Fair Preference-based Reinforcement Learning with Noisy Preferences

Anonymous authors
Paper under double-blind review

Abstract

Preference-based reinforcement learning (PbRL) offers a promising approach for learning policies from human preferences. However, the presence of multiple objectives and noisy preferences lead to significant challenges in ensuring fairness and robustness. In this paper, we address the problem of learning fair policies in multi-objective PbRL with noisy feedback. Our main objective is to design control policies that optimize multiple objectives while treating each objective equitably, even when the feedback contains imperfections and irrationalities. To achieve this, we design three different fairness-enhanced PbRL methods tailored to handle myopic, mistaken, and perturbed preference feedback. Our proposed methods optimize a generalized Gini welfare function to learn fair policies. We evaluate the performance of our methods in various environments. Our results demonstrate that the proposed methods effectively learn fair policies by creating a trade-off between efficiency and fairness.

1 Introduction

In reinforcement learning (RL), an agent interacts with an environment and attempts to make optimal decisions through a process of trial and error. RL has demonstrated remarkable success in various domains, including games (Mnih et al., 2015; Silver et al., 2016), robotics (Peters et al., 2003; Gu et al., 2017), and autonomous vehicles (Cao et al., 2012; Siddique et al., 2024). However, the success of these methods often relies on careful reward function design. In practice, designing an optimal reward function is challenging. For example, in real-world applications such as autonomous vehicles, crafting suitable reward functions or measuring success is difficult. In this context, preference-based RL (PbRL) has emerged as a promising alternative. PbRL leverages human feedback, eliminating the need for manual reward function design, and uses preferences instead of reward functions to reflect human opinions on target objectives.

Despite its advantages, existing work in PbRL, under both ideal and noisy feedback has primarily focused on maximizing a single performance objective, such as behavior alignment in large language model (LLM) (Ouyang et al., 2022), high-quality image in image generation (Lee et al., 2023), and task efficiency in continuous control problems (Christiano et al., 2017). However, real-world missions often involve multiple objectives and the consideration of preferences among diverse users, necessitating a balanced approach. Existing PbRL methods, which primarily focus on maximizing single performance metrics, neglect the crucial aspect of equity or fairness (Stiennon et al., 2020; Wu et al., 2021; Lee et al., 2021a; Kaufmann et al., 2023; Ouyang et al., 2022; Lee et al., 2023). Consequently, the lack of fairness considerations poses a barrier to the widespread deployment of PbRL for systems affecting multiple end-users, where fairness among these users is critical.

Addressing this critical gap under noisy feedback is imperative. Several papers have explored fairness in RL (Weng, 2019; Siddique et al., 2020; Fan et al., 2022; Yu et al., 2023; 2024) by optimizing welfare functions to ensure fairness in the single agent RL setting. A recent work (Siddique et al., 2023) proposed FPbRL to learn fair policies in PbRL; however, this method relies on high-quality feedback, typically assuming expertise from the teachers. Human feedback is often prone to errors (Christiano et al., 2017; Lee et al., 2021a), and in broader applications, feedback is frequently sourced from non-expert users or crowd-sourcing platforms, where quality can be inconsistent and noisy.

In this paper, we investigate how noisy feedback, specifically myopic, mistaken, and perturbed feedback, affects learning fair policies in PbRL settings. We propose three methods: FPbRL-Myopic, FPbRL-Mistake, and FPbRL-Noisy. Similar to the FPbRL (Siddique et al., 2023), our methods learn vector-estimated rewards from teacher feedback. However, in contrast to FPbRL, we learn the reward predictor from noisy feedback, thereby eliminating the need for handcrafted reward functions and high-quality feedback. By doing so, we aim to address fairness in PbRL settings under noisy feedback without compromising its advantages while also providing a realistic approach to preference elicitation.

In summary, our contributions are twofold: First, we incorporate various preference elicitation methods that better mirror irrationality, myopic tendencies, deviations, and perturbed feedback to learn vector rewards associated with multiple objectives by leveraging generalized Gini welfare-based preferences rather than reward-based preferences in (Christiano et al., 2017). Second, we validate the effectiveness of our approach through extensive experiments conducted in three real-world domains, demonstrating that even in the presence of noisy feedback, the proposed methods are able to learn fair solutions.

2 Background

We briefly recall standard PbRL, its multi-objective variant, and our fairness formulation. For a more detailed background on PbRL and fairness, we refer the readers to (Christiano et al., 2017) and, (Siddique et al., 2023) respectively.

2.1 Multi-objective PbRL

In this paper, we consider the multi-objective Markov Decision Process without reward (MOMDP\R) with multiple objectives, namely, multi-objective PbRL. In multi-objective PbRL, the agent learns an estimated reward function that is represented as a vector, where each component of the vector corresponds to the individual utility of a user in our context. Therefore, the reward function can be defined as $\hat{\boldsymbol{r}}(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\mathcal{K}}$, where $\mathcal{K}$ is the number of objectives (users). This differs from traditional PbRL, where the reward is scalar. The agent interaction loop in PbRL remains the same, as the learning agent interacts with the environment through rollout trajectories, where a length-k trajectory segment takes the form $(s_1, a_1, s_1, a_1, \ldots, s_k, a_k)$. The learning agent then tries to learn 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, a human is asked to compare a pair of length-k trajectory segments $\sigma^1 = (s_1^1, a_1^1, s_2^1, a_2^1, \ldots, s_k^1, a_k^1)$ 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. Since the reward function is unavailable, the agent learns an estimated reward function model, $\hat{\boldsymbol{r}}(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\mathcal{K}}$. The reward estimate $\hat{\boldsymbol{r}}(\cdot,\cdot)$ 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),

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

where $\hat{R}(\sigma_i) := \phi_{\boldsymbol{\omega}}(\sum_{t=1}^{k} \gamma^{t-1} \hat{\boldsymbol{r}}(s_t^i, a_t^i))$ is the welfare utility of total discounted vector reward of a trajectory segment σ_i, and (s_t^i, a_t^i) is the t^{th} state-action pair in σ_i., and $\phi_{\boldsymbol{\omega}}$ is our chosen welfare function (see Section 2.2) with $\boldsymbol{\omega}$ weights. One can minimize the cross-entropy loss between the Bradley-Terry preference predictions and true human preferences, given by (Christiano et al., 2017),

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

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 $L(\hat{\boldsymbol{r}})$, an estimated reward function $\hat{\boldsymbol{r}}(\cdot,\cdot)$ can be obtained to help explain human preferences.

In multi-objective optimization, the usual goal is to find all *Pareto non-dominated* solutions (Roijers et al., 2013). However, this approach may not work for large-scale decision-making problems, where the number of *Pareto non-dominated* solutions may grow exponentially with the size of the problem (Perny et al., 2013). Thus, to make a single decision at a given time step, as expected in PbRL, we employ aggregation methods (Llamazares, 2013; Hayes et al., 2022). These methods effectively transform the multi-objective optimization problem into a single objective. Given our objective of finding balanced and, by extension, fair solutions, we employ aggregation methods that are fair (Weymark, 1981).

2.2 Generalized Gini Welfare Function

The fairness concept used in previous work such as (Speicher et al., 2018; Weng, 2019; Siddique et al., 2020; Zimmer et al., 2021) enforces three properties: *efficiency*, *equity*, and *impartiality*. The concept of *efficiency*, states that the solution should be optimal and Pareto dominant. *Equity* refers to the fair distribution of resources or opportunities and follows the Pigou-Dalton principle (Moulin, 2004), which states that transferring rewards from the more advantaged to the less advantaged users can enhance the overall fairness of the solution. *Impartiality* requires that all users be treated equally, without favoritism towards any particular user in the outcome of the solution.

To make this notoin of fairness operational, we employed generalized Gini welfare function (Weymark, 1981):

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

where $\boldsymbol{u} \in \mathbb{R}^{\mathcal{K}}$ represents the utility vector of a size $\mathcal{K}$, $\boldsymbol{\omega} \in \mathbb{R}^{\mathcal{K}}$ is a fixed weight vector with positive components that strictly decrease (i.e., $w_1 > \ldots > w_{\mathcal{K}}$), and $\boldsymbol{u}^{\uparrow}$ denotes the vector obtained by sorting the components of $\boldsymbol{u}$ in increasing order (i.e., $u_1^{\uparrow} \leq \ldots \leq u_{\mathcal{K}}^{\uparrow}$).

3 Preference Elicitation

In PbRL, the goal is to leverage teacher feedback in the form of preferences to guide the learning agent toward better decision-making. However, gathering and leveraging such feedback may be far from ideal, especially when the expert is a human. In practice, the human factor may introduce elements of irrationality and error, potentially causing biases in the learning process. While prior literature (Christiano et al., 2017; Stiennon et al., 2020; Wu et al., 2021; Lee et al., 2021b) has investigated the concept of a rational teacher, relying exclusively on such idealized entities for evaluating preference-based agents may prove infeasible in fair decision-making. Teacher preferences are susceptible to irrational influences, e.g., cognitive biases and inconsistencies, that must be considered. In this context, we simulate these imperfections by intentionally incorporating a broad spectrum of *irrational behaviors* and *biases*. The aim is to promote resilience and flexibility in the agent's performance as it learns fair decision-making using preferences.

Building upon the insights of (Lee et al., 2021b), our method encompasses myopic (short-sighted) and mistake-prone behaviors, supplemented by the inclusion of noisy behavior in preferences. The concept of noisy behavior, in fact, holds broader implications than mere mistakes. Mistakes often manifest as epsilon-greedy approaches, whereas the introduction of noise, e.g., Gaussian noise, can substantially alter feedback preferences. The positioning of Gaussian noise within the center of probability accumulation could lead to significant shifts in preference feedback. These distinct

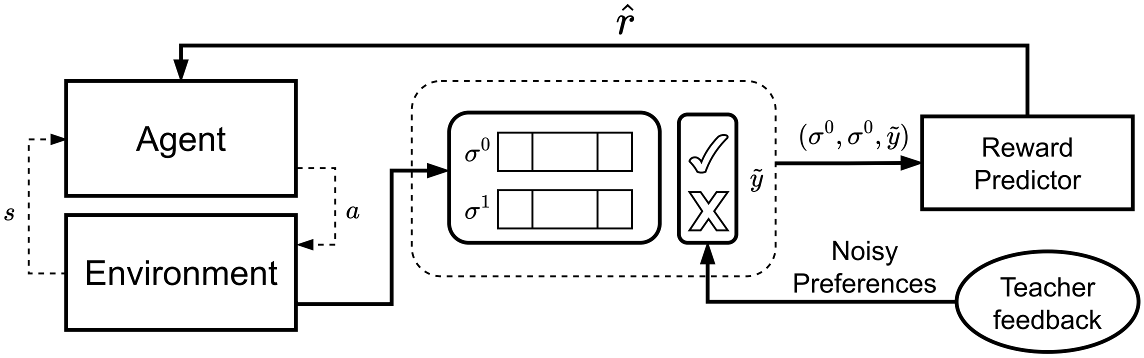


Figure 1: Overview of FPbRL with noisy feedback, where the reward is trained asynchronously from comparisons of trajectory segments and it predicts the vector reward $\hat{\boldsymbol{r}}$.

behavioral dimensions form the foundation of our investigation. Below, we briefly describe the various preference elicitation methods considered in our paper.

Rational (Ideal, Synthetic) Preferences: Rational behavior generates preferences based on actual rewards, (1), and are characterized by deterministic decision-making.

Myopic Tendency: We encapsulate human inclination to emphasize recent observations in the agent's myopic behavior. When expressing preferences, humans often prioritize recent segments of a trajectory. Our model replicates this tendency by assigning greater weight to recent timesteps, expressed as $\sum_{t=1}^{K} \gamma^{K-t} \hat{r}(s_t, a_t)$.

Deviation and Mistake: Acknowledging the inadvertent errors and suboptimal choices in human judgment, we mirror this behavior by introducing a probabilistic inversion of perfect rational preferences. Through this mechanism, a probability of ϵ leads to a $(1, 0)$ preference of $P(\sigma^1 \succ \sigma^2 \mid \hat{r})$, while $(0, 1)$ preferences are attributed to $1 - \epsilon$ probability.

Perturbed Feedback: Unlike previous studies, our exploration incorporates the previously unconsidered domain of noisy behavior. This dimension enhances realism beyond mistakes and myopia, capturing a spectrum of diverse noisy behaviors that include mood swings, fatigue, or external distractions. Notably, slight Gaussian noise centered around the preference probability range of $[0.4, 0.6]$, could dramatically reshape the agent's behavior during fair decision-making, a facet overlooked in (Lee et al., 2021b).

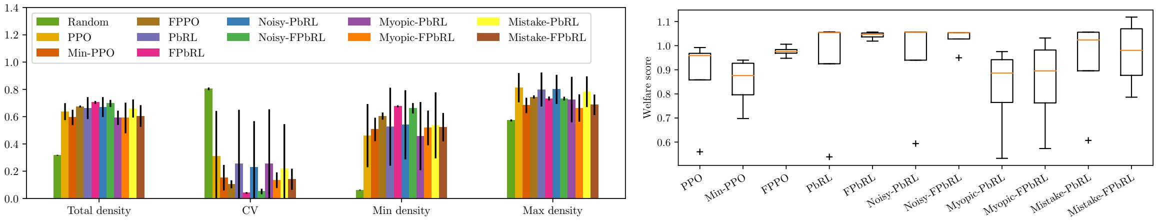
4 Approach

Our approach aligns with FPbRL (Siddique et al., 2023) in that we aim to learn fair policies in a multi-objective PbRL. However, we extend this approach by incorporating noisy feedback rather than relying solely on ideal preferences. At a high level, we solve the following optimization problem:

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

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 $\boldsymbol{J}(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \hat{\boldsymbol{r}}_t\right]$ represents the vectorial objective function that yields the utilities (i.e., $\boldsymbol{u}$) for all users.

Our overall noisy FPbRL framework is illustrated in Figure 1. To estimate $\boldsymbol{J}$, we rely on Proximal Policy Optimization (PPO) (Schulman et al., 2017). During the pretraining phase, a random policy is employed to interact with the environment and to collect preference feedback. This feedback is then used to update the reward model and subsequently refine the policy, which is used to gather further preference feedback. Preferences are solicited from a teacher, who may be rational (ideal, synthetic) or irrational. For synthetic preferences, the feedback is directly proportional to the rewards, aligning



(a) Total density, CV, minimum, and maximum density.

(b) Welfare scores.

Figure 2: Performances of PPO, Min-PPO, PbRL, FPbRL with ideal and noisy preference feedback in species conservation.

with ideal or rational behavior. However, irrational behaviors are prone to errors (Christiano et al., 2017). To better reflect behavior that is close to human, we generate noisy preferences influenced by cognitive biases and inconsistencies. These noisy preferences are then used to train our reward predictor, which estimates the vector reward corresponding to each objective.

Once the $\hat{\boldsymbol{r}}$ is available, the agent updates its parameters by computing the gradient of $\boldsymbol{J}$ as follows:

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\boldsymbol{A}_{\pi_{\boldsymbol{\theta}}}(s,a) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a \mid s)\right], \tag{5}$$

where $\boldsymbol{A}_{\pi_{\boldsymbol{\theta}}}(s,a)$ is the multi-objective variant of advantage function. Using (5), the final policy gradient of the objective (4) can be written as:

$$\nabla_{\boldsymbol{\theta}} \phi_{\boldsymbol{\omega}}(\boldsymbol{J}(\pi_{\boldsymbol{\theta}})) = \nabla_{\boldsymbol{J}(\pi_{\boldsymbol{\theta}})} \phi_{\boldsymbol{\omega}}(\boldsymbol{J}(\pi_{\boldsymbol{\theta}})) \cdot \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\pi_{\boldsymbol{\theta}}) = \boldsymbol{w}_{\sigma}^{\intercal} \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\pi_{\boldsymbol{\theta}}), \tag{6}$$

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.

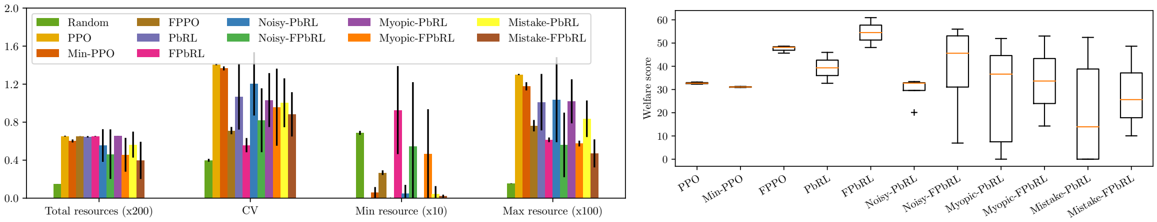
5 Experiments

5.1 Setup

We validate our methods by performing experiments in the same three environments as (Siddique et al., 2023) to ensure a fair comparison. These environments are (i) species conservation, (ii) resource gathering, and (iii) traffic light control. Specifically, in the species conservation environment, the challenge is to incorporate fairness considerations into the conservation efforts of two specific species: sea otters and their prey, the northern abalone. These species face a delicate balance as sea otters consume abalones, and both species are currently endangered. The goal is to maintain the populations of both species equitably. In the resource-gathering environment, the agent collects three different types of resources: gold, gems, and stones. The objective here is to fairly allocate resources to agents, despite the varying values of resources, with stones being the least valuable and diamonds and gold being more valuable. Finally, our last environment is traffic light control, where instead of minimizing the accumulated waiting time in an intersection, the goal is to minimize the waiting time for each side of the road while ensuring fair treatment for all directions.

In all these experiments, we assess the performance of our method in terms of both achieving fairness objectives and maintaining desirable learning outcomes. To ensure that all users/objectives are treated fairly, we use the generalized Gini welfare function with weights $\boldsymbol{w}_i = \frac{1}{2^i}$ for $i = 0, ..., \mathcal{K} - 1$. We ensure the reproducibility of the results by averaging the results over 5+ runs with different random seeds, providing reliable evidence of our method's effectiveness. Each experiment runs for 1 million timesteps. For all algorithms, we optimize hyperparameters using Lightweight HyperParameter Optimizer (LHPO)[1], an open-source library used to run parallel experiments. Our experiments were conducted on a computer equipped with 4 A100 GPUs.

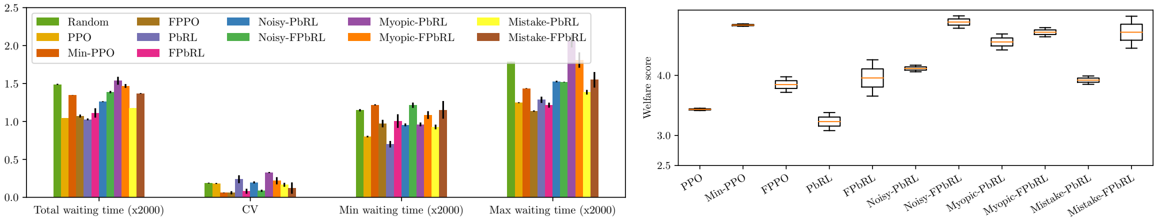
[1] https://github.com/matthieu637/lhpo



(a) Total rewards, CV, minimum, and maximum rewards.

(b) Welfare scores.

Figure 3: Performances of PPO, Min-PPO, PbRL, FPbRL with ideal and noisy preference feedback in resource gathering.



(a) Total waiting time, CV, min, and max waiting time.

(b) Welfare scores.

Figure 4: Performances of PPO, Min-PPO, PbRL, FPbRL with ideal and noisy preference feedback in traffic light control.

5.2 Results

Figures 2 to 4 present the performance of various approaches, including PPO, fairness-enhanced PPO (FPPO) (Siddique et al., 2020), Min-PPO—a multi-objective variant of PPO optimizing exclusively the minimum objective, PbRL, FPbRL, and our noisy versions of PbRL and FPbRL in the species conservation, resource gathering, and traffic light control environments, respectively. We refer our methods as Noisy-FPbRL, Myopic-FPbRL, Mistake-FPbRL. In these environments, we assess our methods based on their Coefficient of Variation (CV), their minimum and maximum objectives value, and their welfare scores to evaluate the effectiveness in optimizing the welfare function. Figures 2a, 3a and 4a display total rewards, CV, minimum and maximum objective values. As expected, the random policy performs worst. PPO and PbRL excel in rewards but have higher CV. Although FPPO also performs well, it should be noted that FPPO assumes true rewards are available and it learns fair solutions by optimizing the Gini welfare function. In contrast, FPbRL algorithms assume that true rewards are not available. Interestingly, in all experiments, our FPbRL algorithms with noisy preferences perform comparably to ideal-preference FPbRL. They achieve lower CV than PbRL, indicating reduced variation among objectives, and effectively maximize the minimum objective for a balanced reward distribution. Finally, Figures 2b, 3b and 4b illustrate welfare scores of all algorithms. Consistent with our previous findings, FPbRL algorithms even with noisy feedback achieve higher welfare scores, demonstrating their ability in optimizing Gini welfare function.

6 Conclusions

In this paper, we present three fair PbRL algorithms designed for reward learning from noisy feedback. Unlike previous works in fairness and PbRL that primarily operate with ideal preference feedback, our approaches address the challenge of noisy feedback. We evaluate the effectiveness and practicality of our methods across three real-world environments where fairness considerations are crucial. Our results demonstrate that our proposed methods can achieve fair and equitable solutions, even when accounting for imperfections in teacher preferences.

References

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Zimeng Fan, Nianli Peng, Muhang Tian, and Brandon Fain. Welfare and fairness in multi-objective reinforcement learning. *arXiv preprint arXiv:2212.01382*, 2022.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation*, pp. 3389–3396. IEEE, 2017.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021a.

Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026*, 2021b.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

B. Llamazares. An analysis of some functions that generalizes weighted means and owa operators. *Intl. J. of Intel. Syst.*, 28:380–393, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Patrice Perny, Paul Weng, Judy Goldsmith, and Josiah Hanna. Approximation of Lorenz-optimal solutions in multiobjective Markov decision processes. In *Association for the Advancement of Artificial Intelligence*, 2013.

Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Reinforcement learning for humanoid robotics. In *Proceedings of the third IEEE-RAS international conference on humanoid robots*, pp. 1–20, 2003.

D.M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

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

Umer Siddique, Abhinav Sinha, and Yongcan Cao. Fairness in preference-based reinforcement learning. *arXiv preprint arXiv:2306.09995*, 2023.

Umer Siddique, Abhinav Sinha, and Yongcan Cao. On deep reinforcement learning for target capture autonomous guidance. In *AIAA SCITECH 2024 Forum*, pp. 0957, 2024.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneerschelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 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.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

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

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

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with preferential treatment. In *ECAI*, 2023.

Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with generalized gini welfare functions. In *Autonomous Agents and Multiagent Systems. Best and Visionary Papers*, pp. 3–29, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-56255-6.

Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2021.