldsymbol{r}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\mathcal{K}}$, where $\mathcal{K}$ denotes the number of objectives. This vector function captures the rewards associated with all objectives, acknowledging the multi-objective nature of the problem at hand. Note that this is different from the scalar reward function $\hat{r}$ in PbRL (Christiano et al., 2017). To formalize the fair policy optimization problem, we integrate the welfare function $\phi_{\boldsymbol{w}}$ into our objective function. Consequently, the goal is to find a policy that generates a fair distribution of rewards over $\mathcal{K}$ objectives given by

$$\max_{\pi_{\boldsymbol{\theta}}} \phi_{\boldsymbol{w}}(\boldsymbol{J}(\pi_{\boldsymbol{\theta}})), \tag{4}$$

where $\pi_{\boldsymbol{\theta}}$ represents a policy parameterized by $\boldsymbol{\theta}$, $\phi_{\boldsymbol{w}}$ denotes a welfare function with fixed weights that requires optimization, and $\boldsymbol{J}(\pi_{\boldsymbol{\theta}})$ represents the vectorial objective function that yields the utilities (i.e., $\boldsymbol{u}$) for all users. It is also worth noting that the chosen welfare function, such as the generalized Gini welfare function, is concave. As a result, the optimization problem presented in (4) can be characterized as a convex optimization problem. This convexity property facilitates the exploration of effective solution methods for achieving equitable policies in model-free RL settings.

Note that optimizing the welfare function defined in (3) is an effective way to address fairness because the weights $\boldsymbol{w}$ are selected such that a higher weight will be assigned for objectives with lower utility values, which will ensure that all objectives are treated fairly than the cases when the weights are assigned without considering the utility values.

Our procedure to optimize the welfare function is an iterative process that integrates the policy update step and reward update step (via the collection of more preferences for reward function estimation). Since the reward function estimation is non-stationary, we focus on policy gradient methods. As a state-of-the-art policy gradient method, we adopt the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) for policy optimization and compute the advantage function via

$$\boldsymbol{A}_{\pi_{\boldsymbol{\theta}}}(s_t, a_t) = \sum_t (\gamma\lambda)^{t-1}\delta_t\,, \tag{5}$$

where δ_t is determined by the expression $\hat{\boldsymbol{r}}_t + \gamma \boldsymbol{V}_{\boldsymbol{\theta}}(s_{t+1}) - \boldsymbol{V}_{\boldsymbol{\theta}}(s_t)$, with $\hat{\boldsymbol{r}}_t$ representing the estimated rewards, and $\boldsymbol{V}_{\boldsymbol{\theta}}(s_t)$ denoting the value function associated with state s_t. In PPO, the objective function $\boldsymbol{J}(\boldsymbol{\theta})$ is designed to limit policy changes after an update, that is,

$$\mathbb{E}_{s\sim \boldsymbol{d}_{\pi}, a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\min(\rho_{\boldsymbol{\theta}}\boldsymbol{A}_{\pi_{\boldsymbol{\theta}}}(s,a), \bar{\rho}_{\boldsymbol{\theta}}\boldsymbol{A}_{\pi_{\boldsymbol{\theta}}}(s,a))\right]\,, \tag{6}$$

where $\rho_{\boldsymbol{\theta}} = \dfrac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_{\boldsymbol{b}}(a|s)}$, $\bar{\rho}_{\boldsymbol{\theta}} = \text{clip}(\rho_{\boldsymbol{\theta}}, 1-\epsilon, 1+\epsilon)$, $\pi_{\boldsymbol{b}}$ represents the policy generating the transitions, and ϵ is a hyperparameter controlling the constraint. To compute the gradient for $\boldsymbol{J}(\boldsymbol{\theta})$, we have

$$\nabla_{\boldsymbol{\theta}}\phi_{\boldsymbol{w}}(\boldsymbol{J}(\pi_{\boldsymbol{\theta}})) = \nabla_{\boldsymbol{J}(\pi_{\boldsymbol{\theta}})}\phi_{\boldsymbol{w}}(\boldsymbol{J}(\pi_{\boldsymbol{\theta}})) \cdot \nabla_{\boldsymbol{\theta}}\boldsymbol{J}(\pi_{\boldsymbol{\theta}}) \tag{7}$$

$$= \boldsymbol{w}_{\sigma}^{\intercal}\nabla_{\boldsymbol{\theta}}\boldsymbol{J}(\pi_{\boldsymbol{\theta}}), \tag{8}$$

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

For reward estimation function update, we ask a human (or a similar mechanism like a synthetic human) to provide preferences for the segments collected by the policy, establishing or expanding the dataset for preferences. The vector

function $\hat{\boldsymbol{r}}$ is learned via minimizing the loss function (2) with a modified preference probability given by

$$P(\sigma^1 \succ \sigma^2 \mid \hat{\boldsymbol{r}}) = \frac{e^{\hat{R}(\sigma^1)}}{e^{\hat{R}(\sigma^1)} + e^{\hat{R}(\sigma^2)}}, \tag{9}$$

where $\hat{R}(\sigma_i) := \phi_{\boldsymbol{w}}(\sum_{t=1}^{k} \gamma^{t-1} \hat{\boldsymbol{r}}(s_t^i, a_t^i))$. This formulation applies the welfare function $\phi_{\boldsymbol{w}}$ to the discounted cumulative vector rewards, resulting in a scalarized $\hat{R}(\sigma_i)$. This scalarized value is then utilized to compute $P(\sigma^1 \succ \sigma^2 \mid \hat{\boldsymbol{r}})$. It is important to note that the key distinction between our proposed approach and PbRL in (Christiano et al., 2017) lies in the utilization of the welfare function to determine preferences, as opposed to relying on segment rewards as done in (Christiano et al., 2017).

5. Experimental Results

To demonstrate the robustness and practicality of our method, we meticulously design and conduct three experiments. Each experiment showcases a unique scenario where fairness plays a pivotal role in RL outcomes. Moreover, at present, our primary emphasis is directed toward investigating synthetic human preferences owing to their convenient acquisition process and their appropriateness for testing objectives. Nonetheless, it is essential to note that our proposed approach is readily applicable in situations that involve human-in-the-loop interactions. Through rigorous analysis and evaluation, we assess the performance of our approach, both in terms of achieving fairness objectives and maintaining desirable learning outcomes in a model-free setting. We assign weights $\boldsymbol{w}_i = \frac{1}{2^i}, i = 0, ..., \mathcal{K} - 1$, and to ensure the reproducibility of the results, and average the results over $5+$ runs with different seeds to provide reliable evidence of our method's effectiveness. All algorithm hyperparameters were optimized using the open-source Lightweight HyperParameter Optimizer (LHPO) (Zimmer, 2018).

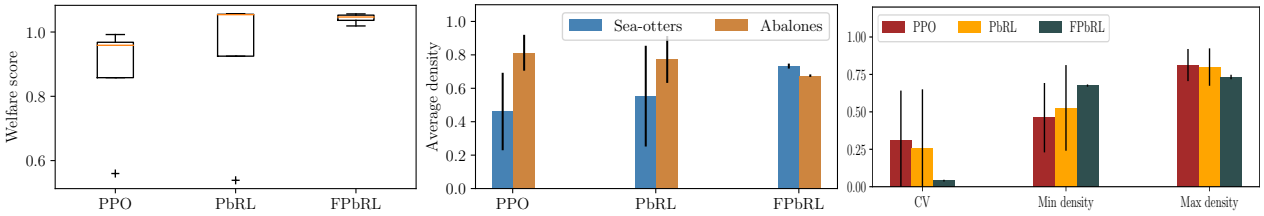
5.1. Species Conservation

Species conservation is a critical domain in the field of ecology, particularly when dealing with the preservation of multiple interacting endangered species. Here, we tackle the challenge of incorporating fairness considerations into the conservation efforts of two specific species: sea otters and their prey, the northern abalone. The sea otter and northern abalone populations face a delicate balance as sea otters consume abalones, both of which are currently endangered. To navigate this complex conservation problem, we adopt the setting proposed in Chadès et al. (2012) and tailor it to address the fairness aspects of this ecosystem. In our conservation problem, the state is defined by the current population numbers of both species. To influence the system, we have five distinct actions at our disposal: introducing sea otters, enforcing antipoaching measures, controlling sea otter populations, implementing a combination of half-antipoaching and half-controlled sea otters, or taking no action. Each action has significant implications, as introducing sea otters is necessary for balancing the abalone population, but if not carefully managed, it can inadvertently drive the abalone species to extinction. Similarly, neglecting any of the other managerial actions would result in the extinction of one of the species, highlighting the importance of a comprehensive approach in terms of equity and fairness. The transition function in this conservation problem incorporates population growth models for both species, accounting for factors such as poaching and oil spills. Through this framework, we strive to optimize not just a single objective but the population densities of both species, thereby dealing with a multidimensional problem where two objectives, sea otter and abalone population densities, need to be simultaneously optimized, leading to $\mathcal{K} = 2$.

In this domain, our primary objective is to assess the effectiveness of our proposed method in optimizing the welfare function, denoted as $\phi_{\boldsymbol{w}}$. To evaluate this, we conduct a comparative analysis of welfare scores between three approaches: PPO, PbRL, and our proposed FPbRL method within this domain. To compute the welfare scores, we employ trained agents and evaluate their performance across 100 trajectories within the given environment. The empirical average vector returns of these trajectories serve as the basis for deriving the welfare score by applying the function $\phi_{\boldsymbol{w}}$. The distribution of welfare scores for PPO, PbRL, and FPbRL is shown in Figure 1a. Our results reveal that FPbRL achieves the highest welfare score, thereby demonstrating its ability to identify fairer solutions compared to PPO and the standard PbRL method. However, recognizing that the welfare score alone may not provide a comprehensive understanding of the objective balance, we present individual density plots in Figure 1b depicting the densities of both species. These plots offer further insights into the distribution of objectives. Consistently, our findings demonstrate that FPbRL yields more balanced solutions in terms of equity, surpassing both PbRL and PPO. In addition, we introduce the Coefficient of Variation (CV) to address scenarios where demonstrating the utility of each objective becomes challenging due to a multitude of objectives. Figure 1c showcases the CV, as well as the minimum and maximum densities. Corresponding with our previous findings, our proposed FPbRL method exhibits the lowest CV, indicating reduced variation between different objectives. Moreover, our method prioritizes maximizing the minimum objective to foster a more equitable distribution of utilities.

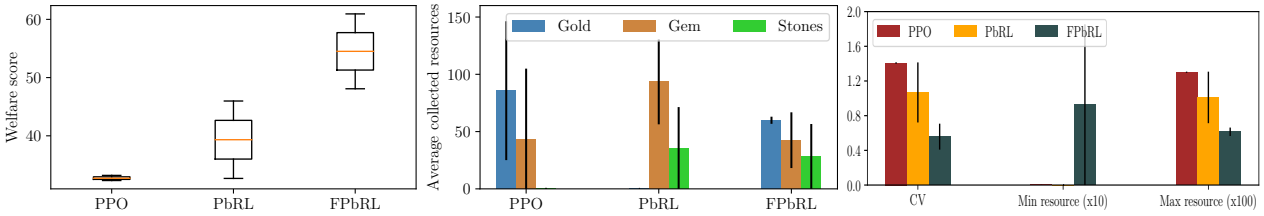
5.2. Resource Gathering

We now consider a resource-gathering environment that encompasses a 5×5 grid world, adapted from the work of (Barrett & Narayanan, 2008). This dynamic environ-



(a) Welfare scores. (b) Individual densities. (c) CV, min and max densities.

Figure 1. Performances of PPO, PbRL, FPbRL in the species conservation problem.

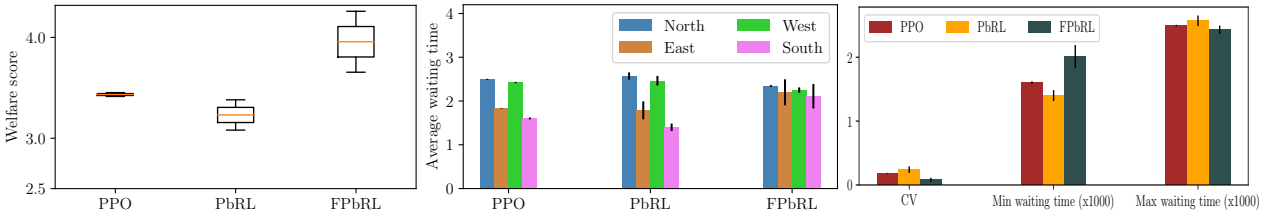


(a) Welfare scores. (b) Individual number of resources collected. (c) CV, min, and max resources collected.

Figure 2. Performances of PPO, PbRL, FPbRL in resource gathering.

ment poses the challenge of resource acquisition, where the agent's objective is to collect three distinct types of resources: gold, gems, and stones, thus $\mathcal{K} = 3$. Within this grid world, the agent is situated at a specific position, while the resources are scattered randomly across various locations. Upon consumption of a resource, it is promptly regenerated at another random location within the grid, ensuring a continuous supply. The state representation in this environment encapsulates the agent's current position within the grid, as well as the cumulative count of each resource type collected throughout the ongoing trajectory. To navigate this complex environment, the agent is equipped with four cardinal direction actions: up, down, left, and right, enabling movement across the grid. However, to introduce an additional layer of intricacy, we assign distinct values to the resources. Gold and gems are endowed with a value of 1, symbolizing their higher significance, while stones, deemed less valuable, are assigned a value of 0.4. This deliberate assignment fosters an unbalanced distribution of resources, with two stones, one gold, and one gem, strategically placed within the grid. Amidst this resource-rich environment, the agent's ultimate goal is twofold: to maximize the accumulation of resources while concurrently maintaining a balanced distribution among the different resource types. By striking this delicate equilibrium, the agent strives to optimize its resource-gathering strategy, maximizing its overall utility and adaptability within this domain.

To demonstrate the efficacy of our proposed approach in maintaining a balanced distribution of resources, we conduct an analysis of welfare scores for the resource collection problem. Through this analysis, we aim to assess the fairness of different approaches and determine the extent to which our proposed method promotes equitable solutions. Figure 2a presents the welfare scores computed for PPO, PbRL, and the proposed FPbRL. These scores were computed over a hundred trajectories during the testing phase. Encouragingly, our proposed method achieved the highest welfare score, signifying a fairer solution when compared to both PPO and the standard PbRL method. To gain a comprehensive understanding of the balance between objectives in resource collection, we also examine the individual number of resources collected (see Figure 2b). Once again, the results reinforce the superiority of FPbRL in producing more balanced solutions. In contrast, PbRL and PPO tend to favor the accumulation of certain resources at the expense of others, highlighting the limitations of a standard approach that solely optimizes the aggregate or weighted sum of objectives. Our proposed method, however, maintains a balanced distribution of different resources, underscoring the significance of fairness considerations in resource collection scenarios. Furthermore, Figure 2c provides additional insights into the performances of PPO, PbRL, and FPbRL by examining the CV as well as the minimum and the maximum number of collected resources. Strikingly, FPbRL outperforms the other algorithms, exhibiting the lowest CV, which indicates a more equitable distribution of objectives. Notably, only FPbRL successfully maximizes the minimum objective utility, whereas PPO and the PbRL method yield the lowest minimum objective values, reflecting a prioritization of maximizing cumulative rewards at the expense of fairness considerations.



(a) Welfare scores. (b) Individual waiting times. (c) CV, min, and max waiting times

Figure 3. Performances of PPO, PbRL, FPbRL in traffic control.

5.3. Traffic Control at Intersections

To thoroughly validate the effectiveness of our proposed method, we also conduct a series of experiments in the demanding real-world domain of traffic light control. This domain presents unique challenges due to the multitude of objectives involved, making it an ideal testbed for evaluating the efficacy of our approach. To simulate a realistic traffic intersection scenario, we employed the widely-used Simulation of Urban MObility (SUMO) platform (Lopez et al., 2018). Specifically, our focus is on a standard 8-lane intersection, with two lanes designated for turning (left or right, depending on the side of the road) and the remaining lanes facilitating straight driving or additional turns. Traditionally, the objective of traffic control is to optimize traffic flow by minimizing the total waiting time for all vehicles approaching the intersection. However, our approach diverges from this conventional perspective. Instead, we adopted a novel viewpoint, aiming to optimize traffic flow for each of the four distinct sides of the road. Each side of the intersection is treated as a separate objective, and our goal is to learn a controller that effectively reduces the expected waiting times for vehicles on each road segment. This multi-objective setup thus takes $\mathcal{K} = 4$, reflecting the four sides of the road that need to be individually optimized. In this challenging problem, a state is defined by several key factors, including the waiting time of vehicles, the car density in the vicinity, and the current phase of the traffic light. The action space comprises four distinct options, each corresponding to a different phase change that influences the traffic flow on a specific side of the road. The transition function governing the evolution of the system is dependent on factors such as the current traffic light phase, the movement of vehicles through the intersection, and the generation of new traffic.

Similar to the previous assessments, we evaluate the efficacy of the proposed method in optimizing the welfare function. The welfare scores obtained during testing for PPO, PbRL, and FPbRL are presented in Figure 3a. To improve readability, the y-axis has been scaled by a factor of 1000, with each tick representing 1000 units. It is evident that FPbRL outperforms both PPO and PbRL, achieving the highest welfare score. This noteworthy result underscores the efficacy of FPbRL in optimizing the welfare function, which is crucial for ensuring fair and equitable treatment of the diverse objectives at hand. To establish the correlation between these high welfare scores and fairer solutions, we examine the waiting times for all sides of the roads, as depicted in Figure 3b. Our proposed method, FPbRL, demonstrates a more balanced distribution of waiting times across all road segments. Although FPbRL exhibits slightly higher total waiting times, it prioritizes lanes with fewer cars, thereby preventing any single vehicle from enduring significantly prolonged waiting periods. In contrast, PPO and PbRL tend to favor lanes with higher car densities in their pursuit of minimizing total waiting times. This observation underscores the importance of fairness considerations, indicating that the attainment of fairness may sometimes come at a cost. However, the cost of fairness is not excessively high, as evidenced in the previous domains (Figures 1b and 2b). Furthermore, we compare the performances of PPO, PbRL, and FPbRL in terms of the CV, minimum waiting time, and maximum waiting time (Figure 3c). Once again, FPbRL emerges as the top-performing algorithm, attaining the lowest CV and achieving a more balanced distribution of objectives. Notably, only FPbRL successfully maximizes the minimum objective and minimizes the maximum objective, thereby promoting equitable outcomes in the context of traffic light control.

6. Conclusions and Future Work

By incorporating fairness into PbRL, we developed a new fairness-induced PbRL (FPbRL) approach that can provide more equitable and socially responsible RL systems. Through our multi-experiment validation, we provided compelling evidence of the effectiveness and practicality of our approach toward its potential applications in real-world scenarios where fairness considerations are imperative. Our findings underscore the effectiveness of our proposed method, FPbRL, in optimizing the welfare function and achieving fairness in the presence of multiple objectives. A detailed investigation of other welfare functions and different impartiality properties, along with actual human feedback, could be interesting to explore in the future.

Acknowledgements

The authors were supported in part by Army Research Lab under grant W911NF2120232, Army Research Office under grant W911NF2110103, and Office of Naval Research under grant N000142212474.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Barrett, L. and Narayanan, S. Learning all optimal policies with multiple criteria. In *Proceedings of the 25th international conference on Machine learning*, pp. 41–47, 2008.

Bäuerle, N. and Ott, J. Markov decision processes with average value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

Brams, S. J. and Taylor, A. D. *Fair Division: From Cake-Cutting to Dispute Resolution*. March 1996.

Busa-Fekete, R., Szörényi, B., Weng, P., and Mannor, S. Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, pp. 625–634. PMLR, 2017.

Chadès, I., Curtis, J. M., and Martin, T. G. Setting realistic recovery targets for two interacting endangered species, sea otter and northern abalone. *Conservation Biology*, 26 (6):1016–1025, 2012.

Chen, J., Wang, Y., and Lan, T. Bringing fairness to actor-critic reinforcement learning for network utility optimization. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2021.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Fan, Z., Peng, N., Tian, M., and Fain, B. Welfare and fairness in multi-objective reinforcement learning. *arXiv preprint arXiv:2212.01382*, 2022.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. In *International conference on machine learning*, pp. 1617–1626. PMLR, 2017.

Jiang, J. and Lu, Z. Learning Fairness in Multi-Agent Systems. In *Advances in neural information processing systems*, 2019.

Ju, P., Ghosh, A., and Shroff, N. B. Achieving fairness in multi-agent markov decision processes using reinforcement learning. *arXiv preprint arXiv:2306.00324*, 2023.

Lee, K., Smith, L., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., and Wießner, E. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018.

Moulin, H. *Fair Division and Collective Welfare*. MIT Press, 2004a.

Moulin, H. *Fair division and collective welfare*. MIT press, 2004b.

Neidhardt, A., Luss, H., and Krishnan, K. R. Data fusion and optimal placement of fixed and mobile sensors. In *2008 IEEE Sensors Applications Symposium*, February 2008.

Nguyen, V. H. and Weng, P. An efficient primal-dual algorithm for fair combinatorial optimization problems. In *International Conference on Combinatorial Optimization and Applications*, pp. 324–339. Springer, 2017.

Ogryczak, W., Perny, P., and Weng, P. A compromise programming approach to multiobjective markov decision processes. *International Journal of Information Technology & Decision Making*, 12(05):1021–1053, 2013.

Ogryczak, W., Luss, H., Pióro, M., Nace, D., and Tomaszewski, A. Fair optimization and networks: A survey. *Journal of Applied Mathematics*, 2014, 2014.

Rawls, J. A theory of justice. In *A theory of justice*. Harvard university press, 2020.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Siddique, U., Weng, P., and Zimmer, M. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, 2020.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, 2018.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Wen, M., Bastani, O., and Topcu, U. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR, 2021.

Weng, P. Fairness in reinforcement learning. In *AI for Social Good Workshop at International Joint Conference on Artificial Intelligence*, 2019.

Weymark, J. Generalized Gini inequality indices. *Mathematical Social Sciences*, 1:409–430, 1981.

Wu, J., Ouyang, L., Ziegler, D. M., Stiennon, N., Lowe, R., Leike, J., and Christiano, P. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

Zimmer, M. *Apprentissage par renforcement developpemental*. PhD thesis, University of Lorraine, January 2018.

Zimmer, M., Glanois, C., Siddique, U., and Weng, P. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2021.