Learning Fair Pareto-Optimal Policies in Multi-Objective Reinforcement Learning

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

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

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

ABSTRACT

Fairness is an important aspect of decision-making in multi-objective reinforcement learning (MORL), where policies must ensure both optimality and equity across multiple, potentially conflicting objectives. While *single-policy* MORL methods can learn fair policies for fixed user preferences using welfare functions such as the *generalized Gini welfare function* (GGF), they fail to provide the diverse set of policies necessary for dynamic or unknown user preferences. To address this limitation, we formalize the fair optimization problem in *multi-policy* MORL, where the goal is to learn a set of Pareto-optimal policies that ensure fairness across all possible user preferences. Our key technical contributions are threefold: (1) We show that for concave, piecewise-linear welfare functions (e.g., GGF), fair policies remain in the *convex coverage set* (CCS), which is an approximated Pareto front for linear scalarization. (2) We demonstrate that non-stationary policies, augmented with accrued reward histories, and stochastic policies improve fairness by dynamically adapting to historical inequities. (3) We propose three novel algorithms, which include integrating GGF with multi-policy multi-objective Q-Learning (MOQL), state-augmented multi-policy MOQL for learning non-statoinary policies, and its novel extension for learning stochastic policies. To validate the performance of the proposed algorithms, we perform experiments in various domains and compare our methods against the state-of-the-art MORL baselines. The empirical results show that our methods learn a set of fair policies that accommodate different user preferences.

KEYWORDS

Multi-objective reinforcement learning, Deep reinforcement learning, Fair optimization, Welfare functions

1 INTRODUCTION

Multi-objective reinforcement learning (MORL) is an important topic in the area of reinforcement learning (RL) that focuses on designing control policies to optimize multiple objectives simultaneously. While traditional MORL methods focus on learning Pareto optimal solutions—ensuring no objective can be improved without sacrificing another—they often neglect fairness, which requires equitable treatment of all objectives or users in our context. For example, in healthcare, a policy may aim to maximize overall patient outcomes (optimality) while ensuring equal treatment across different demographic groups (fairness). A common approach to solving fairness in MORL is to use *utilitarian* welfare functions, where user utilities are aggregated, typically via weighted sum, into a scalarized objective. Despite its simplicity, this approach struggles with fairness, as some users' utilities may be significantly reduced to achieve overall efficiency. An alternative approach is to employ an *egalitarian* welfare function, which prioritizes the least advantaged user by maximizing the minimum utility. While this approach improves fairness, it often leads to inefficient solutions overall, as it optimizes only the lowest utility without ensuring fairness.

Several works have explored fairness in the *single-policy* RL setting [14, 19, 21, 37, 47, 52, 57, 60], where only a single policy is learned. For instance, the work in [52] and [47] enforced fairness by utilizing the generalized Gini social welfare function as a scalarized function and assigning appropriate weights to different objectives to ensure their equitable treatment. Extensions have been explored in multi-agent RL [45, 60] and preferential treatment under known preference weights [57]. Recently, fairness has been studied in multi-policy MORL [17, 32] where Cimpeana et al. [17] defined several fairness notions, while [32] proposed the Lorenz Condition Network (LCN), an extension of the Pareto Conditioned Network (PCN), which trains a policy network in a supervised manner to map states to desired returns. Despite these works, the investigation of fairness in RL still poses some limitations, including (1) learning a *single* fair policy, (2) required knowledge of the welfare function (e.g., scalarized function) with preference weights a prior, and (3) training a conditioning network on specific return targets, limiting their ability to generalize to unseen preferences. Hence, the existing methods operate under fixed/predefined preference weights and cannot be generalized for all possible preferences.

To address these limitations, we propose a novel framework to address fairness in *multi-policy* MORL, rather than the traditional *single-policy* MORL that is the focus of the existing work. Our methods are highly scalable as they leverage a single parameterized network to learn an undominated set of policies, specifically a convex coverage set (CCS), by sampling the entire preference space in MORL. In particular, to address fairness, we apply the welfare function (e.g., GGF) during learning for each sampled preference weight to ensure that each learned policy treats its objectives fairly. We further introduce non-stationary action selection using the state-augmented accrued rewards to enhance fairness by effectively utilizing historical information. We further demonstrate the benefits of learning stochastic policies for fairness. Motivated by hindsight experience replay [3], we incorporate resampling of random preference weights across different preference conditions to improve sample efficiency in MORL, as it is done in [55].

The main contributions of this paper are as follows: (1) We introduce a novel framework for fairness in multi-policy MORL, enabling users to select any fair policy based on their specific preferences, thereby enhancing user satisfaction(Section 3.2). (2) We provide a

Proc. of the Adaptive and Learning Agents Workshop (ALA 2025), Avalos, Aydeniz, Müller, Mohammedalamen (eds.), May 19 – 20, 2025, Detroit, Michigan, USA, ala-workshop.github.io. 2025.

theoretical analysis establishing that for concave, piecewise-linear welfare functions (e.g., GGF), fair policies remain in CCS. Additionally, we demonstrate that non-stationary policies can improve fairness by adapting to historical disparities and that stochastic policies further improve fairness over deterministic policies(Section 4). (3) Building on our theoretical insights, we propose three scalable methods for learning fair policies in MORL using a single parameterized network: (i) an extension to Envelope [55] for learning fair stationary policies, (ii) a non-stationary counterpart that incorporates state-augmented accrued rewards to adaptively improve fairness over time, and (iii) a novel extension for learning stochastic policies, which further enhances fairness(Section 5). (4) We experimentally validate our methods and demonstrate their effectiveness compared to state-of-the-art MORL and fairness methods across three different domains(Section 6).

2 RELATED WORK

Fairness in machine learning (ML) has become a significant research direction [1, 9, 16, 20, 36, 43, 48, 58, 59]. Several studied have addressed fairness in model predictions [49], recommender systems [30], classification [1, 20, 27, 58], and ranking [48]. While much of the literature focuses on the principle of "equal treatment of equals", other aspects, such as proportionality [6] or envy-freeness [15] and its multiple variants (e.g., [7, 11]), have been considered in ML. In contrast, our work is grounded in distributive justice [8, 34, 41], with a focus on optimizing a welfare function for fairness considerations. This principled approach has also been recently advocated in several papers [18, 23, 49].

Recently, fairness in RL has gained significant attention with the work by [24], which ensures fairness in state visitation using scalar rewards. The work of [26], proposed FEN a hierarchical decentralized method using a gossip algorithm to ensure fairness across agents involved in a system. Similarly, [13] proposed to incorporate fairness into actor-critic RL algorithms, optimizing general fairness utility functions for real-world network optimization problems. Considering the multi-objective nature of many RL problems, the study of fairness in multi-objective reinforcement learning (MORL) has been widely studied. In particular, [47] proposed multiple adaptations to deep RL algorithms that optimize the *generalized Gini social welfare.* [44, 60] extended this work to the decentralized co-operative multi-agent setting. [21] proposed to optimize the Nash welfare function using scalarized expected return criterion. [19] proposed a method for generalized Gini welfare function optimization in rankings. [40, 57] proposed methods that learn a fair policy providing preferential treatment to some users while ensuring equal treatment of all others under the assumption that these preferential weights are known in advance. [46] proposed FPbRL, a fairness-enhanced method in preference-based RL to learn fair policies in the absence of true rewards. Recently, fairness has been considered in multi-policy MORL with [32] propose learning Lorenz Condition networks, which ensures fairness through Lorenz domination and adds an extra parameter λ, however, we use the welfare function to learn a set of fair optimal policies.

Despite the significant successes achieved in the field of RL and MORL, existing methods heavily rely on scalarization functions to learn a *single policy* with fixed preference weights. However, such single-policy methods do not work when preferences are unknown or user-specific solutions are required. To address this limitation, several works have been proposed to accommodate user-specific preferences, including but not limited to those proposed by [2, 5, 33, 42, 50, 55]. Notably, these methods aim to learn a set of policies that approximate the Pareto front of optimal solutions. For instance, [5] and [33] proposed methods to compute policies on the Pareto front's convex hull, while [55] introduced envelope Q-learning, learning policies from the CCS. These approaches, however, do not address fairness, which is the focus of this paper.

3 PRELIMINARIES

3.1 Multi-Objective Markov Decision Process

A multi-objective Markov Decision Process (MOMDP) extends the classical Markov Decision Process (MDP) framework to scenarios where an agent must optimize multiple objectives simultaneously. An MDP [39] is a mathematical model commonly used for sequential decision-making problems. Formally, an MDP is defined by a tuple, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions available to the agent, $\mathcal{P}_{a,s,s'} \in [0, 1]$ is the probability of transition from state s to state s' after taking action a, *i.e.*, $\mathcal{P}(s'|s, a) = \mathcal{P}[S_{t+1} = s'|S_t = s, A_t = a]$, $r(s, a) : s \times a \mapsto r$ is the immediate reward obtained by taking action a at state s, and $\gamma \in [0, 1)$ is the discount factor. An MOMDP can be represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \boldsymbol{r}, \gamma, \Omega, f_\Omega)$, in which the definitions of $\mathcal{S}, \mathcal{A}, \mathcal{P}$, and γ are the same as in MDP except that the reward $\boldsymbol{r}$ is now a vector, with each component corresponding to an objective that the agent seeks to optimize. Here, the additional Ω represents the entire space of preferences, and f_Ω is the preference function which takes a linear form, producing a single utility $f_{\boldsymbol{\omega}}(\boldsymbol{r}) = \boldsymbol{\omega}^T \boldsymbol{r}(s, a)$, where $\boldsymbol{\omega}$ is a vector representing the preference weights for different objectives. In MOMDPs, the objectives may be conflicting, and hence it is often difficult to optimize all objectives simultaneously.

The goal of an agent in an MOMDP is to either learn a single policy that balances multiple objectives or a set of policies that optimize different trade-offs among objectives. These approaches are referred to as *single-policy* MORL and *multi-policy* MORL, respectively. A policy π is a strategy that maps states to actions, which can be deterministic (i.e., $\forall s, \pi(s) \in \mathcal{A}$) or stochastic (i.e., $\forall s, a, \pi(a|s)$ denotes the probability of selecting a in s). In MOMDPs, policies are typically *stationary* or *Markovian*, meaning that action selection probabilities depend solely on the current state, irrespective of past states and actions. Conversely, a non-stationary policy $\pi(a|\tau, s)$, also known as an adaptive policy, may depend on the agent's history τ. Standard definitions in MDPs, such as the return $G(\tau)$ and the value functions V or Q, extend naturally to MOMDPs, albeit represented as vectors and matrices respectively. The vector return in an MOMDP is expressed as $\boldsymbol{G}(\tau) = \sum_{t=1}^{\infty} \gamma^{t-1} \boldsymbol{r}_t$, where τ is a trajectory comprising a sequence of states, actions, and rewards following the policy, and $\boldsymbol{r}_t$ is a vector reward obtained at time step t. The state value function of a policy π in an MOMDP is defined as $\boldsymbol{V}^\pi(s) = [V_i^\pi(s)] = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=1}^{\infty} \gamma^{t-1} \boldsymbol{r}_t \mid S_0 = s \right]$, where all operations (addition, product) are applied component-wise.

In MOMDPs, value functions do not offer a complete ordering over the policy space. This means it is possible to encounter scenarios wherein, e.g., $V_i^\pi(s) > V_i^{\pi'}(s)$ for objective i, while

$V_j^{\pi}(s) < V_j^{\pi'}(s)$ for j. Hence, value functions in MOMDPs induce only a partial ordering within the policy space, necessitating additional information into objective prioritization for policy ordering.

Envelope Multi-Objective Q-Learning. The Envelope algorithm [55] learns a CCS by sampling preference weights $\boldsymbol{\omega} \in \Omega$ and optimizing linearly scalarized Q-values: $Q(s, a, \boldsymbol{\omega}) = \boldsymbol{\omega}^T \boldsymbol{Q}(s, a)$, where $\boldsymbol{Q}(s, a) \in \mathbb{R}^D$ is the vector of Q-values for D objectives. The Bellman optimality equation for the Envelope algorithm is: $\boldsymbol{Q}^*(s, a, \boldsymbol{\omega}) = \boldsymbol{r}(s, a) + \gamma \max_{a'} \boldsymbol{\omega}^T \boldsymbol{Q}^*(s', a')$. A single neural network parameterizes $\boldsymbol{Q}(s, a, \boldsymbol{\omega})$ by concatenating $\boldsymbol{\omega}$ to the state s, enabling efficient learning across all preferences. Despite its scalability, Envelope lacks explicit fairness guarantees, as linear scalarization may prioritize dominant objectives.

3.2 Fairness Formulation

In MORL, fairness, rooted in distributive justice [34], is crucial for ensuring equitable distribution of rewards. Prior studies in fair optimization within MORL have primarily focused on learning a *single-policy*, commonly referred to as an average policy [21, 46, 47, 56]. In this paper, we adopt a more inclusive view of fairness, including *efficiency*, *equity*, and *impartiality* to generate fair optimal solutions for user-specific preferences.

Definition 3.1. *Efficiency states that among two solutions, if one solution is (weakly or strictly) preferred by all users, then it should be preferred to the other one, e.g.,* $\boldsymbol{V} \succ \boldsymbol{V'} \Rightarrow \phi(\boldsymbol{V}) > \phi(\boldsymbol{V'})$*, where* $\phi(\boldsymbol{V})$ *is the scalar utility function by using the* ϕ *that specifies the value of a solution.*

The efficiency property specifies that given all else equal, one prefers to increase a user's utility. In the MORL setting, the efficiency property simply means Pareto dominance. More specifically, a solution is considered efficient if it is not dominated by any other solution for all objectives.

Definition 3.2. *For a given pair of solutions* $\boldsymbol{V}, \boldsymbol{V'} \in \mathbb{R}^D$*,* $\boldsymbol{V}$ *weakly Pareto-dominates* $\boldsymbol{V'}$ *if* $\forall i, V_i \geq V_i', \forall i \in \{1, \cdots, D\}$*, where* D *is the total number of objectives. Besides,* $\boldsymbol{V}$ *Pareto-dominates* $\boldsymbol{V'}$ *if* $V_i \geq V_i', \forall i$ *and* $\exists j, V_j > V_j'$*. For brevity, we denote Pareto dominance as* $\geq$ *for the weak form and* $>$ *for the strict form.*

Essentially, a solution $\boldsymbol{V}$ (weakly) Pareto-dominates another solution $\boldsymbol{V'}$ if the former's value $\phi(\boldsymbol{V})$ (weakly) Pareto-dominates that of the latter $\phi(\boldsymbol{V'})$. A solution $\boldsymbol{V}^*$ is said to be *Pareto-optimal* if no other solution $\boldsymbol{V}$ Pareto-dominates it. *Pareto front* ($\mathcal{F}$) is defined as the set of Pareto-optimal solutions, which may consist of infinitely many solutions, especially when policies can be stochastic. A typical way to approximate ($\mathcal{F}$) is to compute the CCS, defined below.

Definition 3.3. *A solution in CCS has a maximal scalarized value in a weighted sense if there exists a weight vector* $\boldsymbol{\omega} \in \Omega$ *such that the scalarized utility* $\boldsymbol{\omega}^T \boldsymbol{V}$ *is weakly preferred to the scalarized utility* $\boldsymbol{\omega}^T \boldsymbol{V'}$ *for all other solutions* $\boldsymbol{V'}$ *in the Pareto front. Formally speaking,* $\boldsymbol{V} \in CCS \iff \exists\, \boldsymbol{\omega} \in \Omega \text{ s.t. } \boldsymbol{\omega}^T \boldsymbol{V} \geq \boldsymbol{\omega}^T \boldsymbol{V'}, \forall\, \boldsymbol{V'} \in \mathcal{F}$.

Next, we discuss the significance of the *equity* property, a stronger property than efficiency and often associated with distributive justice, as it refers to the fair distribution of resources or opportunities. This property ensures that a fair solution follows the *Pigou-Dalton principle* [34], which states the transferring of rewards from more advantaged users to less advantaged users.

Definition 3.4. *A solution satisfies the Pigou-Dalton principle if for all* $\boldsymbol{V}, \boldsymbol{V'}$ *equal except for* $V_i = V_i' + \delta$ *and* $V_j = V_j' - \delta$ *where* $V_i' - V_j' > \delta > 0$*,* $\phi(\boldsymbol{V}) > \phi(\boldsymbol{V'})$*.*

Finally, the *impartiality* property, which is rooted in the principle of "equal treatment of equals" states that individuals sharing similar characteristics should be treated similarly.

Definition 3.5. *In a system, individuals with similar characteristics should be treated similarly, i.e., the solution should be independent of the order of its arguments* $\phi(\boldsymbol{V}) = \phi(\boldsymbol{V}_\sigma)$*, where* σ *is a permutation and* $\boldsymbol{V}_\sigma$ *is the vector obtained from vector* $\boldsymbol{V}$ *permuted by* σ*.*

To ensure fairness that satisfies the above three properties, we use a well-known generalized Gini welfare function (GGF) [53], which can be defined as:

$$\phi_{\boldsymbol{\omega}}(\boldsymbol{u}) = \sum_{i \in D} \omega_i u_i^{\uparrow}, \tag{1}$$

$\boldsymbol{u} \in \mathbb{R}^D$ represents the utility vector of a size D for D objectives, $\boldsymbol{\omega} \in \mathbb{R}^D$ is a fixed weight vector with positive components that strictly decrease (i.e., $\omega_1 > \ldots > \omega_D$) with $\sum_i w_i = 1$, and $\boldsymbol{u}^{\uparrow}$ denotes the vector by sorting the components of $\boldsymbol{u}$ in an increasing order (i.e., $u_1^{\uparrow} \leq \ldots \leq u_D^{\uparrow}$). GGF satisfies the aforementioned three fairness properties. As the weights are positive, it is monotonic with respect to Pareto dominance, thus satisfying the efficiency property. Since the utility vector is reordered, it is also symmetric and therefore satisfies the impartiality property. Furthermore, the positive and decreasing weights ensure that GGF is Schur-concave, i.e., monotonic with respect to Pigou-Dalton transfers, and therefore satisfies the impartiality property.

Moreover, the GGF is a piecewise-linear concave function and can be equivalently expressed as:

$$\phi_{\boldsymbol{\omega}}(\boldsymbol{V}^{\pi}(s)) = \min_{\sigma \in \mathbb{S}_D} \left\{ \boldsymbol{\omega}_\sigma^\top \boldsymbol{V}^{\pi}(s) \right\} \tag{2}$$

where $\mathbb{S}_D$ is the symemetric group of degree D and $\boldsymbol{\omega}_\sigma^\top$ denotes the permutation of weights corresponding to σ. For a fixed permutation, $\boldsymbol{\omega}_\sigma^\top \boldsymbol{V}^{\pi}(s)$ defines an affine function of $\boldsymbol{V}^{\pi}(s)$, which guarantees piecewise linearity and concavity of GGF.

GGF has been studied and used in MORL extensively [31, 40, 47, 56], however, all of these works used it for single-policy setting. We are the first ones to use it in a multi-policy MORL setting. In multi-policy MORL, the usual approach is to find all Pareto non-dominated solutions [35, 51]. This approach may work for small problems, however, for large-scale problems, the Pareto non-dominated solutions grow exponentially. A better way to achieve scalable and multiple solutions to approximate the Pareto front is possibly to arrive at the solutions that form the convex envelope and thus form a CCS.

4 FAIRNESS IN MORL

Since we are in a multi-policy MORL setting, where an agent learns a set of Pareto optimal policies, fairness becomes more important as different stakeholders may have different preferences and during

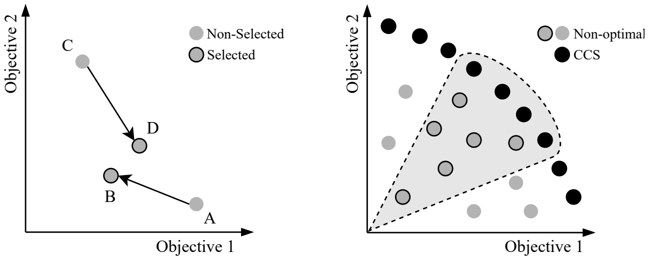


Figure 1: Examples of 2-objective MOMDP where GGF leads to fairer outcomes.

inference, any solution can be used from the Pareto non-dominated solutions given the stakeholder preferences. We formalize this sophisticated multi-policy fair optimization problem as:

$$\forall \boldsymbol{\omega} \in \Omega, \quad \max_{\pi\in\Pi} \ \phi_{\boldsymbol{\omega}}(\boldsymbol{J}(\pi)), \tag{3}$$

where Ω is the set of valid preference weights sorted in descending order, $\boldsymbol{J}(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_t]$ is the expected discounted return, and $\phi_{\boldsymbol{\omega}}(\boldsymbol{J}) = \sum_{i=1}^{D} w_i J_{(i)}$ with $J_{(1)} \leq \cdots \leq J_{(n)}$. The concavity of GGF makes problem (3) as convex optimization problem, enabling efficient solutions within the CCS. Below, we establish three foundational results, which show that it is always feasible to obtain optimal solutions in the CCS corresponding to GGF fair optimization. Next, we demonstrate that a non-stationary policy based on accrued rewards is beneficial in yielding improved fairness when compared with its stationary counterpart. Here, a policy yields improved fairness or is fairer if a higher welfare score, defined in (1), is achieved. Lastly, we show that a stochastic policy may yield fairer solutions than a deterministic one.

Sufficiency of Optimal Solutions in the CCS. The first question relates to the learning of fair policies in a multi-policy MORL setting is which subset of policies may be optimal among the set of all (possibly non-stationary) policies. Indeed, for linear scalarization function, CCS contains the set of Pareto front solutions. Below, we formally state it:

Lemma 4.1. *For any MOMDP with linear preferences over objectives, the CCS contains an optimal policy for any linear combination of the objectives.*

While GGF introduces non-linear fairness objectives, its piecewise linearity and concavity allow representation as a maximum of linear functions, which ensures that solutions lie within the CCS. The following proposition establishes the sufficiency of the CCS in representing optimal policies for $\phi_{\boldsymbol{\omega}}$ preference weights.

Proposition 4.1. *For any $s \in \mathcal{S}$ in an MOMDP and a piecewise-linear concave welfare function $\phi_{\boldsymbol{\omega}}$ (e.g., GGF) that can be represented as, $\phi_{\boldsymbol{\omega}}(\boldsymbol{V}^\pi(s)) = \min_{\sigma\in\mathbb{S}_D}\{\boldsymbol{\omega}_\sigma^\top \boldsymbol{V}^\pi(s)\}$, there exists a policy $\pi^* \in$ CCS such that $\phi_{\boldsymbol{\omega}}(\boldsymbol{V}^{\pi^*}(s)) \geq \phi_{\boldsymbol{\omega}}(\boldsymbol{V}^\pi(s)), \quad \forall\pi \in \Pi$.*

Example 4.1. *To illustrate how the GGF function ensures fairness in MORL, consider a two-objective MOMDP with objective values $\boldsymbol{V}_1 = (3, 1)$ and $\boldsymbol{V}_2 = (2, 3)$ and weights $(1, 2)$. For $\boldsymbol{V}_1$, two weighted combinations are possible: **A)** $(3, 1) \cdot (2, 1) = (6, 1)$ with scalar sum $6+1 = 7$, **B)** $(3, 1) \cdot (1, 2) = (3, 2)$ with scalar sum $3+2 = 5$. Since the GGF is defined as $\phi_{\boldsymbol{\omega}}(\boldsymbol{V}^\pi(s)) = \min_{\sigma\in\mathbb{S}_D}\{\boldsymbol{\omega}_\sigma^\top \boldsymbol{V}^\pi(s)\}$, it selects the lower scalar utility, preferring B over A (see left figure of Figure 1). Similarly, for $\boldsymbol{V}_2$: **C)** $(2, 3) \cdot (1, 2) = (2, 6)$ with scalar sum $2 + 6 = 8$, **D)** $(2, 3) \cdot (2, 1) = (4, 3)$ with scalar sum $4+3 = 7$. Here, D is preferred over C. This mechanism directs the solutions toward the fairer region (gray dotted area in the right figure of Figure 1), demonstrating that maximizing the GGF leads to fair Pareto-optimal solutions.*

Fairness of Non-Stationary Policies. In fair MORL, learning non-stationary policies can be particularly beneficial, as they leverage historical information to make more informed decisions and adapt over time.

Proposition 4.2. *Let the reward $\boldsymbol{r}$ be nonnegative, and Π_S and Π_{NS} be the sets of stationary and non-stationary policies, respectively. For any $s \in \mathcal{S}$ in an MOMDP and a given $\phi_{\boldsymbol{\omega}}$, there exists a non-stationary policy $\pi_{NS} \in \Pi_{NS}$ that achieves a higher welfare score than any stationary policy $\pi_S \in \Pi_S$, i.e., $\exists \pi_{NS} \in \Pi_{NS}$: $\phi_{\boldsymbol{\omega}}(\boldsymbol{V}^{\pi_{NS}}(s)) \geq \max_{\pi_S\in\Pi_S} \phi_{\boldsymbol{\omega}}(\boldsymbol{V}^{\pi_S}(s))$.*

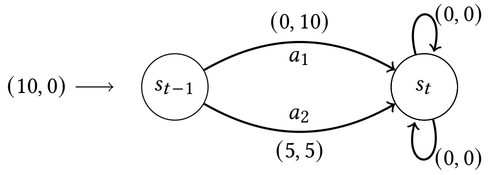


Figure 2: Example of MOMDP where actions lead to different rewards.

Example 4.2. *To illustrate the value of learning a non-stationary policy, consider a 2-objective MOMDP, shown in Fig. 2. At timestep $t > 0$, the agent has accrued a vector reward $\boldsymbol{r}_{acc} = (10, 0)$ for two objectives. The preference weights, encapsulated within the welfare function ϕ, denote decreasing weights, such as $(0.8, 0.2)$. With two potential actions, each leading to a final state, action a_1 yields a reward of $(0, 10)$, while action a_2 yields $(5, 5)$. Since s_t is the absorbing state, we can set the discount factor $\gamma = 1$. Under the given welfare function ϕ defined in Equation (1), executing a_1 yields a welfare score of 2, whereas executing a_2 yields a score of 5 if only future rewards are considered. However, considering historical data, i.e., $\boldsymbol{r}_{acc}$, a_1 yields a higher accrued episodic return of $(10, 10)$ and a welfare score of 10. Similarly, a_2 yields $(15, 5)$ and 7 episodic return and welfare scores, respectively. Note that action a_1 is a fairer choice in this case since it balances the two objectives, unlike action a_2, which fails to achieve a more equitable outcome. Hence, employing historical data, namely, accrued rewards in this case, is critical to enable fair policy learning.*

Optimality of Stochastic Policies for Fairness. Unlike single-objective RL, in MORL, a deterministic policy may not be fair optimal. A fairer solution can often be achieved through randomization.

Proposition 4.3. *Let Π_{ST} be the set of stochastic policies and Π_D be the set of deterministic policies. For an MOMDP $\mathcal{M}$ and a concave welfare function such as $\phi_{\boldsymbol{\omega}}$, there exists a stochastic policy $\pi_{ST} \in \Pi_{ST}$ such that $\phi_{\boldsymbol{\omega}}(\boldsymbol{V}^{\pi_{ST}}) \geq \max_{\pi_D\in\Pi_D} \phi_{\boldsymbol{\omega}}(\boldsymbol{V}^{\pi_D})$.*

The proofs of the above lemma and propositions can be found in Appendix A. Figure 3 (left) illustrates GGF on a two-objective

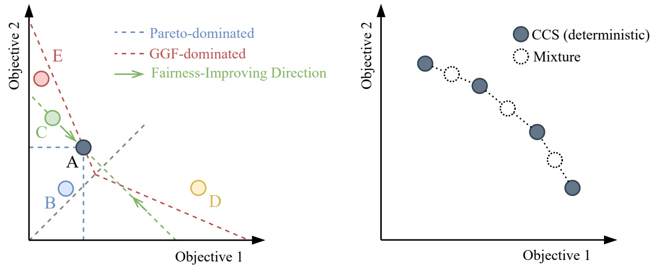


Figure 3: Left Figure: Point A is always preferred to B due to Pareto dominance; A is always preferred to C due to the Pigou-Dalton transfer principle (fairer solution); depending on the weights of GGF, Points D and E can be dominated or non-dominated by A (w.r.t. GGF); with weights (0.3, 0.7), A is preferred to E but not to D. Right Figure: Black points refer to deterministic policy that in CCS and stochastic policy can be obtained with the mixture of deterministic policies in the CCS, shown in dotted point. Demonstrate that stochastic policy may achieve a fairer solution, which a deterministic policy cannot.

optimization task. The optimality of stochastic policies implies that restricting the search for fair solutions to deterministic policies is insufficient. Stochastic policies offer a broader range of solutions and may better capture the trade-offs among multiple objectives, enhancing the overall fairness, as shown in Figure 3 (right).

5 PROPOSED ALGORITHMS

In this section, we introduce three novel algorithms that incorporate fairness into MORL based on our technical analysis Section 4. These algorithms optimize the GGF welfare function defined in (1) to ensure fairness across D fixed users with varying preferences. Our proposed methods are scalable and sample-efficient as they utilize a single parameterized network to estimate Q-values for all objectives while maintaining a diverse set of Pareto-optimal policies. We present three distinct algorithms: Fair Multi-Objective Deep Q-Learning (F-MDQ), its extension with non-stationary policies (FN-MDQ), and a novel extension incorporating stochastic policies (FNS-MDQ). This progression from stationary to non-stationary to stochastic and non-stationary policies demonstrates our systematic approach to enhancing fairness in MORL algorithms, with each method building upon and improving the previous one.

F-MDQ. F-MDQ builds on the Envelope algorithm [55] by replacing the linear scalarization function with the GGF ϕ. This ensures fairness while learning policies across all preferences $\boldsymbol{\omega} \in \Omega$. The Bellman optimality equation for F-MDQ is given by:

$$\boldsymbol{Q}^*(s,a,\boldsymbol{\omega}) = \mathbb{E}[\boldsymbol{r}(s,a) + \gamma \boldsymbol{Q}^*(s', \sup_{a'\in\mathcal{A}} \phi_{\boldsymbol{\omega}}(\boldsymbol{r}(s,a)+ Q^*(s',a',\boldsymbol{\omega}),\boldsymbol{\omega}) \mid s,a],$$

where $\boldsymbol{Q}^{\pi}(s,a,\boldsymbol{\omega})$ represents the expected return vector for policy π, conditioned on preference $\boldsymbol{\omega}$. As the MO Q-function is parameterized, it can be learned by minimizing the loss function $\mathcal{L} = \mathbb{E}_{(s,a,\boldsymbol{r},s',\omega)\sim\mathcal{D}}\left[\|\boldsymbol{y} - \boldsymbol{Q}(s,a,\boldsymbol{\omega})\|_2^2\right]$, where the expectation is taken over experiences sampled from the replay buffer $\mathcal{D}$. Given that the loss function includes an expectation over $\boldsymbol{\omega}$, the preference weights are sampled randomly and are decoupled from the transitions, allowing increased sample efficiency through a resampling scheme similar to Hindsight Experience Replay (HER) [3]. The target $\boldsymbol{y}$ is F-MDQ is computed as

$$\boldsymbol{y} = \boldsymbol{r}(s,a) + \gamma \boldsymbol{Q}'(s', \sup_{a'\in\mathcal{A}} \phi_{\boldsymbol{\omega}}(\boldsymbol{r}(s,a) + \gamma Q(s',a',\boldsymbol{\omega})),\boldsymbol{\omega}),$$

where Q' represents the target multi-objective Q-function, and the supremum is applied over the GGF welfare function $\phi_{\boldsymbol{\omega}}$ instead of a linear weighted sum. This ensures that actions are selected based on higher welfare scores rather than simply maximizing Q-values.

FN-MDQ. FN-MDQ extends F-MDQ by incorporating accrued rewards into the state to learn non-stationary policies, as discussed in Proposition A.2. It augments the observed state with accrued rewards, allowing the agent to balance reward distribution across users more effectively (as demonstrated in Example 2). The augmented state is defined as $\mathfrak{s}_t = (s_t, \boldsymbol{r}_{\text{acc}})$, where $\boldsymbol{r}_{\text{acc}} = \sum_{i=1}^{t-1}\gamma^{i-1}\boldsymbol{r}_i$ is the discounted total reward received in the current trajectory. The regression target for FN-MDQ is then given by

$$\boldsymbol{r}(s_t,a_t) + \gamma \boldsymbol{Q}'(\mathfrak{s}_{t+1}, \sup_{a'\in\mathcal{A}} \phi_{\boldsymbol{\omega}}(Q(\mathfrak{s}_{t+1},a',\boldsymbol{\omega})),\boldsymbol{\omega}).$$

Here, the immediate reward $\boldsymbol{r}(s_t,a_t)$ is excluded from the optimal action computation since this signal is already included in the augmented state as part of the discounted total reward. This extension enables the agent to identify and prioritize users who have received insufficient rewards within an episode.

FNS-MDQ. Given that stochastic policies can outperform deterministic ones (see Proposition A.3), the performance of FN-MDQ can be enhanced by incorporating stochastic policies. We now explain how stochastic policies can be integrated into the FN-MDQ.

Under the stochastic policies, the target Q-value is adjusted to account for the expected Q-values, which reformulates the update:

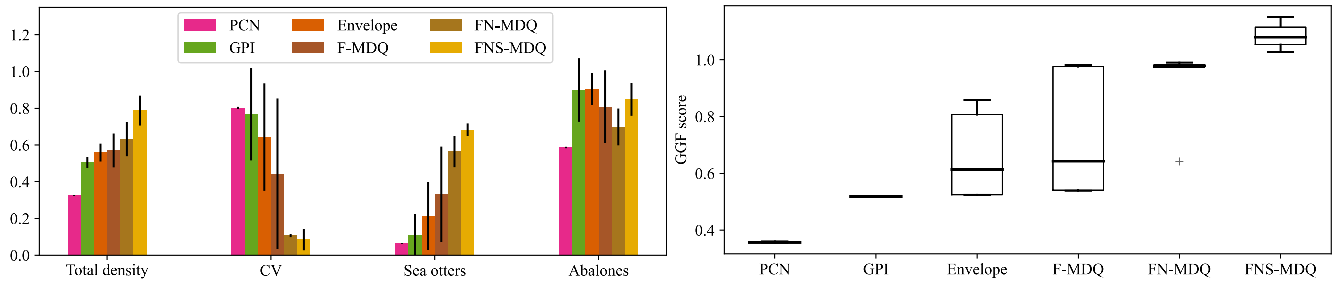
$$\boldsymbol{r}(s_t,a_t) + \gamma \boldsymbol{Q}'(\mathfrak{s}_{t+1}, \sum_{a'\in\mathcal{A}} \phi_{\boldsymbol{\omega}}(\pi(a' \mid \mathfrak{s}_{t+1})Q(\mathfrak{s}_{t+1},a',\boldsymbol{\omega})),\boldsymbol{\omega}),$$

where $\pi(a' \mid \mathfrak{s}_{t+1})$ is the probability of taking action a' given the augmented state $\mathfrak{s}_{t+1}$. This reformulation considers the distribution of possible actions rather than selecting a single best deterministic action, aligning with our theoretical insights.

Unlike F-MDQ and FN-MDQ, which rely on deterministic action selection, FNS-MDQ samples actions from a probability distribution over Q-values. This stochastic action selection improves fairness by enabling more balanced policy exploration and reducing biases that arise from always selecting the highest Q-value action. Note that, during the training phase, all algorithms employ an ϵ-greedy policy during training, however, FNS-MDQ differs in its action-selection strategy by using the best learned stochastic policy rather than a deterministic greedy approach. This increased flexibility and randomness can lead to more equitable solutions.

6 EXPERIMENTS

To evaluate the proposed methods, we conduct experiments across three domains, each characterized by varying levels of complexity in terms of the number of objectives. These domains, ranging from low to high in terms of the number of objectives, include species



(a) Total density, CV, min density, and max density.

(b) GGF scores.

Figure 4: Performances of multi-policy MORL baselines and our methods in species conservation ($D = 2$)

.

conservation, resource gathering, and multi-product web advertising. Each environment presents unique challenges where fairness plays a critical role. We first briefly describe each environment and then present our experimental results.

6.1 Environments

Our first domain is a species conservation (SC) environment, which addresses a critical ecological challenge: balancing the populations of two highly interacting endangered species, the sea otter and the northern abalone. Both species are at risk of extinction, requiring sophisticated management strategies to ensure their survival. We adopt the model proposed by [10], which simulates the predation relationship between the species, where sea otters prey on abalones. This dynamic presents a unique preservation challenge, as the survival of one species could potentially drive the other to extinction if not properly managed. In this environment, the state space comprises the current population of both species. The action space includes *introducing sea otters*, *enforcing anti-poaching measures*, *controlling sea otters*, *implementing a combination of half-antipoaching and half-controlled sea otters*, or *taking no action*. Each action has significant ecological implications. For instance, introducing sea otters may help balance the abalone population, but if mismanaged, could lead to abalone extinction. The reward function is defined by the population densities of both species, i.e., $D = 2$. Fairness in this context is interpreted as achieving a balanced distribution of species densities to ensure their preservation.

Our second environment is a resource-gathering (RG) problem, which is a 5×5 grid world that contains three types of resources: gold, gems, and stones. These resources are randomly positioned on the grid and regenerate randomly upon consumption. The main challenge here is to collect these resources, where each resource has a different value: gold and gems are valued at 1, while stones have a lower value of 0.4. This creates an intentionally uneven resource distribution, with two stones, one gold and one gem. In this environment, the state is defined by the agent's current location on the grid and the cumulative count of each resource collected during its trajectory. The agent can take four actions: up, down, left, and right. The reward function is defined as a vector representing the rewards collected for each type of resource, i.e., $D = 3$. Fairness here is defined as the equitable collection of resources despite their differing values. Note that, this problem is particularly important for validating whether the proposed methods can achieve fairer solutions while still reaching Pareto optimal solutions.

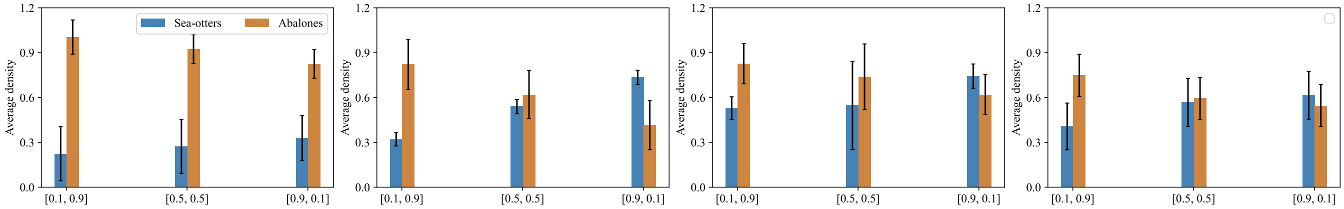
Our third domain is a multi-product web advertising (MWP) problem that involves an online store offering $D = 7$ distinct products. Here, the agent decides which advertisement to display: a product-specific advertisement for one of the products $i \in [0, ..., D-1]$, or a general advertisement that is not tailored to any specific product. In this environment, the state space includes the number of products available in the store, as well as the number of visits, purchases, and exits. The action space is $D + 1$, where actions 0 through $D - 1$ correspond to displaying advertisements for specific products, and action D involves showing a general advertisement. This additional action adds complexity, requiring the agent to decide the optimal moment to transition between states. The reward function is designed so that the agent receives a reward of 1 in the i^{th} dimension of the reward vector if a product of the type i is sold after displaying its advertisement. In this environment, fairness is defined as balancing the frequency of advertisements shown for each product, ensuring no single product is overly prioritized. The challenge lies in increasing overall rewards while maintaining a fair distribution of advertisement exposure across all products.

6.2 Baselines

We compare our methods against several multi-policy MORL baselines. Generalized Policy Improvement Linear Support (GPI-LS) [2] employs GPI [4] to combine policies within its learned CCS and prioritize the weight vectors on which agents should train at each moment. The Envelope algorithm [55] uses a single neural network conditioned on a weight vector to approximate the CCS. Pareto Conditioned Networks (PCN) [42] utilizes a neural network conditioned on a desired return per objective and is trained via supervised learning to predict actions that yield the desired return. Hyperparameters for each method were optimized, and experiments were run for five different seeds, with average results reported. Further details on experimental configurations and hyperparameters are provided in the Appendix D.

6.3 Results

In this section, we present the experimental results across the three environments presented above. The primary objective of these experiments is to assess the effectiveness of our proposed methods



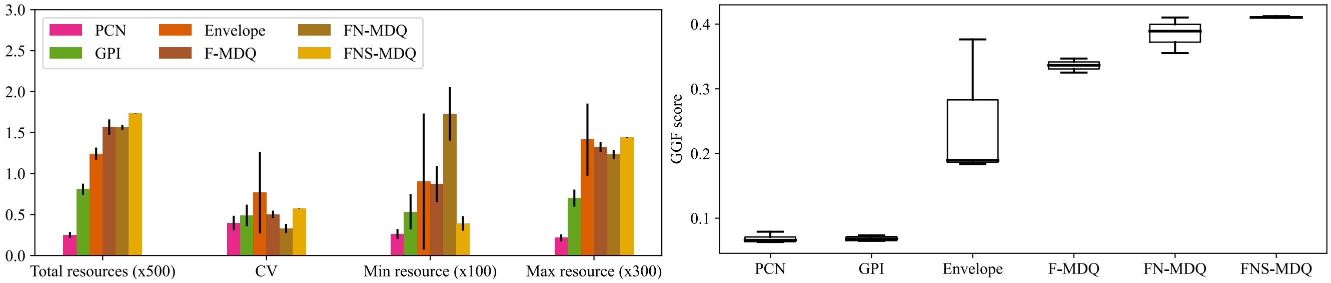
(a) Envelope.

(b) F-MDQ.

(c) FN-MDQ.

(d) FNS-MDQ.

Figure 5: Individual densities of Envelope, and our proposed methods during testing with unseen preferences.



(a) Total resources, CV, min and max resources.

(b) GGF scores.

Figure 6: Performances of multi-policy MORL baselines and our methods in resource gathering ($D = 3$).

by addressing the following key research questions: **(A)** How effective are our methods in learning fairer solutions compared to multi-policy MORL baselines? **(B)** Can our methods generate fair solutions across different preference settings during inference? **(C)** To what extent can our proposed algorithms achieve comparable performance in terms of hypervolume and cardinality relative to multi-policy MORL approaches? **(D)** What is the impact of our approach on the diversity and quality of non-dominated solutions that satisfy fairness criteria? **(E)** Does the incorporation of stochastic policies in MO Q-learning based algorithms contribute to improved fairness or overall performance?

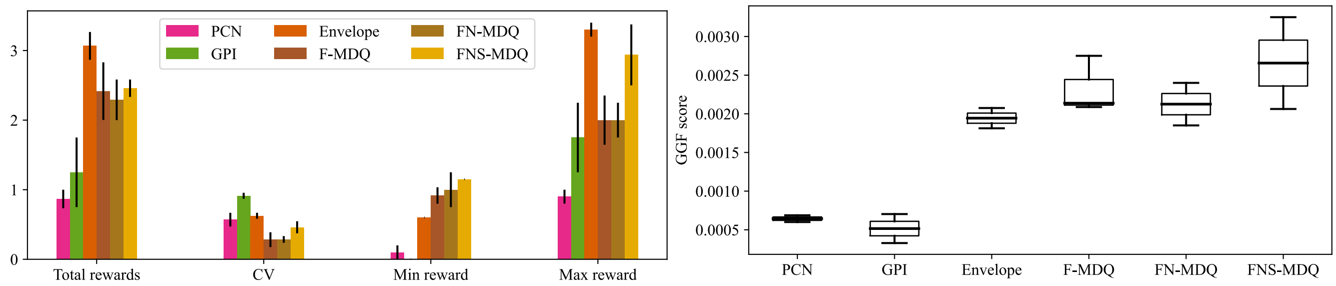
Question (A). To evaluate how effective our methods are in learning fair solutions, we conducted experiments in the SC, RG, and MWP domains, as shown in Figures 4a, 6a and 7a. We compare our proposed methods (F-MDQ, FN-MDQ, and FMS-MDQ) with multi-policy MORL baselines during the training phase. We choose these baselines as they are the current state-of-the-art MORL baselines. The Key evaluation metrics used include total rewards, Coefficient of Variation (CV) indicating the variations in different objectives' utilities, and the minimum and maximum objective utilities. Moreover, GGF welfare scores were computed to quantify fairness. As we are in a multi-policy MORL, an agent learns a set of Pareto optimal policies during learning. To show the results, we computed these metrics over the last 50 trajectories for all the Pareto optimal policies and reported their normalized scores. Note that, during the last 50 trajectories, all the agents are converged so it ensures a fair comparison for multi-policy MORL methods.

As shown in Figure 4a, PCN performs the worst. GPI outperforms PCN, likely due to its TD3-based [22] architecture and efficient prioritization scheme in learning the Pareto front $\mathcal{F}$. The Envelope algorithm performs better than PCN and GPI as it achieves higher total density and, interestingly, lower CV. However, our proposed algorithms outperform all other methods by achieving the lowest CV and highest welfare scores Figure 4b, with FN-MDQ outperforming F-MDQ, underscoring the value of non-stationary policies. Furthermore, FNS-MDQ outperforms both F-MDQ and FN-MDQ as it maximizes the minimum objective utility and demonstrates better fairness through optimizing the welfare function ϕ_{ω}. Similar results are observed in RG Figure 6a, where PCN performs the worst as it collects the least resources, likely due to its limitations in deterministic environments [42]. Although GPI performs better than PCN, both exhibit low CV alongside poor overall performance and GGF welfare utility Figure 6b. The Envelope algorithm achieves better performance in terms of rewards but suffers from the highest CV and lower GGF utility scores. In contrast, our proposed methods attain a lower CV compared to all baselines, and they achieve the highest GGF scores, highlighting their effectiveness in identifying fair policies through welfare function optimization. Interestingly, FNS-MDQ exhibits a higher CV due to its higher maximum objective and the total resources collected. Nevertheless, it also achieves the highest welfare scores. Consistent with our previous results, our proposed methods in MVP environment Figure 7a achieve the highest welfare scores, indicating their capacity to ensure an equitable distribution of rewards across all objectives. Moreover, they maintain the lowest CV, highlighting their robustness in learning fair policies, even in highly stochastic environments with a higher number of objectives. Once again, PCN, and GPI perform the worst, further underscoring the efficacy of our methods in this context.

Question (B). To check whether our methods can generate fair solutions across different preference settings, we evaluated our

Table 1: Hypervolume (HV) and Cardinality (CD) of different MORL methods across SC, RC, and MWP domains.

Methods	SC		RC		MWP	
	HV (10^4)↑	CD↑	HV (10^5)	CD	HV (10^9)	CD
PCN	1.81 ± 0.14	19.67 ± 2.99	11.69 ± 0.90	6.0 ± 1.27	10.17 ± 0.22	43.5 ± 1.06
GPI	2.82 ± 0.03	12.0 ± 2.05	7.33 ± 0.19	43.0 ± 2.62	10.44 ± 0.86	41.0 ± 2.83
Envelope	2.35 ± 0.18	5.6 ± 1.04	17.51 ± 3.73	19.75 ± 6.79	10.55 ± 1.96	51.5 ± 1.06
F-MDQ	2.22 ± 0.19	6.6 ± 1.31	16.92 ± 1.63	31.33 ± 7.84	10.45 ± 2.40	48.0 ± 2.12
FN-MDQ	2.34 ± 0.07	11.68 ± 1.05	20.38 ± 1.49	33.54 ± 8.29	10.51 ± 2.42	52.2 ± 2.44
FNS-MDQ	2.91 ± 0.20	15.38 ± 1.10	24.40 ± 2.22	36.11 ± 8.96	10.62 ± 2.45	51.05 ± 2.30



(a) Total rewards, CV, min reward, and max reward.

(b) GGF scores.

Figure 7: Performances of multi-policy MORL baselines and our proposed methods in the MPW ($D = 7$).

algorithms with unseen preferences during testing in the SC environment. As shown in Figure 5, which presents the individual species densities (sea otters and abalones) for preference configurations $(0.1, 0.9)$, $(0.5, 0.5)$, $(0.9, 0.1)$, the Envelope algorithm fails to produce fair solutions, suggesting its limitation in generating fair optimal policies across varying preferences. In contrast, F-MDQ generates more balanced solutions, while FN-MDQ and FNS-MDQ achieve even fairer outcomes, further validating our earlier findings.

Question (C). To answer this question, we evaluate algorithms in terms of MORL metrics, such as cardinality and hypervolume (HV) in all three environments. A higher cardinality indicates greater policy diversity within $\mathcal{F}$, while HV measures both the convergence rate and policy diversity [29]. Recall that, HV is defined as for any given $\mathcal{F}'$ an approximation of $\mathcal{F}$ and a reference point (the worst-possible return), it measures the volume of the hypercube spanned by the reference point and estimated return in a trajectory. Table 1 presents the HV and cardinality in all environments. These results show that our proposed methods perform on par with multi-policy MORL baselines.

Question (D). The results discussed in previous questions suggest that our methods can generate a range of Pareto *non-dominated* solutions across varied preference configurations, which indicates better coverage of the objective space, thus improving performance across multiple objectives. For quality, our proposed algorithms consistently achieve the lowest CV and highest GGF scores across all three domains, performing on par in terms of HV and CD which indicates that our solutions exhibit more equitable distribution of objective utilities while maintaining Pareto optimality.

Question (E). Finally, to assess the impact of incorporating stochastic policies in MO Q-learning algorithms, we refer to the results in Figures 4, 6 and 7, where stochastic policies consistently improve both efficiency and fairness. Moreover, as shown in Table 1 incorporating stochastic policies also enhances MORL metrics, including HV and cardinality, validating the contribution of stochasticity to both fairness and overall performance.

7 CONCLUSIONS AND LIMITATIONS

In this paper, we presented a novel approach to addressing fairness in the context of multi-policy MORL. Our proposed methods leverage a single parameterized network to learn optimized policies across the entire space of possible preferences. Both theoretical and empirical analyses demonstrate that learning a non-stationary policy significantly improves fairness. Additionally, we highlighted the importance of stochastic policies in achieving fair outcomes. Experimental evaluations in three domains validated the effectiveness of our approach in yielding more equitable policies compared to state-of-the-art MORL and fair baselines.

Our approach also has some limitations. First, it is limited to MOMDPs with discrete action spaces. Second, it assumes that preference weights are linear to learn the CCS, which may not capture the concave regions of the Pareto front. Third, the current formulation is focused on individual fairness. Given that optimizing a welfare function is a broad framework applicable to various real-world MORL problems involving general utilities, an important direction for future research is to extend this approach to accommodate more sophisticated objective functions, particularly those related to group-level fairness, safety, and risk sensitivity.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research under Grant N000142412405 and the Army Research Office under Grants W911NF2110103 and W911NF2310363.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML*.

[2] Lucas N Alegre, Ana LC Bazzan, Diederik M Roijers, Ann Nowé, and Bruno C da Silva. 2023. Sample-Efficient Multi-Objective Learning via Generalized Policy Improvement Prioritization. *arXiv preprint arXiv:2301.07784* (2023).

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. *NeurIPS* 30 (2017).

[4] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. 2017. Successor features for transfer in reinforcement learning. *NeurIPS* 30 (2017).

[5] Leon Barrett and Srini Narayanan. 2008. Learning All Optimal Policies with Multiple Criteria. In *ICML*.

[6] X. Bei, S. Liu, C.K. Poon, and H. Wang. 2022. Candidate selections with proportional fairness constraints. In *AAMAS*.

[7] Aurélie Beynier, Yann Chevaleyre, Laurent Gourvès, Ararat Harutyunyan, Julien Lesca, Nicolas Maudet, and Anaëlle Wilczynski. 2019. Local envy-freeness in house allocation problems. *AAMAS* (2019).

[8] Steven J. Brams and Alan D. Taylor. 1996. *Fair Division: From Cake-Cutting to Dispute Resolution.* Cambridge University Press.

[9] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. 2017. Multiobjective bandits: Optimizing the generalized gini index. In *ICML*. 625–634.

[10] Iadine Chadès, Janelle MR Curtis, and Tara G Martin. 2012. Setting realistic recovery targets for two interacting endangered species, sea otter and northern abalone. *Conservation Biology* 26, 6 (2012), 1016–1025.

[11] M. Chakraborty, A. Igarashi, W. Suksompong, and Y. Zick. 2021. Weighted envy-freeness in indivisible item allocation. *TEAC* 9, 3 (2021), 1–39.

[12] Satya R. Chakravarty. 1990. *Ethical Social Index Numbers.* Springer Verlag.

[13] Jingdi Chen, Yimeng Wang, and Tian Lan. 2021. Bringing Fairness to Actor-Critic Reinforcement Learning for Network Utility Optimization. In *IEEE Conference on Computer Communications.* 1–10.

[14] Violet Xinying Chen and JN Hooker. 2021. A guide to formulating equity and fairness in an optimization model. *Preprint* (2021), 162–174.

[15] Yann Chevaleyre, Paul E Dunne, Michel Lemaître, Nicolas Maudet, Julian Padget, Steve Phelps, and Juan A Rodríguez-aguilar. 2006. Issues in Multiagent Resource Allocation. *Computer* 30 (2006), 3–31.

[16] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. *NeurIPS* 30 (2017).

[17] Alexandra Cimpeana, Catholijn Jonkerb, Pieter Libina, and Ann Nowéa. 2023. A Multi-objective Framework For Fair Reinforcement Learning. In *Multi-Objective Decision Making Workshop 2023.*

[18] Cyrus Cousins. 2021. An axiomatic theory of provably-fair welfare-centric machine learning. In *NeurIPS*.

[19] Virginie Do and Nicolas Usunier. 2022. Optimizing generalized Gini indices for fairness in rankings. *arXiv preprint arXiv:2204.06521* (2022).

[20] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.* 214–226.

[21] Zimeng Fan, Nianli Peng, Muhang Tian, and Brandon Fain. 2022. Welfare and Fairness in Multi-objective Reinforcement Learning. *arXiv preprint arXiv:2212.01382* (2022).

[22] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *ICML*. 1582–1591.

[23] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *NeurIPS*.

[24] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *ICML*. 1617–1626.

[25] N. Jensen. 1967. An introduction to bernoullian utility theory, I: utility functions. *Swedish Journal of Economics* 69 (1967), 163–183.

[26] Jiechuan Jiang and Zongqing Lu. 2019. Learning Fairness in Multi-Agent Systems. In *NeurIPS*.

[27] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 247–254.

[28] David Kurokawa, Ariel D. Procaccia, and Nisarg Shah. 2015. Leximin Allocations in the Real World. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation.* 345–362. https://doi.org/10.1145/2764468.2764490

[29] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. 2002. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary Computation* 10, 3 (2002), 263–282.

[30] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018.* 101–102.

[31] Debmalya Mandal and Jiarui Gan. 2022. Socially Fair Reinforcement Learning. *arXiv preprint arXiv:2208.12584* (2022).

[32] Dimitris Michailidis, Willem Röpke, Diederik M Roijers, Sennay Ghebreab, and Fernando P Santos. 2024. Scalable Multi-Objective Reinforcement Learning with Fairness Guarantees using Lorenz Dominance. *arXiv preprint arXiv:2411.18195* (2024).

[33] Kristof Van Moffaert and Ann Nowé. 2014. Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies. *JMLR* 15 (2014), 3663–3692.

[34] H. Moulin. 2004. *Fair Division and Collective Welfare.* MIT Press.

[35] Yusuke Mukai, Yasuaki Kuroe, and Hitoshi Iima. 2012. Multi-Objective Reinforcement Learning Method for Acquiring All Pareto Optimal Policies Simultaneously. In *IEEE International Conference on Systems, Man, and Cybernetics.*

[36] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning Optimal Fair Policies. In *ICML*.

[37] Samer B Nashed, Justin Svegliato, and Su Lin Blodgett. 2023. Fairness and sequential decision making: Limits, lessons, and opportunities. *arXiv preprint arXiv:2301.05753* (2023).

[38] Patrice Perny, Paul Weng, Judy Goldsmith, and Josiah Hanna. 2013. Approximation of Lorenz-optimal solutions in multiobjective Markov decision processes. In *International Conference on Uncertainty in Artificial Intelligence (UAI).*

[39] M.L. Puterman. 1994. *Markov decision processes: discrete stochastic dynamic programming.* Wiley.

[40] Junqi Qian, Umer Siddique, Guanbao Yu, and Paul Weng. 2025. From fair solutions to compromise solutions in multi-objective deep reinforcement learning. *Neural Computing and Applications* (2025), 1–31.

[41] John Rawls. 1971. *The Theory of Justice.* Havard university press.

[42] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. 2022. Pareto conditioned networks. *arXiv preprint arXiv:2204.05036* (2022).

[43] Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. 2019. Average Individual Fairness: Algorithms, Generalization and Experiments. In *NeurIPS*.

[44] Umer Siddique, Peilang Li, and Yongcan Cao. 2024. Fairness in Traffic Control: Decentralized Multi-agent Reinforcement Learning with Generalized Gini Welfare Functions. In *Multi-Agent reinforcement Learning for Transportation Autonomy.*

[45] Umer Siddique, Peilang Li, and Yongcan Cao. 2024. Towards Fair and Equitable Policy Learning in Cooperative Multi-Agent Reinforcement Learning. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop.*

[46] Umer Siddique, Abhinav Sinha, and Yongcan Cao. 2023. Fairness in Preference-based Reinforcement Learning. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning.*

[47] 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.*

[48] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *NeurIPS*.

[49] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (London, United Kingdom).

[50] Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. 2013. Scalarized multi-objective reinforcement learning: Novel design techniques. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL).* 191–199.

[51] Kristof Van Moffaert and Ann Nowé. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research* 15, 1 (2014), 3483–3512.

[52] Paul Weng. 2019. Fairness in Reinforcement Learning. In *AI for Social Good Workshop at International Joint Conference on Artificial Intelligence.*

[53] J.A. Weymark. 1981. Generalized Gini inequality indices. *Mathematical Social Sciences* 1 (1981), 409–430.

[54] R.R. Yager. 1988. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. on Syst., Man and Cyb.* 18 (1988), 183–190.

[55] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. 2019. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *NeurIPS* 32 (2019).

[56] Guanbao Yu, Umer Siddique, and Paul Weng. 2023. Fair Deep Reinforcement Learning with Generalized Gini Welfare Functions. In *Adaptive and Learning Agents (ALA) Workshop.*

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

[58] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-Based Notions of

Fairness in Classification. In *NIPS*.

[59] Xueru Zhang and Mingyan Liu. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 525–555.

[60] Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. 2021. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. In *ICML*.