# Inference-Time Policy Alignment for Fair Reinforcement Learning

**Umer Siddique, Peilang Li, Conor Wallace, Yongcan Cao**

**Keywords:** Reinforcement learning, Inference-time alignment, Policy shaping, Fair optimization

## Summary

Deep reinforcement learning (RL) agents achieve strong performance by optimizing scalar reward functions. However, once deployed, the policies of these RL agents are often rigid and costly to adapt to new performance criteria. For instance, an agent trained to maximize expected cumulative reward may not accommodate previously unknown stakeholder preferences. Existing approaches to achieve fairness, a type of preference, in RL typically assume that such preferences are known *a priori* and require complete retraining of the policy under a fairness-oriented metric. Inspired by inference-time alignment in large language models, we investigate the problem of steering a pretrained RL policy toward welfare-based fairness objectives at inference time without updating the base policy's parameters. We formalize inference-time fairness alignment as a policy shaping problem and propose a multiplicative policy shaping framework that adjusts action probabilities using action-dependent welfare scores, thus requiring no modification to the base policy. Our framework is general and compatible with any deep RL agent. Through extensive experiments across multiple domains, we demonstrate that inference-time policy shaping substantially improves welfare-based fairness objectives while preserving core task performance.

## Contribution(s)

1. We formalize *inference-time fairness alignment* in RL as a policy shaping problem, enabling a frozen reward-maximizing policy to be steered toward welfare-based fairness objectives at deployment without retraining.
   **Context:** While prior work on welfare-based fairness in RL (Siddique et al., 2020; Fan et al., 2022; Yu et al., 2023a) strictly requires training-time optimization, and inference-time alignment techniques (Lu et al., 2023; Mujtaba et al., 2025) have largely focused on autoregressive generation, we bridge this gap and achieve inference-time fairness alignment in sequential decision-making without retraining the base policy.
2. We construct a welfare-augmented MDP that restores Markovianity for non-linear welfare functions and derive, in closed form, the shaped policy that maximizes a KL-regularized welfare surrogate, an exponential reweighting of the base policy which provides a local lower bound on the welfare improvement under the KL constraint.
   **Context:** Although the KL-regularized policy improvement is a foundational result in RL (Schulman et al., 2015), its application to non-linear, trajectory-level welfare functions is previously unexplored. We provide a formal theoretical guarantee that the state augmentation forms a valid augmented MDP and show that welfare-based fairness can be optimized through this structure by introducing an augmented state that tracks cumulative rewards.
3. We propose QFair, a lightweight welfare critic trained offline on base-policy rollouts, which evaluates actions based on both the current state and the evolving cumulative reward vector.
   **Context:** By proving a trajectory equivalence result between the original and augmented MDPs, we demonstrate that QFair can be learned from base-policy rollouts with light exploration, requiring no welfare-driven environment interaction. This avoids the computationally expensive or unsafe welfare-oriented training interactions required by standard multi-objective fair RL methods (Siddique et al., 2020; Zimmer et al., 2021; Ju et al., 2023; Kim et al., 2025).

# Inference-Time Policy Alignment for Fair Reinforcement Learning

**Umer Siddique, Peilang Li, Conor Wallace, Yongcan Cao**
{muhammadumer.siddique,peilang.li,conor.wallace}@my.utsa.edu,
yongcan.cao@utsa.edu

**Department of Electrical and Computer Engineering,**
**University of Texas at San Antonio**

## Abstract

Deep reinforcement learning (RL) agents achieve strong performance by optimizing scalar reward functions. However, once deployed, the policies of these RL agents are often rigid and costly to adapt to new performance criteria. For instance, an agent trained to maximize expected cumulative reward may not accommodate previously unknown stakeholder preferences. Existing approaches to achieve fairness, a type of preference, in RL typically assume that such preferences are known *a priori* and require complete retraining of the policy under a fairness-oriented metric. Inspired by inference-time alignment in large language models, we investigate the problem of steering a pretrained RL policy toward welfare-based fairness objectives at inference time without updating the base policy's parameters. We formalize inference-time fairness alignment as a policy shaping problem and propose a multiplicative policy shaping framework that adjusts action probabilities using action-dependent welfare scores, thus requiring no modification to the base policy. Our framework is general and compatible with any deep RL agent. Through extensive experiments across multiple domains, we demonstrate that inference-time policy shaping substantially improves welfare-based fairness objectives while preserving core task performance.

## 1 Introduction

Deep RL has demonstrated remarkable success in domains ranging from robotics (Peters et al., 2003) to game playing (Silver et al., 2017), typically by optimizing for a single scalar reward function (Sutton & Barto, 1998). These achievements are often built on fixed task definitions and carefully engineered training pipelines, where RL agents learn policies that maximize return by interacting with an environment. However, such policies can be rigid and brittle at deployment and hence, incorporating new constraints, accommodating evolving stakeholder preferences, or handling out-of-distribution (OOD) shifts (Haider et al., 2023; Ajay et al., 2022) typically requires costly retraining or delicate fine-tuning, with no guarantee of robustness. This becomes particularly challenging in socially impactful domains, where the objective extends beyond cumulative reward to include fairness, safety, and multi-stakeholder welfare. For instance, in autonomous driving and decision-making systems, efficiency may need to be traded off for risk sensitivity and safety. Similarly, in resource allocation and transportation, equitable outcomes across user groups may be required. Crucially, such safety or fairness preferences can evolve depending on context, regulatory changes, or societal expectations. Yet, in most deep RL methods, these requirements are addressed during training, and adapting to new preferences or new stakeholders often requires retraining. In practice, such retraining is often impractical due to the associated computational cost, time, and data requirements (Mujtaba et al., 2025). Consequently, there is a need for techniques that can adapt *pretrained* RL policies to evolving

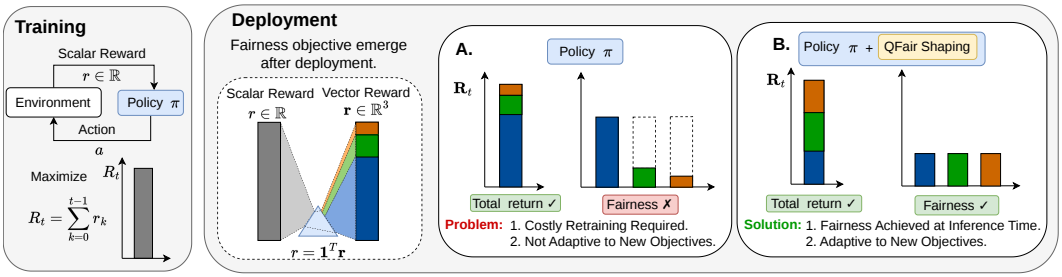



Figure 1: Motivation for inference-time fairness alignment. Left: During training, the policy $\pi$ optimises a scalar return via standard RL. Center: At deployment, the scalar reward is revealed to be structurally decomposable into a vector reward $\mathbf{r} \in \mathbb{R}^n$. Right (A): Deploying the frozen policy $\pi$ directly yields high total return but inequitable allocation across objectives. Right (B): Our approach attaches a lightweight shaping module (QFair) that reweights action probabilities at inference time using welfare-aware Q-values conditioned on the cumulative reward vector. This achieves fair outcomes while preserving the base policy's total return, without retraining the base policy, and with the ability to adapt to new fairness objectives at deployment.

fairness objectives without retraining the base policy. For instance, when a single base policy serves many clients with heterogeneous fairness needs, or when regulatory changes demand compliance before retraining is feasible.

In contrast to the static nature of deployed RL agents, recent progress in large language models (LLMs) suggests that post-training alignment can be feasible (Kumar et al., 2025). These inference-time alignment techniques aim to steer frozen trained models toward new objectives without updating their parameters. Methods such as reward-guided decoding, representation steering, and lightweight policy adaptation have demonstrated that frozen models can be flexibly directed toward new objectives via auxiliary scoring or controlled decoding strategies (Li et al., 2023; Xu et al., 2024; Lin et al., 2025). Other methods, including RL from human feedback (RLHF) (Ouyang et al., 2022; Kaufmann et al., 2024), direct preference optimization (DPO) (Rafailov et al., 2023), and activation steering (Turner et al., 2023), further illustrate how behavior can be reshaped at deployment time through preference models or inference-time control. These techniques are effective mainly for autoregressive LLMs, where generation can be viewed as a short-horizon decision process with immediate feedback and where past decisions (generated tokens) are automatically carried forward in the model's context. However, the applicability of this approach in RL is an open problem as the objective, such as fairness, is typically defined over long horizons (e.g., expectation over trajectories) rather than as instantaneous, per-step signals. Moreover, standard deployed RL policies are Markovian, meaning that they only condition on the current state. Unlike autoregressive generation, past actions do not necessarily remain available to the policy unless they are encoded into the state. Therefore, it remains unclear whether and how inference-time alignment can solve sequential decision-making problems. This paper addresses precisely this question: ***Given a fixed, pretrained RL policy, can we steer its long-term behavior toward objectives such as fairness at deployment time, while preserving its core performance?***

Addressing this alignment problem is both practically and ethically important. From a deployment perspective, inference-time alignment reduces retraining costs, improves adaptability, and mitigates training-deployment mismatches (Mujtaba et al., 2025). From a societal perspective, learning human-aligned and fair solutions is essential in high-stakes systems, especially when these systems impact multiple end-users. Existing approaches to aligned and fair RL, including multi-objective RL (MORL) with scalarization (Siddique et al., 2020; Hayes et al., 2022), constrained Markov decision processes (CMDPs) (Bei et al., 2022), and reward shaping (Ng et al., 1999; Kumar & Yeoh, 2025), operate at training time and incorporate fairness directly into the learning objective (Jabbari et al., 2017; Weng, 2019; Siddique et al., 2023; 2024). While these approaches can produce effective solutions, they generally assume that the desired preferences are known *a priori* and require

complete retraining or maintaining a portfolio of policies for different scalarizations (Busa-Fekete et al., 2017; Zimmer et al., 2021; Do & Usunier, 2022; Yu et al., 2023a). Generalization techniques such as domain randomization (Tobin et al., 2017) can improve robustness, but they do not enable *post hoc* objective alignment. Meanwhile, policy shaping methods (Knox & Stone, 2010; Griffith et al., 2013) and inference-time alignment strategies in LLMs (Ouyang et al., 2022; Rafailov et al., 2023; Balashankar et al., 2024; Xu et al., 2024; Lin et al., 2025) demonstrate inference-time behavioral control, yet they do not formalize fairness alignment in RL via policy shaping. To date, inference-time alignment has largely remained confined to autoregressive generative models rather than long-horizon sequential control with fairness objectives.

In this paper, we bridge this gap by proposing an *inference-time fairness alignment* approach in RL via policy shaping, enabling a frozen base policy to be adapted toward fairness objectives at deployment, with only a lightweight critic trained offline. We assume access to a base policy $\pi(a \mid s)$ trained to maximize a standard scalar return, and consider settings in which the environment's feedback can be decomposed into multiple objective components that support a fairness (welfare) formulation. Our goal is to construct a shaped policy $\pi'$ that preserves the base policy's capabilities (i.e., core reward performance) while steering its *long-term* behavior toward fairness. To achieve this, we introduce a general multiplicative policy-shaping mechanism that reweights action probabilities at inference time using an action-dependent critic. We instantiate this critic as a neural network $Q_\phi$, trained offline to approximate a fairness (welfare) action-value signal. Through extensive experiments across multiple domains, we demonstrate that this inference-time shaping can steer long-term behavior toward fairness objectives while preserving the base policy's performance.

Our primary contributions are summarized as follows:

- We formalize the novel problem of inference-time fairness alignment in deep RL as a policy shaping problem (Section 4).
- We provide theoretical justification that explains when and why inference-time welfare-based shaping can improve fairness objectives without updating the base policy parameters (Section 4.3).
- We propose a multiplicative policy-shaping framework that performs test-time action reweighting using a learned fairness critic, enabling adaptation without retraining the base policy (Section 5).
- We develop **QFair**, a history-dependent fairness $Q$-network that conditions on reward vectors to address the non-stationarity inherent in optimizing non-linear welfare functions (Section 5.1).
- We provide extensive empirical validation showing that inference-time policy shaping substantially improves welfare while preserving competitive return (Section 6).

## 2 Related Work

**Inference Time Alignment.** Inference-time alignment studies how pretrained models can be steered toward new objectives at deployment without updating their core parameters. This line of work has been particularly prominent in LLMs, where frozen models are guided via decoding strategies, auxiliary reward models, or representation-level interventions (Guan et al., 2025; Lu et al., 2023; Scalena et al., 2024; Mujtaba et al., 2025). Motivated by the linear representation hypothesis, Sharkey et al. (2025) studied activation steering, which posits that high-level concepts correspond to directions in latent space; steering is performed by injecting scaled concept vectors into intermediate activations during the forward pass (Siddique et al., 2025b). Recent work further optimizes these steering vectors using RL, showing that training low-dimensional intervention vectors while freezing the backbone can rival fine-tuned models on reasoning benchmarks such as GSM8K and MATH (Sinii et al., 2025). Complementary approaches include Inference-time Policy Adapters (IPA) (Lu et al., 2023) and reward-guided decoding methods such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), as well as autoregressive reward modeling frameworks like GemARM and PARM (Xu et al., 2024; Lin et al., 2025). Despite their success, these methods are largely confined to autoregressive generation, where alignment can be framed as short-horizon scor-

ing. In contrast, we extend the paradigm of inference-time alignment to RL with trajectory-level fairness objectives.

**Policy Shaping and Test-Time Intervention.** Policy shaping and test-time intervention in RL address how learned policies can be modified without retraining, often through external guidance or auxiliary evaluators. Early work by Griffith et al. (2013) formalized policy shaping as a mechanism for incorporating human feedback into action selection, while Ferreira & Lefevre (2015) integrated social signals into dialogue management systems to refine policies online. Preference-based RL further addressed reward misspecification by inferring reward functions from pairwise human feedback (Wirth et al., 2016; Christiano et al., 2017). Human-in-the-loop corrective strategies such as COACH (Najar & Chetouani, 2021) demonstrated that lightweight action-level corrections can guide exploration without modifying the base objective. More recently, Mujtaba et al. (2025) explicitly introduced test-time policy shaping to mitigate unethical behaviors in text-based agents, using classifiers to penalize undesirable actions during deployment. In parallel, token-level steering approaches for LLMs, including SPI (Zhang et al., 2025) and decoding-time reward combination via Legendre transforms (Shi et al., 2024), highlight the broader applicability of inference-time control. Latent activation editing methods (Das et al., 2025; Sinii et al., 2025) intervene directly in representation space, whereas policy shaping operates explicitly over the action distribution. Building on this line of work, our framework formulates fairness alignment in RL as a principled inference-time policy-shaping problem for frozen policies under trajectory-level welfare objectives.

**Fairness in RL** Fairness in RL means ensuring equitable outcomes over sequential interactions, where fairness is inherently trajectory-level and temporally coupled. Foundational work in fairness in ML established formal notions such as demographic parity, equalized odds, and individual fairness (Dwork et al., 2012; Zafar et al., 2017; Agarwal et al., 2018; Speicher et al., 2018; Zhang & Liu, 2021). Beyond static prediction, distributive justice perspectives emphasize welfare-based aggregation and equitable resource allocation (Rawls, 1971; Brams & Taylor, 1996; Moulin, 2004; Heidari et al., 2018; Cousins, 2021). In RL, Jabbari et al. (2017) introduced fairness constraints on state visitation frequencies, while Jiang & Lu (2019) proposed decentralized fairness via gossip-based coordination. Actor–critic formulations incorporating fairness utilities were explored by Chen et al. (2021). Liu et al. (2018) studied the delayed impact of fair decisions, showing that static fairness criteria can harm protected groups under one-step feedback dynamics, while Icarte et al. (2018) employ reward machines to expose automaton-structured reward decompositions to the agent. However, these methods either work with static classifiers or encode fixed task structures, rather than operating at the inference-time level and steering a pre-trained policy towards fairness. MORL has provided a natural framework for fairness via scalarization of vector rewards; notably, Siddique et al. (2020) optimized the generalized Gini welfare function (GGF) in deep RL, satisfying the Pigou–Dalton transfer principle, and subsequent extensions addressed decentralized MARL and broader welfare formulations (Zimmer et al., 2021; Siddique et al., 2024; Fan et al., 2022; Do & Usunier, 2022; Yu et al., 2023b; Qian et al., 2025; Michailidis et al., 2024; Siddique et al., 2025a; 2026). These approaches, however, assume that fairness preferences are specified during training or require explicitly learning fair policies. In contrast, our work departs from training-time fairness optimization and instead enables fairness alignment at deployment, steering a pretrained reward-maximizing policy toward welfare-based objectives without updating the base policy's parameters.

## 3 Preliminary

### 3.1 Reinforcement Learning with Decomposable Rewards

We consider finite-horizon Markov decision processes (MDPs) in which the environment feedback admits a natural decomposition across multiple objectives (e.g., stakeholders, resource types, or user groups). Concretely, we consider *multi-objective MDP* (MOMDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \boldsymbol{r}, T, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is a finite action space, $P(s' \mid s, a)$ is the transition kernel, $T$ is the horizon, and $\gamma \in (0, 1]$ is the discount factor. Here, the re-

ward is a vector $\boldsymbol{r} = (r_1, r_2, ..., r_N)$, where each component $r_i$ corresponds to objective $i$. The corresponding (discounted) return vector for a trajectory is given by, $\sum_{t=0}^{T-1} \gamma^t \boldsymbol{r}(s_t, a_t)$. In many domains, the scalar reward used for training is a simple decomposition of these components. In particular, a common choice is the utilitarian sum, which can be defined as $r(s,a) = h(\boldsymbol{r}(s,a)) = \mathbf{1}^\top \boldsymbol{r}(s,a) = \sum_{i=1}^{N} r_i(s,a)$, yielding a standard scalar-reward MDP for training. Given a scalar reward $r$, the usual RL objective is to maximize expected discounted return $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)\right]$. The state-value and action-value functions under $\pi$ are given by, $V^\pi(s) = \mathbb{E}_\pi \left[\sum_{k=t}^{T-1} \gamma^{k-t} r_k \;\middle|\; s_t = s\right]$, $Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{k=t}^{T-1} \gamma^{k-t} r_k \;\middle|\; s_t = s,\, a_t = a\right]$. When the state or action spaces are large, deep RL methods approximate these quantities with neural networks. Value-based methods such as DQN (Mnih et al., 2015) estimate $Q^\pi$ and derive policies via an $\epsilon$-greedy approach. Policy-gradient methods directly optimize parameterized policies $\pi_\theta$ via gradient ascent on $J(\pi_\theta)$ Sutton & Barto (1998). Actor–critic methods such as A2C (Mnih et al., 2016) combine both approaches by learning a critic $V_\phi$ to reduce variance in policy-gradient updates, while PPO (Schulman et al., 2017) stabilizes training via clipped surrogate objectives. In this paper, these algorithms are used only to obtain pretrained, reward-maximizing base policies; consequently, we adapt these policies at inference time without retraining.

### 3.2 Fairness Formulation

To evaluate fairness across objectives, we adopt the welfare-function framework from distributive justice and multi-objective optimization (Rawls, 1971; Moulin, 2004; Speicher et al., 2018; Siddique et al., 2020). A welfare function $\phi_{\boldsymbol{\omega}} : \mathbb{R}^N \to \mathbb{R}$ aggregates the return vector into a scalar and provides a measure of social desirability. Following prior work (Weng, 2019; Siddique et al., 2020; Zimmer et al., 2021), we require $\phi_{\boldsymbol{\omega}}$ to satisfy three axiomatic properties, namely efficiency, impartiality, and equity. Linear aggregation functions such as $\mathbf{1}^\top \boldsymbol{u}$ satisfy efficiency and impartiality but generally fail to satisfy equity, motivating non-linear welfare functions. Therefore, in this paper, we focus on the *generalized Gini welfare function* (GGF) (Weymark, 1981), defined as

$$\phi_{\boldsymbol{\omega}}(\boldsymbol{u}) = \sum_{i=1}^{N} w_i u_i^{\uparrow}, \tag{1}$$

where $\boldsymbol{u}^{\uparrow}$ denotes the sorted utility vector in ascending order and $\boldsymbol{w} \in \mathbb{R}^N_{>0}$ is a strictly decreasing positive weight vector. GGF satisfies all the above properties as it is symmetric, Pareto-monotonic, and Schur-concave. It interpolates between utilitarian and max–min objectives through the choice of weights. While our framework applies to any concave welfare function satisfying the axioms above, we adopt GGF throughout as a canonical choice. When evaluating fairness, an important distinction arises in how the expectation is applied to the GGF. While prior works (Fan et al., 2022) optimize the expected welfare, this formulation cannot theoretically ensure fairness due to the non-linearity of the welfare function. Instead, to properly guarantee the fair treatment across objectives, we must optimize the welfare of expected return. Thus, our fairness objective can be expressed as,$J_{\boldsymbol{\omega}}(\pi) = \phi_{\boldsymbol{\omega}}(\boldsymbol{J}(\pi))$. Since $\phi_{\boldsymbol{\omega}}$ is non-linear, evaluating the welfare of expected returns is inherently a trajectory-level property that cannot be decomposed into per-step scalar rewards. Consequently, a base policy strictly optimized for the step-wise utilitarian objective $J(\pi)$ generally fails to maximize $J_{\mathbf{w}}(\pi)$. This highlights the necessity for inference-time alignment at deployment.

## 4 Inference-Time Fairness Alignment

We consider the problem of aligning or shaping a pretrained RL policy toward fair outcomes at inference time without updating its learned parameters. The key observation is that, although the base policy was trained with a scalar reward $r = h(\boldsymbol{r})$, the environment exposes a richer *vector* reward $\boldsymbol{r} \in \mathbb{R}^N$ that can be leveraged post hoc to optimize a welfare-based fairness objective.

**Assumptions.** We require that (i) a frozen pretrained base policy $\pi(a \mid s)$ whose action distribution (or action values) can be queried, (ii) observability of the vector reward $\boldsymbol{r} \in \mathbb{R}^N$ at deployment, i.e., the scalar training reward is structurally decomposable and its components are exposed by the environment, and (iii) an offline phase in which QFair is trained from base-policy rollouts with light exploration. Hence, no welfare-driven environment interaction and base policy updates are required.

Formally, our inference-time alignment problem can be written as

$$\max_{\pi'} J_{\boldsymbol{\omega}}(\pi') = \max_{\pi'} \phi_{\boldsymbol{\omega}}\left(\boldsymbol{J}(\pi')\right), \tag{2}$$

where the expectation is taken with respect to the shaped policy $\pi'$. The goal is therefore to find a policy $\pi'$ that maximizes the welfare objective $J_{\boldsymbol{\omega}}(\pi')$ subject to the constraint that $\pi'$ is obtained strictly by reshaping the action distribution of $\pi$ at inference time. However, solving this problem introduces some unique challenges. Since $\phi_{\boldsymbol{\omega}}$ is non-linear, it cannot in general be expressed as a sum of per-step scalar rewards on the original state space. Moreover, the welfare objective depends on the cumulative return vector accumulated along the trajectory, rather than solely on the current environment state. As a result, the welfare-optimal decision at time $t$ depends on the history of rewards obtained so far. Our approach addresses these challenges through a sequence of principled reductions. First, we show that the non-linearity of $\phi_{\boldsymbol{\omega}}$ induces non-Markovianity in the original MDP, because the welfare objective depends on the full trajectory return. Second, we construct a welfare-augmented MDP $\widetilde{\mathcal{M}}$ that restores the Markov property (Section 4.1). Third, we formally relate the pretrained base policy to this augmented environment through a policy lifting construction. Finally, within $\widetilde{\mathcal{M}}$, we derive the shaped policy in closed form as the solution to a KL-regularized welfare surrogate that locally lower-bounds the true welfare improvement (Section 4.3). We then show how this solution can be estimated in practice using a welfare critic learned from base-policy trajectories. The full proofs of our propositions and theorems are provided in Appendix A.

### 4.1 Restoring Markovianity via a Welfare-Augmented MDP

The core difficulty in optimizing $J_{\boldsymbol{\omega}}(\pi')$ is that the welfare function $\phi_{\boldsymbol{\omega}}$ is a non-linear function applied to the full trajectory return. Consequently, the objective cannot be decomposed into independent per-step rewards. This means that the welfare-optimal action at time $t$ depends not only on the current state $s_t$ but also on the history of rewards already accrued. For simplicity, we hereafter assume $\gamma = 1$ and thus $J$ becomes the cumulative sum of rewards, denoted as $\boldsymbol{R}$.

**Proposition 4.1** (Non-Markovianity of Welfare Optimization)**.** *When $\phi_{\boldsymbol{\omega}}$ involves sorting (e.g., the GGF defined in (1)), the welfare-optimal policy is generally not Markovian in $\mathcal{S}$. In particular, there exist states $s$ and distinct reward histories $\boldsymbol{R} \neq \boldsymbol{R}'$ for which the welfare-optimal actions at $(s, \boldsymbol{R})$ and $(s, \boldsymbol{R}')$ are different.*

**Example 4.1** *To illustrate Proposition 4.1, we consider an MDP with $N = 2$ objectives and GGF weights $w_1 > w_2 > 0$. As seen in Figure 2, at state $s_{t-1}$, action $a_1$ yields $\boldsymbol{r}_1 = (1, 0)$ and $a_2$ yields $\boldsymbol{r}_2 = (0, 1)$, both leading to a terminal state $s_t$. If the accrued reward in this trajectory is $\boldsymbol{R} = (10, 2)$, then $a_2$ produces a final return of $(10, 3)$, which sorts $(3, 10)$ yielding a welfare score of $3w_1 + 10w_2$. Action $a_1$ produces $(11, 2)$, yielding a welfare score of $2w_1 + 11w_2$. Since $w_1 > w_2$ and the difference is $w_1 - w_2 > 0$, $a_2$ is optimal. Conversely, if $\boldsymbol{R}' = (2, 10)$, the same analysis shows $a_1$ is optimal. Thus, the optimal action at the* same state $s$ *reverses depending on cumulative reward $\boldsymbol{R}$.*

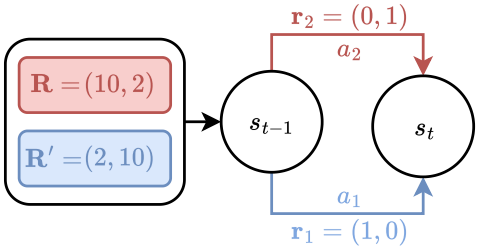



Figure 2: Example where different cumulative rewards lead to different optimal actions.

To restore the Markov property, we enrich the state representation with the cumulative reward vector and define a welfare-augmented MDP.

**Definition 4.1** (Welfare-Augmented MDP)**.** *Given an MOMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \boldsymbol{r}, T, \gamma)$ and welfare function $\phi_{\boldsymbol{\omega}}$, define the undiscounted finite-horizon MDP $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{S}}, \mathcal{A}, \widetilde{P}, \tilde{r}, T)$ where the state space is augmented with the cumulative reward vector, $\widetilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}^N$, providing an augmented state is $\tilde{s}_t = (s_t, \boldsymbol{R}_t)$. The transition dynamics update the accumulator deterministically: $\widetilde{P}\big(\tilde{s}' \mid \tilde{s}, a\big) = \widetilde{P}((s', \mathbf{R}') \mid (s, \mathbf{R}), a) = P(s' \mid s, a) \cdot \mathbf{1}[\mathbf{R}' = \mathbf{R} + \mathbf{r}(s, a)]$. The scalar reward in $\widetilde{\mathcal{M}}$ is defined as the marginal welfare contribution $\tilde{r}\big((s, \boldsymbol{R}),\, a\big) \;=\; \phi_{\boldsymbol{w}}\big(\boldsymbol{R} + \boldsymbol{r}(s, a)\big) - \phi_{\boldsymbol{w}}(\boldsymbol{R})$.*

**Theorem 4.1** (Augmented MDP Equivalence)**.** *$\widetilde{\mathcal{M}}$ is Markovian, and for any policy $\tilde{\pi}$ on $\widetilde{\mathcal{S}}$:*

$$\mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{T-1} \tilde{r}(\tilde{s}_t, a_t)\right] \;=\; \mathbb{E}_{\tilde{\pi}}\big[\phi_{\boldsymbol{w}}(\boldsymbol{R}_T)\big] - \phi_{\boldsymbol{w}}(\mathbf{0}). \tag{3}$$

*Hence, maximizing the scalar return in $\widetilde{\mathcal{M}}$ is equivalent to maximizing the welfare objective $J_{\boldsymbol{\omega}}(\pi')$.*

The augmented MDP reduces the non-Markovian welfare optimization to a standard scalar-reward MDP. We define the welfare value functions in $\widetilde{\mathcal{M}}$: $\widetilde{V}^{\tilde{\pi}}(\tilde{s}) = \mathbb{E}_{\tilde{\pi}}\left[\sum_{k=t}^{T-1} \tilde{r}_k \;\middle|\; \tilde{s}_t = \tilde{s}\right]$, $\widetilde{Q}^{\tilde{\pi}}(\tilde{s}, a) = \tilde{r}(\tilde{s}, a) + \mathbb{E}_{s'}[\widetilde{V}^{\tilde{\pi}}(\tilde{s}')]$, and the welfare advantage $\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a) = \widetilde{Q}^{\tilde{\pi}}(\tilde{s}, a) - \widetilde{V}^{\tilde{\pi}}(\tilde{s})$.

**Remark 4.1.** *Theorem 4.1 extends to $\gamma < 1$ by augmenting the state with the discounted cumulative vector $\boldsymbol{R}_t = \sum_{k=0}^{t-1} \gamma^k \boldsymbol{r}(s_k, a_k)$ together with the running discount $\gamma^t$, and defining the marginal welfare reward $\tilde{r}\big((s, \boldsymbol{R}),\, a\big) = \phi_{\boldsymbol{w}}\big(\boldsymbol{R} + \gamma^t \boldsymbol{r}(s, a)\big) - \phi_{\boldsymbol{w}}(\boldsymbol{R})$. The telescoping argument in the proof of Theorem 4.1 remains unchanged, and the augmented return again equals $\mathbb{E}_{\tilde{\pi}}\big[\phi_{\boldsymbol{w}}(\boldsymbol{R}_T)\big] - \phi_{\boldsymbol{w}}(\mathbf{0})$.*

### 4.2 Anchoring on the base policy via policy lifting

Since the base policy $\pi$ is trained in $\mathcal{M}$ with scalar reward, but its shaped policy $\pi'$ must operate in $\widetilde{\mathcal{M}}$, we therefore define its canonical embedding into the augmented state space. Note that both of these MDPs share the same dynamics and action space but differ fundamentally in state representation and objective. To reason about $\pi$ within $\widetilde{\mathcal{M}}$, its canonical embedding is defined as below.

**Definition 4.2** (Lifting)**.** *The lifted policy $\tilde{\pi}$ on $\widetilde{\mathcal{S}}$ is $\tilde{\pi}(a \mid \tilde{s}) := \pi(a \mid s), \forall \tilde{s} = (s, \boldsymbol{R}) \in \widetilde{\mathcal{S}},\ a \in \mathcal{A}$.*

While $\tilde{\pi}$ is generally a suboptimal policy in $\widetilde{\mathcal{M}}$ (because it ignores the $\boldsymbol{R}$-component necessary for welfare optimality), it perfectly preserves the base policy's interaction with the environment.

**Proposition 4.2** (Trajectory Equivalence)**.** *Running $\pi$ in original MDP $\mathcal{M}$ and running $\tilde{\pi}$ in augmented MDP $\widetilde{\mathcal{M}}$ induce same distributions over base-state trajectories $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$. That is, $\Pr^{\pi}_{\mathcal{M}}(\tau) \;=\; \Pr^{\tilde{\pi}}_{\widetilde{\mathcal{M}}}(\tau)$, where $\Pr(\tau)$ is the probability of an entire trajectory $\tau$ occurring.*

A direct and powerful consequence of Proposition 4.2 is that rollouts collected by the base policy can be *directly reused* for offline training in $\widetilde{\mathcal{M}}$. We simply annotate each transition with the cumulative reward vector $\boldsymbol{R}_t$ (computed post-hoc from the recorded vector rewards) and the marginal welfare reward. Furthermore, although $\pi$ was trained to maximize $J(\pi)$, it induces a well-defined welfare $J_{\boldsymbol{\omega}}(\pi)$ that provides a guaranteed performance baseline. The following result quantifies why a good base policy is a good starting point for welfare improvement.

**Proposition 4.3** (Welfare Floor)**.** *If the training aggregation is the utilitarian sum ($h = \mathbf{1}^\top$), then for any policy $\pi$:*

$$J_{\boldsymbol{\omega}}(\pi) \;\geq\; w_N \cdot J(\pi), \tag{4}$$

*where $w_N > 0$ is the smallest GGF weight.*

Essentially, these results, combined with Theorem 4.1, provide welfare improvement achievable by any shaped policy $\pi'$: $J_{\boldsymbol{\omega}}(\pi') - J_{\boldsymbol{\omega}}(\pi) = \widetilde{J}(\tilde{\pi}') - \widetilde{J}(\tilde{\pi})$, where $\widetilde{J}(\cdot)$ denotes the expected return in $\widetilde{\mathcal{M}}$. This identity provides a formal bridge, which is *improving fairness over the base policy is mathematically equivalent to a standard policy improvement problem over the lifted policy in $\widetilde{\mathcal{M}}$.*

### 4.3 Optimal KL-Regularized Inference-Time Policy Shaping

We now derive the shaped policy as the closed-form solution to a KL-regularized surrogate optimization problem in $\widetilde{\mathcal{M}}$. The KL-regularization serves two purposes. First, it keeps the shaped policy $\pi'$ close to the original policy $\pi$, and second, it ensures the surrogate objective remains a valid approximation to the true welfare gain. Solving this optimization problem and using the performance difference lemma (Kakade & Langford, 2002) applied to $\widetilde{\mathcal{M}}$, we can express the welfare gap as $\widetilde{J}(\tilde{\pi}') - \widetilde{J}(\tilde{\pi}) = T\mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}'}}\Big[\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a)\Big]$. Since $d^{\tilde{\pi}'}$ depends on the unknown policy, we replace it with $d^{\tilde{\pi}}$ to obtain a tractable welfare surrogate that shares the same gradient at $\tilde{\pi}' = \tilde{\pi}$ and provides a local lower bound on the true improvement under a KL penalty (Schulman et al., 2015). Therefore, to maximize this objective while preserving the environmental competence guaranteed by the welfare floor, we restrict the shaped policy to a KL-divergence trust region around $\tilde{\pi}$:

$$\max_{\tilde{\pi}'} \quad T\,\mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}}}\Big[\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a)\Big] \tag{5}$$

$$\text{s.t.} \quad \mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}}}\Big[\mathrm{KL}\big(\tilde{\pi}'(\cdot|\tilde{s}) \,\big\|\, \tilde{\pi}(\cdot|\tilde{s})\big)\Big] \le \varepsilon, \qquad \tilde{\pi}'(\cdot|\tilde{s}) \in \Delta(\mathcal{A}) \;\; \forall\, \tilde{s}, \tag{6}$$

We emphasize that problem (5)-(6) is a surrogate objective as it maximizes a local lower bound on the true welfare improvement, valid within the KL trust region (Schulman et al., 2015), rather than the true objective (2). The following closed form is therefore optimal for the surrogate, and constitutes one step of KL-regularized policy improvement over the lifted base policy.

**Theorem 4.2** (Optimal Shaped Policy)**.** *The unique solution to* (5)-(6) *is:*

$$\boxed{\pi'(a \mid s, \boldsymbol{R}) \;=\; \frac{\pi(a \mid s)\;\exp\Big(\frac{1}{\beta}\,\widetilde{Q}^{\tilde{\pi}}\big((s, \boldsymbol{R}),\, a\big)\Big)}{\displaystyle\sum_{a'} \pi(a' \mid s)\;\exp\Big(\frac{1}{\beta}\,\widetilde{Q}^{\tilde{\pi}}\big((s, \boldsymbol{R}),\, a'\big)\Big)}} \tag{7}$$

*where $\beta > 0$ is the Lagrange multiplier for the KL constraint, determined by $\varepsilon$. The shaped policy $\pi'$ is computed directly from $\pi$ and the welfare value function $\widetilde{Q}^{\tilde{\pi}}$ with no iterative update.*

## 5 Training Welfare Critic and Inference-Time Shaping

To solve the problem (2) of finding a fair policy at inference, an agent must evaluate the welfare action-value function $\widetilde{Q}^{\tilde{\pi}}(\tilde{s}, a)$ derived in Theorem 4.2. Because computing this quantity exactly is intractable in environments with large or continuous state spaces, we introduce **QFair**, a lightweight neural network welfare critic, parametrized by $Q_\phi$ that learns to approximate $\widetilde{Q}^{\tilde{\pi}}$ from offline data. The design and integration of QFair bridge our theoretical framework developed in Section 4 and enables practical inference-time fairness alignment with standard deep RL algorithms. Crucially, QFair is the only component in our framework that requires training. Once learned, it is frozen and utilized purely for inference-time scoring.

### 5.1 Learning the Welfare Critic (QFair)

We train QFair to approximate the welfare Q-function in $\widetilde{\mathcal{M}}$ from offline data, using an architecture that is influenced by the non-Markovianity of the welfare objective (as established in Proposition 4.1). A critic conditioned only on the base state $s$ cannot correctly assess welfare contributions, as the same action can yield opposite fairness impacts depending on the cumulative reward. Therefore, $Q_\phi$ is constructed to take the fully augmented state $\tilde{s} = (s, \boldsymbol{R})$ as input. The network outputs a welfare score for each action, which approximates $\widetilde{Q}^{\tilde{\pi}}(\tilde{s}, a)$. We parameterize $Q_\phi$ as a three-layer MLP in which the state and accrued reward are concatenated at the input layer.

To train QFair from base policy rollouts with light exploration and without requiring additional environment interactions, we rely on the trajectory equivalence result established in Proposition 4.2.

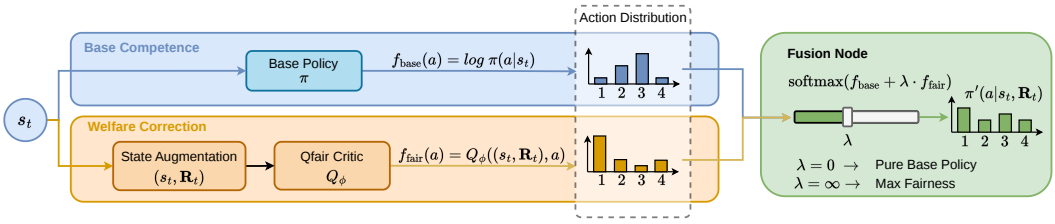



Figure 3: Inference-time policy shaping pipeline. At each step, the current state $s_t$ is processed along two parallel paths. The base policy produces the base competence signal, while the QFair critic receives the augmented state and outputs the welfare correction signal . The fusion node combines both signals with a softmax function, where the alignment strength $\lambda$ interpolates continuously between the pure base policy and maximum fairness. Both the base policy and the QFair critic are frozen and no parameters are updated during deployment.

This result shows that the trajectories from the base policy in original MDP $\mathcal{M}$ induce the same state-action distribution as the lifted policy in the augmented MDP $\widetilde{\mathcal{M}}$. Therefore, data collected in the original environment can be reused to train QFair. To collect offline data, we roll out the behavior policy $\mu(a \mid s) = (1 - \beta_{\text{explore}})\ \pi(a \mid s) + \beta_{\text{explore}}\ (1/|\mathcal{A}|)$ that mixes the frozen base policy with uniform exploration. Each transition is then annotated post-hoc with the running accumulator $\boldsymbol{R}_t$ and the marginal welfare reward $\tilde{r}_t = \phi_{\boldsymbol{w}}(\boldsymbol{R}_{t+1}) - \phi_{\boldsymbol{w}}(\boldsymbol{R}_t)$, yielding a dataset $\mathcal{D} = \{(\tilde{s}_t, a_t, \tilde{r}_t, \tilde{s}_{t+1}, d_t)\}$ of transitions in $\widetilde{\mathcal{M}}$. Note that light uniform exploration is included only because near-deterministic base policies under-explore the augmented state space, which is insufficient for TD generalization.

Because Theorem 4.1 guarantees that $\widetilde{\mathcal{M}}$ is a proper MDP, the Bellman equation holds, and we can train QFair using TD learning $\widetilde{Q}^{\mu}(\tilde{s}, a) = \tilde{r}(\tilde{s}, a) + \gamma\, \mathbb{E}_{s' \sim P}\Big[\sum_{a'} \mu(a'|s')\, \widetilde{Q}^{\mu}(\tilde{s}', a')\Big]$. Then we minimize the mean-squared Bellman error with Expected SARSA targets:

$$y = \tilde{r} + \gamma(1 - d) \sum_{a'} \mu(a'|s')\, Q_\phi(\tilde{s}', a'), \qquad \mathcal{L}(\phi) = \mathbb{E}_{(\tilde{s}, a, \tilde{r}, \tilde{s}', d) \sim \mathcal{D}}\Big[(Q_\phi(\tilde{s}, a) - y)^2\Big]. \tag{8}$$

Expected SARSA averages over the behaviour policy's action distribution at $\tilde{s}'$, avoiding the maximization bias of Q-learning while remaining consistent with the data collection process. At convergence, $Q_\phi$ approximates the welfare Q-function $\widetilde{Q}^{\mu}$ in $\widetilde{\mathcal{M}}$, which is close to $\widetilde{Q}^{\tilde{\pi}}$ and required for policy shaping. The full algorithm for training QFair is provided in Algorithm 1 (see Appendix B).

### 5.2 Inference-time Policy Shaping

With QFair trained offline, the final step is to steer the pretrained policy. At inference time, the agent observes $s_t$ and maintains a running accumulator $\boldsymbol{R}_t$ (initialised to $\mathbf{0}$ at episode start), and computes the shaped action distribution via a softmax combination:

$$\pi'(a \mid s_t, \boldsymbol{R}_t) \;=\; \text{softmax}\Big(f_{\text{base}}(a) + \lambda\, f_{\text{fair}}(a)\Big), \tag{9}$$

where the two signal components are $f_{\text{base}}(a)$ which is a standard RL base policy and $f_{\text{fair}}(a) = \bar{Q}_\phi\big((s_t, \boldsymbol{R}_t),\, a\big)$. To ensure numerical stability and comparable scaling between the two signals, $\bar{Q}_\phi$ denotes the welfare scores normalized to zero-mean and unit-variance across actions.

This softmax formulation mathematically implements the theoretical exponential reweighting derived in Eq. (7). Thus, this formulation provides a general framework for policy shaping at test time and is compatible with both value-based and policy gradient methods. In our experiments, we demonstrate the validity of this framework by successfully aligning policies trained with DQN, A2C, and PPO. The full algorithm of inference-time policy shaping is provided in Algorithm 2.

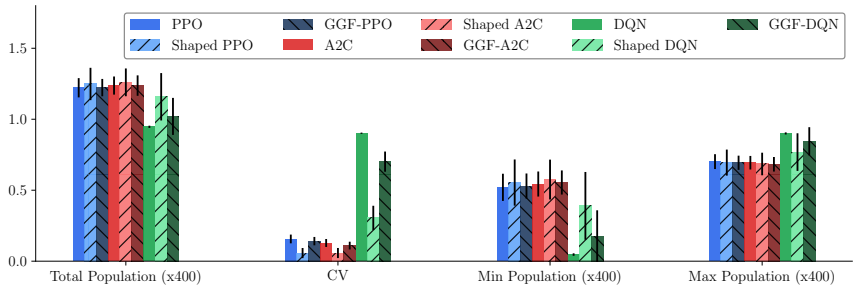



Figure 4: Performance comparison of PPO, A2C, and DQN with their inference-time shaped variants and GGF-trained counterparts in the Species Conservation environment in terms of total reward, CV, min population, and max population

## 6 Experiments

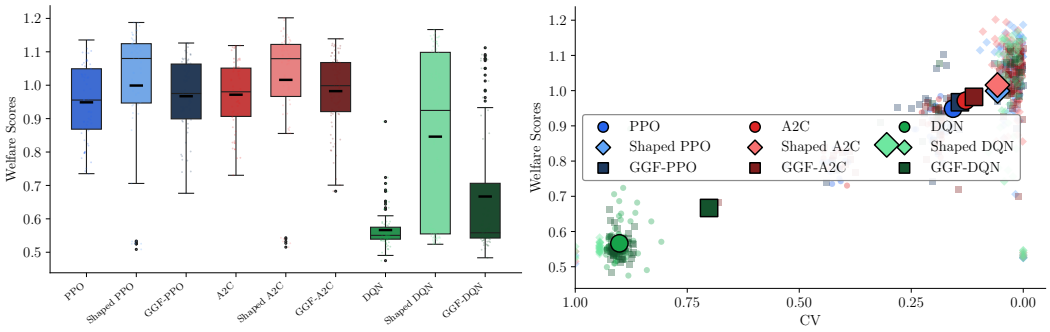
We evaluate our framework across three domains of increasing complexity and number of objectives. In all environments, fairness is defined at the user (objective) level. During training, standard RL agents optimize the scalar sum of rewards. At inference time, we decompose this scalar reward into its vector components and apply our shaping mechanism to enforce welfare-based fairness. To ensure a comprehensive evaluation, we select domains with varying numbers of objectives and dynamics, making fairness non-trivial. The domains include species conservation, four-room environment, and harvest–regrow. Each domain introduces distinct structural challenges where balancing objectives is essential. Due to space constraints, the environment description and full empirical results for the four-room domain have been deferred to Supplementary Materials C.

### 6.1 Setup and Baselines

We reshape three widely used standard deep RL algorithms, including PPO, A2C, and DQN, covering both policy-based and value-based methods. For each algorithm, we compare (i) standard RL algorithms trained to maximize cumulative scalar reward, (ii) GGF adaptations for PPO, A2C, and DQN that are fairness baselines trained directly to optimize the GGF welfare function following Siddique et al. (2020), and (iii) Shaped algorithms for PPO, A2C, and DQN that were trained for reward maximization and aligned toward GGF at inference time via policy shaping. The GGF-based methods serve as true fairness baselines, as they directly optimize the welfare objective during training. In contrast, our shaped agents are trained purely for reward maximization and only steered toward welfare objectives at deployment. All algorithms are evaluated over 20 random seeds in Species Conservation, 10 in Four-Room, and 5 in Harvest–Regrow. Hyperparameters are tuned using Optuna, and full configurations are reported in Appendix D.

### 6.2 Species Conservation

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 (Chadès et al., 2012), which simulates the predation relationship between the species, where sea otters prey on abalones. The state space is composed of the current population sizes of sea otters and northern abalones. The action space includes introducing sea otters, enforcing anti-poaching measures, controlling sea otter populations, implementing a combination of half-antipoaching and half-controlled sea otters, or taking no action. The reward function is defined by the population densities of both species, i.e., $N = 2$. Fairness in this context is interpreted as achieving a balanced distribution of species densities to ensure their preservation. During training, agents optimize the scalar sum of species populations.




(a) Welfare scores.

(b) Pareto front plot.

Figure 5: Performance comparison of PPO, A2C, DQN, their inference-time shaped variants, and GGF-trained counterparts in the Species Conservation.

At inference time, we decompose this sum into its two components and apply policy shaping toward the GGF objective as described in Section 5.

Figure 4 reports total reward, coefficient of variation (CV), minimum population, and maximum population across different standard RL algorithms, GGF-based methods, and our Shaped RL algorithms. Our results show that standard PPO and A2C achieve higher total reward than DQN. As expected, standard RL methods that are trained to maximize reward tend to increase the dominant species population while ignoring the weaker species, resulting in high maximum values and low minimum values. In contrast, GGF-based methods achieve lower CV and higher minimum population levels in comparison to their standard RL algorithms. This means that GGF-based methods achieve more balanced ecological outcomes. Since lower CV corresponds to a more equitable distribution, this confirms that GGF-based training enforces fairness at the expense of some reward optimality. Our shaped agents consistently achieve the lowest CV among all methods while maintaining competitive total reward. In particular, they produce the highest minimum population levels, indicating improved protection of the worst-off species. These results demonstrate that inference-time shaping can yield solutions that are both balanced and performance-preserving.

Figure 5a presents welfare scores across 100 evaluation trajectories. For each experiment, we compute the empirical average vector return and apply the welfare function to obtain the final score. Our results show that, although GGF-based methods are trained to directly optimize welfare, our shaped agents often achieve comparable or, in some cases, higher welfare scores. This indicates that inference-time shaping effectively approximates welfare optimization without retraining. Figure 5b shows the Pareto trade-off between CV (x-axis) and welfare score (y-axis). Lower CV indicates more balanced distributions. The results show that our shaped RL methods shift the Pareto frontier toward the desirable region of lower inequality and higher welfare. Notably, Shaped-DQN achieves a substantially lower CV than both standard and GGF-trained DQN while maintaining higher welfare. PPO and A2C show similar improvements. These results highlight that inference-time shaping can move pretrained agents toward fairness without modifying learned parameters.

### 6.3 Four Room

Our second domain is a four-room (FR) grid environment, which models a multi-objective navigation and resource-collection problem under spatial constraints and stochastic outcomes. The environment is a $13 \times 13$ maze divided by walls into four connected regions, with multiple spawn locations and three resource types distributed across the map. These resources are represented by distinct shapes and colors: a blue square, a red circle, and a green triangle and the dynamics of these resources are intentionally heterogeneous to create imbalance among objectives. Details of this environment and its empirical results are provided in Supplementary Materials C.

Table 1: Performance comparison in Harvest-Regrow. We report total reward, CV, minimum objective return, and maximum objective return (mean ± std over evaluation runs).

| Algorithm | Total reward | CV | Min crop reward | Max crop reward |
|---|---|---|---|---|
| PPO | 700.32 ± 9.03 | 0.9997 ± 0.0003 | 0.01 ± 0.02 | 700.15 ± 9.04 |
| Shaped PPO | 804.14 ± 19.11 | 0.8797 ± 0.0005 | 0.03 ± 0.09 | 723.54 ± 17.25 |
| GGF-PPO | 483.12 ± 34.14 | 0.2783 ± 0.1289 | 52.12 ± 27.75 | 203.46 ± 55.35 |
| A2C | 599.80 ± 155.73 | 0.9995 ± 0.0004 | 0.01 ± 0.02 | 599.59 ± 155.71 |
| Shaped A2C | 694.97 ± 116.38 | 0.9194 ± 0.0008 | 0.02 ± 0.07 | 625.19 ± 104.67 |
| GGF-A2C | 613.50 ± 99.95 | 0.9996 ± 0.0004 | 0.00 ± 0.00 | 613.33 ± 100.04 |
| DQN | 529.89 ± 30.63 | 0.9461 ± 0.1115 | 0.09 ± 0.11 | 505.61 ± 68.94 |
| Shaped DQN | 638.92 ± 25.28 | 0.9267 ± 0.0035 | 0.22 ± 0.32 | 573.61 ± 22.83 |
| GGF-DQN | 618.93 ± 69.12 | 0.8730 ± 0.1921 | 0.05 ± 0.08 | 558.65 ± 157.50 |

### 6.4 Harvest Regrow

Our third and final domain is a harvest-regrow (HR) environment, which captures a resource allocation challenge in a spatial ecosystem, where the goal of the agent is to collect resources from multiple renewable but heterogeneous resource types while avoiding strong imbalances across objectives. The environment is a grid of $10 \times 10$ with four crop patches. These crops include apple, melon, berry, and wheat, each with distinct scarcity/value dynamics due to their varying regrowth rates and rewards. In particular, melon yields higher rewards but regrows slowly; wheat provides lower rewards but regenerates quickly; and berry yields are stochastic. In this domain, the agent's state consists of its current grid position together with four binary harvest indicators specifying whether each crop type has been collected at the current location. The action space comprises the four cardinal movements (up, down, left, right). Rewards are vector-valued with $N = 4$, corresponding to per-step crop rewards $[r_{\text{apple}}, r_{\text{melon}}, r_{\text{berry}}, r_{\text{wheat}}]$. Fairness in this setting is interpreted as balanced performance across crop objectives, discouraging policies that over-exploit only the most accessible or highest-value crops. During training, agents optimize the scalarized sum of all crop rewards. At inference time, we decompose the scalar reward into its four components and apply policy shaping toward the GGF objective, as described in Section 5.

Table 1 reports the total reward, CV, minimum crop return, and maximum crop return during the testing phase. The results show that although CV values remain high overall, inference-time shaping consistently reduces inequality relative to the corresponding base algorithms. For PPO, A2C, and DQN, the shaped variants achieve lower CV and higher total reward than their standard algorithms. This suggests that inference-time alignment improves fairness without sacrificing efficiency, and in fact yields higher overall returns. Among all methods, GGF-PPO achieves the lowest CV and higher welfare, as expected since it is trained directly to optimize the GGF objective. However, this highest fairness comes at a cost as it has the lowest total reward. In contrast, our shaped methods create a trade-off and learn solutions that are not only fair but also efficient.

## 7 Conclusions

In this paper, we address the problem of inference-time fairness alignment in deep RL. By formalizing fairness alignment as a policy shaping problem, we achieve test-time adaptation without retraining the base policy, using our novel history-dependent welfare critic, **QFair**. We instantiate this multiplicative shaping method across multiple deep RL algorithms and demonstrate through extensive experiments that inference-time policy shaping substantially improves welfare objectives while preserving competitive task performance. In the future, we plan to extend this inference-time shaping framework to accommodate dynamic preferences beyond welfare-based fairness, such as continuous human alignment or interactive human-guided adaptation.

**Acknowledgments**

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

## References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018.

Anurag Ajay, Abhishek Gupta, Dibya Ghosh, Sergey Levine, and Pulkit Agrawal. Distributionally adaptive meta reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 25856–25869, 2022.

Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, Adrian Hutter, Jong Lee, Chirag Nagpal, Flavien Prost, Aradhana Sinha, et al. Infalign: Inference-aware language model alignment. *arXiv preprint arXiv:2412.19792*, 2024.

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

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

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. Multi-objective bandits: Optimizing the generalized gini index. In *ICML*, pp. 625–634, 2017.

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

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017.

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

Satyajeet Das, Darren Chiu, Zhehui Huang, Lars Lindemann, and Gaurav S Sukhatme. Latent activation editing: Inference-time refinement of learned policies for safer multirobot navigation. *arXiv preprint arXiv:2509.20623*, 2025.

Virginie Do and Nicolas Usunier. Optimizing generalized gini indices for fairness in rankings. *arXiv preprint arXiv:2204.06521*, 2022.

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

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

Emmanuel Ferreira and Fabrice Lefevre. Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards. *Computer Speech & Language*, 34(1):256–274, 2015.

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26, 2013.

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. A survey on personalized alignment—the missing piece for large language models in real-world applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5313–5333, 2025.

Tom Haider, Karsten Roscher, Felippe Schmoeller da Roza, and Stephan Günnemann. Out-of-distribution detection for reinforcement learning agents with probabilistic dynamics models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 851–859, 2023.

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.

Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *NeurIPS*, 2018.

Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, pp. 2107–2116. PMLR, 2018.

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

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

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

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2024.

Woosung Kim, Jinho Lee, Jongmin Lee, and Byung-Jun Lee. Fairdice: Fairness-driven offline multi-objective reinforcement learning. *arXiv preprint arXiv:2506.08062*, 2025.

W Bradley Knox and Peter Stone. Combining manual feedback with subsequent mdp reward signals for reinforcement learning. In *AAMAS*, volume 10, pp. 5–12, 2010.

Ashwin Kumar and William Yeoh. Decaf: Learning to be fair in multi-agent resource allocation. *arXiv preprint arXiv:2502.04281*, 2025.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Baijiong Lin, Weisen Jiang, Yuancheng Xu, Hao Chen, and Ying-Cong Chen. Parm: Multi-objective test-time alignment via preference-aware autoregressive reward model. *arXiv preprint arXiv:2505.06274*, 2025.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Ximing Lu, Faeze Brahman, Peter West, Jaehun Jung, Khyathi Chandu, Abhilasha Ravichander, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, et al. Inference-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6863–6883, 2023.

Dimitris Michailidis, Willem Röpke, Diederik M Roijers, Sennay Ghebreab, and Fernando P Santos. Scalable multi-objective reinforcement learning with fairness guarantees using lorenz dominance. *arXiv preprint arXiv:2411.18195*, 2024.

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.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

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

Dena Mujtaba, Brian Hu, Anthony Hoogs, and Arslan Basharat. Aligning machiavellian agents: Behavior steering via test-time policy shaping. *arXiv preprint arXiv:2511.11551*, 2025.

Anis Najar and Mohamed Chetouani. Reinforcement learning with human advice: a survey. *Frontiers in Robotics and AI*, 8:584075, 2021.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

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.

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.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

John Rawls. *The Theory of Justice*. Havard university press, 1971.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603, 2024.

J. Schulman, S. Levine, P. Abbeel, M.I. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

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.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A Smith, and Simon S Du. Decoding-time language model alignment with multiple objectives. *Advances in Neural Information Processing Systems*, 37:48875–48920, 2024.

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. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

Umer Siddique, Peilang Li, and Yongcan Cao. Fairness in traffic control: Decentralized multi-agent reinforcement learning with generalized gini welfare functions. In *Multi-Agent reinforcement Learning for Transportation Autonomy*, 2024.

Umer Siddique, Peilang Li, and Yongcan Cao. Towards fair and efficient policy learning in cooperative multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 2744–2746, 2025a.

Umer Siddique, Peilang Li, and Yongcan Cao. Learning fair pareto-optimal policies in multi-objective reinforcement learning. *arXiv preprint arXiv:2606.18111*, 2026.

Zara Siddique, Irtaza Khalid, Liam D Turner, and Luis Espinosa-Anke. Shifting perspectives: Steering vectors for robust bias mitigation in llms. *arXiv preprint arXiv:2503.05371*, 2025b.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Viacheslav Sinii, Alexey Gorbatovski, Artem Cherepanov, Boris Shaposhnikov, Nikita Balagansky, and Daniil Gavrilov. Steering llm reasoning through bias-only adaptation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9213–9222, 2025.

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 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

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.

Christian Wirth, Johannes Fürnkranz, and Gerhard Neumann. Model-free preference-based reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *arXiv preprint arXiv:2410.08193*, 2024.

Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with generalized gini welfare functions. In *Adaptive and Learning Agents (ALA) Workshop*, 2023a.

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

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*, 2017.

Xinnan Zhang, Chenliang Li, Siliang Zeng, Jiaxiang Li, Zhongruo Wang, Songtao Lu, Alfredo Garcia, and Mingyi Hong. Reinforcement learning in inference time: A perspective from successive policy iterations. In *Workshop on Reasoning and Planning for Large Language Models*, 2025.

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

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

# Supplementary Materials

*The following content was not necessarily subject to peer review.*

## A Theoretical Analysis of Inference-Time Fairness Alignment

In this appendix, we provide the complete theoretical analysis of our inference-time fairness alignment framework. For clarity and self-containment, we recall the core problem formulation and restate all relevant propositions and theorems from the main text alongside their detailed proofs.

As introduced in Section 4, we consider the problem of steering or shaping a pretrained RL policy toward fair outcomes at inference time without updating its learned parameters. The key observation is that, although the base policy was trained with a scalar reward $r = h(\boldsymbol{r})$, the environment exposes a richer *vector* reward $\boldsymbol{r} \in \mathbb{R}^N$ that can be leveraged post hoc to optimize a welfare-based fairness objective. Formally, we can write this problem statement as:

$$\max_{\pi'} J_{\boldsymbol{\omega}}(\pi') = \max_{\pi'} \phi_{\boldsymbol{\omega}} \left( \boldsymbol{J}(\pi') \right), \tag{10}$$

where the expectation is taken with respect to the shaped policy $\pi'$. This means that our goal is to find a shaped policy $\pi'$ that maximizes $J_{\boldsymbol{\omega}}(\pi')$ subject to the constraint that $\pi'$ is obtained strictly by reshaping the action distribution of $\pi$ at inference time. However, solving this problem raises some unique challenges. Since $\phi_{\boldsymbol{\omega}}$ is non-linear, it cannot be written as a sum of per-step scalar rewards on the original state space. and the fairness depends on the *cumulative* return vector accrued so far, not only on the current environment state. Our approach solves these challenges through a sequence of principled reductions. First, the core difficulty of non-linearity of $\phi_{\boldsymbol{\omega}}$ is that it is a function of the full trajectory return, which makes it non-decomposable into per-step rewards and non-Markovian in the original state space (Section A.1). We address this by constructing an *augmented MDP* $\widetilde{\mathcal{M}}$ that restores the Markov property, then formally connecting the base policy to this augmented setting. Within $\widetilde{\mathcal{M}}$, we derive the *optimal* shaped policy in closed form as the solution to a KL-constrained welfare maximization problem (Section A.3). Finally, we describe how this can be estimated via our welfare critic-based approach.

### A.1 Restoring Markovianity via a Welfare-Augmented MDP

The core difficulty in optimizing $J_{\boldsymbol{\omega}}(\pi')$ is that the welfare function $\phi_{\boldsymbol{\omega}}$ is a non-linear function applied to the full trajectory return. Consequently, the objective cannot be decomposed into independent per-step rewards. This means that the welfare-optimal action at time $t$ depends not only on the current state $s_t$ but also on the history of rewards already accrued. For simplicity, next we assume $\gamma = 1$ and thus $J$ becomes the cumulative sum of rewards, as denoted as $\boldsymbol{R}$.

**Proposition A.1** (Non-Markovianity of Welfare Optimization)**.** *When $\phi_{\boldsymbol{\omega}}$ involves sorting (e.g., the GGF (1)), the welfare-optimal policy is not Markovian in $\mathcal{S}$. There exist states $s$ and distinct reward histories $\boldsymbol{R} \neq \boldsymbol{R}'$ for which the welfare-optimal actions at $(s, \boldsymbol{R})$ and $(s, \boldsymbol{R}')$ are different.*

**Example A.1.** *To illustrate Proposition A.1, we consider an MDP with $N = 2$ objectives and GGF weights $w_1 > w_2 > 0$. As seen in Figure 2, at state $s_{t-1}$, action $a_1$ yields $\boldsymbol{r}_1 = (1, 0)$ and $a_2$ yields $\boldsymbol{r}_2 = (0, 1)$, both leading to a terminal state $s_t$. If the accrued reward in this trajectory is $\boldsymbol{R} = (10, 2)$, then $a_2$ produces a final return of $(10, 3)$, which sorts $(3, 10)$ yielding a welfare score of $3w_1 + 10w_2$. Action $a_1$ produces $(11, 2)$, yielding a welfare score of $2w_1 + 11w_2$. Since $w_1 > w_2$ and the difference is $w_1 - w_2 > 0$, $a_2$ is optimal. Conversely, if $\boldsymbol{R}' = (2, 10)$, the same analysis shows $a_1$ is optimal. Thus, the optimal action at the* same state *$s$ reverses depending on cumulative reward $\boldsymbol{R}$.*

To restore the Markov property, we enrich the state representation with the cumulative reward vector and define a welfare-augmented MDP.

**Definition A.1** (Welfare-Augmented MDP)**.** *Given an MOMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \boldsymbol{r}, T, \gamma)$ and welfare function $\phi_{\boldsymbol{\omega}}$, define the undiscounted finite-horizon MDP $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{S}}, \mathcal{A}, \widetilde{P}, \tilde{r}, T)$ where the state space is augmented with the cumulative reward vector, $\widetilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}^N$, providing an augmented state is $\tilde{s}_t = (s_t, \boldsymbol{R}_t)$. The transition dynamics update the accumulator deterministically: $\widetilde{P}\big(\tilde{s}' \mid \tilde{s}, a)\big) = \widetilde{P}((s', \mathbf{R}') \mid (s, \mathbf{R}), a) = P(s' \mid s, a) \cdot \mathbf{1}[\mathbf{R}' = \mathbf{R} + \mathbf{r}(s, a)]$. The scalar reward in $\widetilde{\mathcal{M}}$ is defined as the marginal welfare contribution:*

$$\tilde{r}\big((s, \boldsymbol{R}),\, a\big) \;=\; \phi_{\boldsymbol{w}}\big(\boldsymbol{R} + \boldsymbol{r}(s, a)\big) - \phi_{\boldsymbol{w}}(\boldsymbol{R}). \tag{11}$$

**Theorem A.1** (Augmented MDP Equivalence)**.** *$\widetilde{\mathcal{M}}$ is Markovian, and for any policy $\tilde{\pi}$ on $\widetilde{\mathcal{S}}$:*

$$\mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{T-1} \tilde{r}(\tilde{s}_t, a_t)\right] \;=\; \mathbb{E}_{\tilde{\pi}}\big[\phi_{\boldsymbol{w}}(\boldsymbol{R}_T)\big] - \phi_{\boldsymbol{w}}(\mathbf{0}). \tag{12}$$

*Hence, maximizing the standard scalar return in $\widetilde{\mathcal{M}}$ is equivalent to maximizing the welfare fairness objective $J_{\boldsymbol{\omega}}(\pi')$.*

*Proof.* Markovianity: $\tilde{r}$ and $\widetilde{P}$ depend on the current augmented state $\tilde{s}_t = (s_t, \boldsymbol{R}_t)$ and action $a_t$ only. Equivalence: $\sum_t \tilde{r}(\tilde{s}_t, a_t) = \sum_t [\phi_{\boldsymbol{w}}(\boldsymbol{R}_{t+1}) - \phi_{\boldsymbol{w}}(\boldsymbol{R}_t)] = \phi_{\boldsymbol{w}}(\boldsymbol{R}_T) - \phi_{\boldsymbol{w}}(\mathbf{0})$ by telescoping. □

The augmented MDP reduces the non-Markovian welfare problem to a standard scalar-reward MDP. We define the welfare value functions in $\widetilde{\mathcal{M}}$:

$$\widetilde{V}^{\tilde{\pi}}(\tilde{s}) = \mathbb{E}_{\tilde{\pi}}\left[\sum_{k=t}^{T-1} \tilde{r}_k \;\middle|\; \tilde{s}_t = \tilde{s}\right], \quad \widetilde{Q}^{\tilde{\pi}}(\tilde{s}, a) = \tilde{r}(\tilde{s}, a) + \mathbb{E}_{s'}[\widetilde{V}^{\tilde{\pi}}(\tilde{s}')], \tag{13}$$

and the welfare advantage $\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a) = \widetilde{Q}^{\tilde{\pi}}(\tilde{s}, a) - \widetilde{V}^{\tilde{\pi}}(\tilde{s})$.

### A.2 Anchoring on the base policy via policy lifting

Since the base policy $\pi$ is trained in the standard MDP with scalar reward, but its shaped policy $\pi'$ must operate in $\widetilde{\mathcal{M}}$, we therefore define its canonical embedding into the augmented state space. Note that both of these MDPs share the same dynamics and action space but differ fundamentally in state representation and objective. To reason about $\pi$ within $\widetilde{\mathcal{M}}$, its canonical embedding is defined as below.

**Definition A.2** (Lifting)**.** *The lifted policy $\tilde{\pi}$ on $\widetilde{\mathcal{S}}$ is $\tilde{\pi}(a \mid \tilde{s}) \;:=\; \pi(a \mid s)$ for all $\tilde{s} = (s, \boldsymbol{R}) \in \widetilde{\mathcal{S}},\ a \in \mathcal{A}$.*

While $\tilde{\pi}$ is generally a suboptimal policy in $\widetilde{\mathcal{M}}$ (because it ignores the $\boldsymbol{R}$-component necessary for welfare optimality), it perfectly preserves the base policy's interaction with the environment.

**Proposition A.2** (Trajectory Equivalence)**.** *Running $\pi$ in original MDP $\mathcal{M}$ and running $\tilde{\pi}$ in augmented MDP $\widetilde{\mathcal{M}}$ induce same distributions over base-state trajectories $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$. That is, $\Pr_M^{\pi}(\tau) \;=\; \Pr_{\widetilde{\mathcal{M}}}^{\tilde{\pi}}(\tau)$, where $\Pr(\tau)$ is the probability of an entire trajectory $\tau$ occurring.*

*Proof.* Both trajectory probabilities factor as $\rho(s_0) \prod_{t=0}^{T-1} \pi(a_t \mid s_t)\, P(s_{t+1} \mid s_t, a_t)$. The lifting gives $\tilde{\pi}(a_t \mid \tilde{s}_t) = \pi(a_t \mid s_t)$, and $\widetilde{P}$ marginalises to $P(s' \mid s, a)$ since the $\boldsymbol{R}$-update is deterministic. □

A direct and powerful consequence of Proposition A.2 is that rollouts collected by the base policy can be *directly reused* for offline training in $\widetilde{\mathcal{M}}$. We simply annotate each transition with the cumulative reward vector $\boldsymbol{R}_t$ (computed post-hoc from the recorded vector rewards) and the marginal welfare reward (11). Furthermore, although $\pi$ was trained to maximize $J(\pi)$, it induces a well-defined

welfare $J_{\boldsymbol{\omega}}(\pi)$ that provides a guaranteed performance baseline. The following result quantifies why a good base policy is a good starting point for welfare improvement.

**Proposition A.3** (Welfare Floor)**.** *If the training aggregation is the utilitarian sum ($h = \mathbf{1}^\top$), then for any policy $\pi$:*

$$J_{\boldsymbol{\omega}}(\pi) \;\geq\; w_N \cdot \boldsymbol{J}(\pi), \tag{14}$$

*where $w_N > 0$ is the smallest GGF weight.*

*Proof.* By definition of the GGF, $\phi_{\boldsymbol{\omega}}(\boldsymbol{u}) = \sum_i w_i u_i^{\uparrow} \geq w_N \sum_i u_i^{\uparrow}$, applying it to $J_{\boldsymbol{\omega}}(\pi) = \phi_{\boldsymbol{\omega}}\left(\boldsymbol{J}(\pi)\right) = \sum_{i=1}^{N} w_i J_i(\pi)^{\uparrow} \geq w_N \sum_{i=1}^{N} J_i(\pi)^{\uparrow} = w_N \boldsymbol{J}(\pi)$ yield the result. □

The bound reveals a separation of concerns, which means a trained base policy contributes *environmental competence* (high total return $J(\pi)$) to a welfare floor. What $\pi$ lacks is equity. By combining trajectory equivalence with Theorem A.1, we can characterize the welfare improvement achievable by any shaped policy $\pi'$:

$$J_{\boldsymbol{\omega}}(\pi') - J_{\boldsymbol{\omega}}(\pi) = \widetilde{J}(\tilde{\pi}') - \widetilde{J}(\tilde{\pi}), \tag{15}$$

where $\widetilde{J}(\cdot)$ denotes the expected return in $\widetilde{\mathcal{M}}$. This identity provides a formal bridge, which is *improving fairness over the base policy is mathematically equivalent to a standard policy improvement problem over the lifted policy in $\widetilde{\mathcal{M}}$.*

### A.3 Optimal KL-Regularized Inference-Time Policy Shaping

We now derive the shaped policy as the optimal solution to a KL-regularized optimization problem in $\widetilde{\mathcal{M}}$. The KL-regularization serves two purposes. First, it keeps the shaped policy $\pi'$ close to the original policy $\pi$, and second, it ensures the surrogate objective remains a valid approximation to the true welfare gain. Solving this optimization problem and using the performance difference lemma (Kakade & Langford, 2002) applied to $\widetilde{\mathcal{M}}$, we can express the welfare gap as $\widetilde{J}(\tilde{\pi}') - \widetilde{J}(\tilde{\pi}) = T\mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}'}}\Big[\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a)\Big]$. Since $d^{\tilde{\pi}'}$ depends on the unknown policy, we replace it with $d^{\tilde{\pi}}$ to obtain a tractable welfare surrogate that shares the same gradient at $\tilde{\pi}' = \tilde{\pi}$ and provides a local lower bound on the true improvement under a KL penalty (Schulman et al., 2015). Therefore, to maximize this objective while preserving the environmental competence guaranteed by the welfare floor, we restrict the shaped policy to a KL-divergence trust region around $\tilde{\pi}$:

$$\max_{\tilde{\pi}'} \quad T\,\mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}}}\Big[\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a)\Big] \tag{16}$$

$$\text{s.t.} \quad \mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}}}\Big[\mathrm{KL}\big(\tilde{\pi}'(\cdot|\tilde{s}) \,\big\|\, \tilde{\pi}(\cdot|\tilde{s})\big)\Big] \leq \varepsilon, \qquad \tilde{\pi}'(\cdot|\tilde{s}) \in \Delta(\mathcal{A}) \;\; \forall\, \tilde{s}. \tag{17}$$

**Theorem A.2** (Optimal Shaped Policy)**.** *The unique solution to (16) is:*

$$\boxed{\pi'(a \mid s, \boldsymbol{R}) \;=\; \frac{\pi(a \mid s)\;\exp\Big(\frac{1}{\beta}\,\widetilde{Q}^{\tilde{\pi}}\big((s, \boldsymbol{R}),\, a\big)\Big)}{\displaystyle\sum_{a'} \pi(a' \mid s)\;\exp\Big(\frac{1}{\beta}\,\widetilde{Q}^{\tilde{\pi}}\big((s, \boldsymbol{R}),\, a'\big)\Big)}} \tag{18}$$

*where $\beta > 0$ is the Lagrange multiplier for the KL constraint, determined by $\varepsilon$. The shaped policy $\pi'$ is computed directly from $\pi$ and the welfare value function $\widetilde{Q}^{\tilde{\pi}}$ with no iterative update.*

*Proof.* The proof proceeds in four parts: (i) form the Lagrangian, (ii) solve pointwise for each $\tilde{s}$, (iii) replace the advantage with the Q-function, and (iv) verify uniqueness and KKT conditions.

**Part 1: Lagrangian.**

We introduce a dual variable $\alpha \geq 0$ for the KL constraint and per-state multipliers $\zeta(\tilde{s})$ for the normalisation constraints $\sum_a \tilde{\pi}'(a|\tilde{s}) = 1$. The non-negativity constraints $\tilde{\pi}'(a|\tilde{s}) \geq 0$ will be satisfied automatically by the form of the solution. The Lagrangian is:

$$\begin{aligned}\mathcal{L}(\tilde{\pi}', \alpha, \zeta) = \mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}}}\left[\sum_a \tilde{\pi}'(a|\tilde{s})\, \widetilde{A}^{\tilde{\pi}}(\tilde{s}, a)\right] \\ - \alpha\left(\mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}}}\left[\sum_a \tilde{\pi}'(a|\tilde{s}) \log \frac{\tilde{\pi}'(a|\tilde{s})}{\tilde{\pi}(a|\tilde{s})}\right] - \varepsilon\right) \\ + \sum_{\tilde{s}} \zeta(\tilde{s})\left(1 - \sum_a \tilde{\pi}'(a|\tilde{s})\right). \end{aligned} \tag{19}$$

Expanding the KL divergence and writing $p(a) = \tilde{\pi}'(a|\tilde{s})$ and $q(a) = \tilde{\pi}(a|\tilde{s}) = \pi(a|s)$ for readability:

$$\mathcal{L} = \sum_{\tilde{s}} d^{\tilde{\pi}}(\tilde{s}) \sum_a p(a)\left[\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a) - \alpha \log \frac{p(a)}{q(a)}\right] + \alpha\varepsilon + \sum_{\tilde{s}} \zeta(\tilde{s})\left(1 - \sum_a p(a)\right). \tag{20}$$

**Part 2: Pointwise optimisation.**

Since the objective and constraints decompose over states (the KL decomposes as $\mathbb{E}_{\tilde{s}}[\mathrm{KL}(\tilde{\pi}'(\cdot|\tilde{s})\|\tilde{\pi}(\cdot|\tilde{s}))]$ and the simplex constraints are per-state), we can solve for $\tilde{\pi}'(\cdot|\tilde{s})$ independently at each $\tilde{s}$.

Fix $\tilde{s}$ and differentiate $\mathcal{L}$ with respect to $p(a) = \tilde{\pi}'(a|\tilde{s})$:

$$\frac{\partial \mathcal{L}}{\partial p(a)} = d^{\tilde{\pi}}(\tilde{s})\left[\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a) - \alpha\left(\log \frac{p(a)}{q(a)} + 1\right)\right] - \zeta(\tilde{s}). \tag{21}$$

Setting this to zero:

$$\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a) - \alpha \log \frac{p(a)}{q(a)} - \alpha = \frac{\zeta(\tilde{s})}{d^{\tilde{\pi}}(\tilde{s})}. \tag{22}$$

Solving for $\log p(a)$:

$$\log p(a) = \log q(a) + \frac{1}{\alpha}\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a) - 1 - \frac{\zeta(\tilde{s})}{\alpha\, d^{\tilde{\pi}}(\tilde{s})}. \tag{23}$$

Denoting $\beta = \alpha$ and collecting the terms that do not depend on $a$ into a single constant:

$$c(\tilde{s}) \;:=\; -1 - \frac{\zeta(\tilde{s})}{\beta\, d^{\tilde{\pi}}(\tilde{s})}, \tag{24}$$

we get:

$$\log \tilde{\pi}'(a|\tilde{s}) = \log \pi(a|s) + \frac{1}{\beta}\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a) + c(\tilde{s}). \tag{25}$$

Exponentiating:

$$\tilde{\pi}'(a|\tilde{s}) = \pi(a|s) \cdot \exp\left(\frac{1}{\beta}\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a)\right) \cdot \exp\big(c(\tilde{s})\big). \tag{26}$$

**Part 3: Normalisation and replacing advantage with Q-function.**

The constant $c(\tilde{s})$ is determined by the normalisation constraint $\sum_a \tilde{\pi}'(a|\tilde{s}) = 1$. Summing (26) over $a$:

$$1 = \exp\big(c(\tilde{s})\big) \sum_a \pi(a|s)\, \exp\left(\frac{1}{\beta}\widetilde{A}^{\tilde{\pi}}(\tilde{s}, a)\right), \tag{27}$$

so:

$$\exp\big(c(\tilde{s})\big) = \frac{1}{\sum_a \pi(a|s)\,\exp\Big(\frac{1}{\beta}\widetilde{A}^{\tilde{\pi}}(\tilde{s},a)\Big)}. \tag{28}$$

Substituting back into (26):

$$\tilde{\pi}'(a|\tilde{s}) = \frac{\pi(a|s)\,\exp\Big(\frac{1}{\beta}\widetilde{A}^{\tilde{\pi}}(\tilde{s},a)\Big)}{\sum_{a'} \pi(a'|s)\,\exp\Big(\frac{1}{\beta}\widetilde{A}^{\tilde{\pi}}(\tilde{s},a')\Big)}. \tag{29}$$

Now we replace the advantage $\widetilde{A}^{\tilde{\pi}}$ with the Q-function $\widetilde{Q}^{\tilde{\pi}}$. Since $\widetilde{A}^{\tilde{\pi}}(\tilde{s},a) = \widetilde{Q}^{\tilde{\pi}}(\tilde{s},a) - \widetilde{V}^{\tilde{\pi}}(\tilde{s})$, substituting into (29):

$$\begin{aligned}
\tilde{\pi}'(a|\tilde{s}) &= \frac{\pi(a|s)\,\exp\Big(\frac{1}{\beta}\big[\widetilde{Q}^{\tilde{\pi}}(\tilde{s},a) - \widetilde{V}^{\tilde{\pi}}(\tilde{s})\big]\Big)}{\sum_{a'} \pi(a'|s)\,\exp\Big(\frac{1}{\beta}\big[\widetilde{Q}^{\tilde{\pi}}(\tilde{s},a') - \widetilde{V}^{\tilde{\pi}}(\tilde{s})\big]\Big)} \\
&= \frac{\pi(a|s)\,\exp\Big(\frac{1}{\beta}\widetilde{Q}^{\tilde{\pi}}(\tilde{s},a)\Big)\cdot\exp\Big(-\frac{1}{\beta}\widetilde{V}^{\tilde{\pi}}(\tilde{s})\Big)}{\sum_{a'} \pi(a'|s)\,\exp\Big(\frac{1}{\beta}\widetilde{Q}^{\tilde{\pi}}(\tilde{s},a')\Big)\cdot\exp\Big(-\frac{1}{\beta}\widetilde{V}^{\tilde{\pi}}(\tilde{s})\Big)} \\
&= \frac{\pi(a|s)\,\exp\Big(\frac{1}{\beta}\widetilde{Q}^{\tilde{\pi}}(\tilde{s},a)\Big)}{\sum_{a'} \pi(a'|s)\,\exp\Big(\frac{1}{\beta}\widetilde{Q}^{\tilde{\pi}}(\tilde{s},a')\Big)},
\end{aligned} \tag{30}$$

where the last step uses the fact that $\exp\big(-\frac{1}{\beta}\widetilde{V}^{\tilde{\pi}}(\tilde{s})\big)$ is independent of $a$ and cancels between numerator and denominator. Writing $\tilde{s} = (s, \mathbf{R})$ gives exactly 18.

**Part 4: Uniqueness and KKT verification.**

*Strict concavity.* The objective is linear in $\tilde{\pi}'$, and the KL divergence $\mathrm{KL}(\tilde{\pi}'\|\tilde{\pi})$ is strictly convex in $\tilde{\pi}'$ (since $x\log x$ is strictly convex). Therefore, the feasible set $\{\tilde{\pi}' : \mathrm{KL} \le \varepsilon,\ \tilde{\pi}' \in \Delta(\mathcal{A})\}$ is a convex set, and the Lagrangian objective $\mathcal{L}$ is strictly concave in $\tilde{\pi}'$ for any $\alpha > 0$. This guarantees the solution is unique.

*Non-negativity.* Since $\pi(a|s) > 0$ for $a \in \mathrm{supp}(\pi)$ and $\exp(\cdot) > 0$, the solution (30) satisfies $\tilde{\pi}'(a|\tilde{s}) > 0$ for all $a \in \mathrm{supp}(\pi)$ and $\tilde{\pi}'(a|\tilde{s}) = 0$ for $a \notin \mathrm{supp}(\pi)$. The non-negativity constraints are automatically satisfied.

*KKT conditions.* The stationarity condition is satisfied by construction (Part 2). The primal feasibility ($\sum_a \tilde{\pi}' = 1, \mathrm{KL} \le \varepsilon$) is ensured by normalisation and by choosing $\alpha$ to satisfy the KL constraint. Dual feasibility ($\alpha \ge 0$) holds by assumption. Complementary slackness ($\alpha(\mathbb{E}[\mathrm{KL}] - \varepsilon) = 0$) determines $\alpha$: either $\alpha = 0$ (the unconstrained optimum already satisfies $\mathrm{KL} \le \varepsilon$) or the KL constraint binds (equality holds) and $\alpha > 0$. In practice, $\alpha > 0$ and the constraint is active.

*Determining $\beta$.* The parameter $\beta = \alpha > 0$ is implicitly defined by the constraint:

$$\mathbb{E}_{\tilde{s}\sim d^{\tilde{\pi}}}\Big[\mathrm{KL}\big(\tilde{\pi}'_\beta(\cdot|\tilde{s}) \,\big\|\, \tilde{\pi}(\cdot|\tilde{s})\big)\Big] = \varepsilon, \tag{31}$$

where $\tilde{\pi}'_\beta$ is the solution (30) parameterised by $\beta$. The left-hand side is a continuous, strictly monotonically decreasing function of $\beta$: as $\beta \to \infty$, $\tilde{\pi}' \to \tilde{\pi}$ and $\mathrm{KL} \to 0$; as $\beta \to 0^+$, $\tilde{\pi}'$ concentrates on the welfare-maximising action and $\mathrm{KL} \to \max$. By the intermediate value theorem, a unique $\beta > 0$ satisfying (31) exists for any $\varepsilon \in (0, \mathrm{KL}_{\max})$. □

The shaped policy (18) has three structural properties that make it well-suited for inference-time alignment:

(i) **Support preservation.** Since $\exp(\cdot) > 0$, the shaped policy assigns positive probability to exactly those actions in the support of $\pi$: no actions are introduced or eliminated.

(ii) **Smooth interpolation.** The parameter $\lambda := 1/\beta$ controls alignment strength. As $\lambda \to 0$, we recover the base policy $\pi$; as $\lambda \to \infty$, $\pi'$ concentrates on the welfare-maximising action within the support of $\pi$. The intermediate regime provides a continuum between performance and fairness.

(iii) **Log-space decomposition.** The log-policy decomposes additively:

$$\log \pi'(a \mid s, \boldsymbol{R}) \;=\; \underbrace{\log \pi(a \mid s)}_{\text{base competence}} \;+\; \underbrace{\lambda\, \widetilde{Q}^{\tilde{\pi}}\big((s, \boldsymbol{R}), a\big)}_{\text{welfare correction}} \;-\; \log Z(s, \boldsymbol{R}), \tag{32}$$

where $Z$ is the normalisation constant. The welfare Q-function acts as a *logit correction* that tilts the base distribution toward equitable actions.

# B Algorithms

**Algorithm 1** QFAIR: OFFLINE WELFARE CRITIC TRAINING

**Require:** Base policy $\pi$, environment $\mathcal{M}$, welfare $\phi_{\boldsymbol{w}}$, exploration $\beta_{\text{explore}}$, episodes $N$, epochs $E$, learning rate $\eta$
**Ensure:** Trained welfare critic $Q_\phi$

Initialise $Q_\phi$ with random parameters $\phi$
$\mathcal{D} \leftarrow \emptyset$ ▷ Replay buffer for augmented transitions
**for** episode $= 1, \ldots, N$ **do**
  $s_0 \leftarrow \mathcal{M}.\text{reset}(); \quad \boldsymbol{R}_0 \leftarrow \mathbf{0}$
  **for** $t = 0, 1, \ldots, T-1$ **do**
    $a_t \sim \mu(\cdot \mid s_t)$ ▷ Behaviour policy
    $s_{t+1}, \boldsymbol{r}_t, d_t \leftarrow \mathcal{M}.\text{step}(a_t)$ ▷ $\boldsymbol{r}_t \in \boldsymbol{r}^N$
    $\boldsymbol{R}_{t+1} \leftarrow \boldsymbol{R}_t + \gamma^t\, \boldsymbol{r}_t$
    $\tilde{r}_t \leftarrow \phi_{\boldsymbol{w}}(\boldsymbol{R}_{t+1}) - \phi_{\boldsymbol{w}}(\boldsymbol{R}_t)$ ▷ Marginal welfare (11)
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{((s_t, \boldsymbol{R}_t),\ a_t,\ \tilde{r}_t,\ (s_{t+1}, \boldsymbol{R}_{t+1}),\ d_t)\}$
  **end for**
**end for**
**for** epoch $= 1, \ldots, E$ **do**
  **for** mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    $y_i \leftarrow \tilde{r}_i + \gamma(1 - d_i) \sum_{a'} \mu(a' | s'_i)\, Q_\phi(\tilde{s}'_i, a')$ for each $i \in \mathcal{B}$ ▷ Expected SARSA (8)
    $\phi \leftarrow \phi - \eta\, \nabla_\phi \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (Q_\phi(\tilde{s}_i, a_i) - y_i)^2$
  **end for**
**end for**

**Algorithm 2** INFERENCE-TIME WELFARE-ALIGNED ACTION SELECTION

**Require:** Frozen base policy $\pi$, frozen critic $Q_\phi$, alignment strength $\lambda$

$s_0 \leftarrow \mathcal{M}.\text{reset}(); \quad \boldsymbol{R}_0 \leftarrow \mathbf{0}$
**for** $t = 0, 1, \ldots$ **until** episode ends **do**
  $f_{\text{base}}(a) \leftarrow$ base policy signal for all $a \in \mathcal{A}$
  $f_{\text{fair}}(a) \leftarrow \bar{Q}_\phi\big((s_t, \boldsymbol{R}_t),\ a\big)$ for all $a \in \mathcal{A}$ ▷ Normalised QFair scores
  $\pi'(\cdot) \leftarrow \text{softmax}\big(f_{\text{base}} + \lambda\, f_{\text{fair}}\big)$ ▷ Shaped policy (9)
  Select $a_t \leftarrow \text{argmax}_a\, \pi'(a)$ or sample $a_t \sim \pi'$ ▷ Greedy or stochastic
  $s_{t+1}, \boldsymbol{r}_t, d_t \leftarrow \mathcal{M}.\text{step}(a_t)$
  $\boldsymbol{R}_{t+1} \leftarrow \boldsymbol{R}_t + \gamma^t\, \boldsymbol{r}_t$ ▷ Update accumulator
**end for**

# C Additional Expriments

In this section, we present the additional experimental results that were not included in the main paper. As mentioned in Section 6, we present the complete environmental details and empirical results for the four-room domain here.

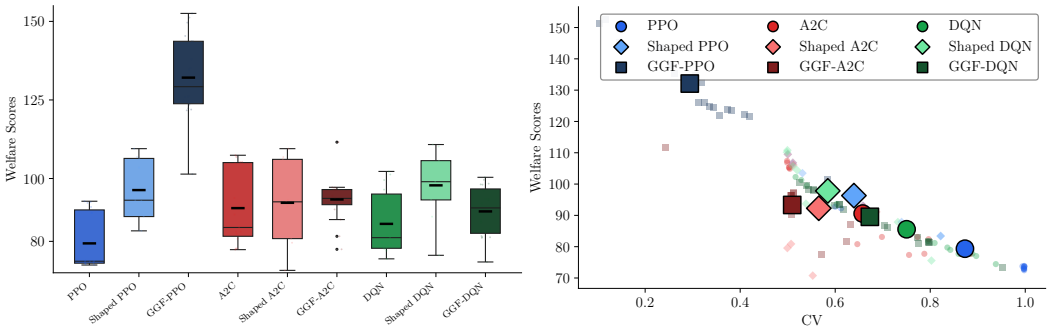
## C.1 Four Room

We first provide the details of the four-room environment, which models a multi-objective navigation and resource-collection problem under spatial constraints and stochastic outcomes. The environment is a $13 \times 13$ maze divided by walls into four connected regions, with multiple spawn locations and three resource types distributed across the map. These resources are represented by distinct shapes and colors: a blue square, a red circle, and a green triangle. In this domain, the state space contains agent's current grid position together with local binary indicators that signal whether the current cell yields a red, green, or blue resource. The action space includes the four cardinal movements (up, down, left, right), where the movement is constrained by walls and map boundaries, so policy quality depends not only on objective prioritization but also on efficient navigation through the maze. The reward is a vector consisting of $N = 3$, which corresponds to the collection of each resource type. The resource dynamics are intentionally heterogeneous: green and blue provide stable unit rewards upon collection, whereas red yields stochastic rewards (success or failure), introducing uncertainty and making balanced collection more challenging. In this environment, fairness is interpreted as avoiding collapse to a single easily obtainable objective and instead maintaining balanced performance across all three reward dimensions. During training, agents optimize the scalar sum of objective rewards. At inference time, we decompose this scalar signal into its three components and apply policy shaping toward the GGF objective as described in Section 5.

Table 2 shows the total reward, CV, minimum reward, and maximum reward during the evaluation phase. The results show that although our inference-time shaped methods achieve total rewards comparable to the original RL algorithms, they consistently produce lower CV values than their corresponding base policies, indicating more balanced objective outcomes. As expected, the GGF-based methods, such as GGF-PPO and GGF-A2C, achieve even lower CV values, since they are explicitly optimized for fairness during training. In contrast, our approach performs fairness alignment purely at inference time without retraining the policy. Interestingly, our shaped PPO achieves the highest total reward among all methods, suggesting that optimizing directly for the GGF objective can sometimes lead to overly conservative solutions with reduced overall reward. On the other hand, our inference-time shaping approach maintains high task performance while improving balance across objectives.

Table 2: Performance comparison across algorithms. We report total reward, coefficient of variation (CV), minimum objective return, and maximum objective return.

| Algorithm | Total | CV | Min reward | Max reward |
|---|---|---|---|---|
| PPO | $289.97 \pm 4.65$ | $0.8731 \pm 0.1829$ | $0.03 \pm 0.04$ | $262.56 \pm 43.70$ |
| Shaped PPO | $296.49 \pm 0.54$ | $0.6394 \pm 0.1331$ | $0.22 \pm 0.21$ | $208.13 \pm 42.04$ |
| GGF-PPO | $280.14 \pm 6.21$ | $0.2939 \pm 0.1203$ | $57.05 \pm 14.43$ | $145.96 \pm 23.71$ |
| A2C | $281.22 \pm 10.41$ | $0.6578 \pm 0.1342$ | $0.18 \pm 0.14$ | $200.50 \pm 41.14$ |
| Shaped A2C | $269.15 \pm 40.69$ | $0.5654 \pm 0.0795$ | $0.33 \pm 0.15$ | $169.89 \pm 44.81$ |
| GGF-A2C | $257.21 \pm 12.47$ | $0.5095 \pm 0.0860$ | $2.79 \pm 10.74$ | $146.70 \pm 21.19$ |
| DQN | $286.42 \pm 1.58$ | $0.7503 \pm 0.1527$ | $0.31 \pm 0.10$ | $231.21 \pm 38.21$ |
| Shaped DQN | $287.43 \pm 11.93$ | $0.5844 \pm 0.1094$ | $0.25 \pm 0.10$ | $184.06 \pm 32.33$ |
| GGF-DQN | $285.77 \pm 1.93$ | $0.6730 \pm 0.1180$ | $0.24 \pm 0.17$ | $213.89 \pm 29.08$ |




(a) Welfare scores.

(b) Pareto front plot.

Figure 6: Performance comparison of PPO, A2C, DQN, their inference-time shaped variants, and GGF-trained counterparts in the Four Room.

To further evaluate fairness outcomes, we compute welfare scores using the GGF function applied to the empirical average return vectors. As shown in Figure 6a, the inference-time shaped variants consistently achieve higher welfare scores than their standard counterparts. In some cases, such as DQN, the shaped variant even attains higher welfare scores than the GGF-trained model. This result highlights the effectiveness of inference-time alignment for improving fairness without modifying the training objective. We further validate this by showing our results in the Pareto analysis shown in Figure 6b. These results demonstrate that our inference-time shaping methods shift policies toward the desirable region of lower inequality and higher welfare. While GGF-PPO achieves the strongest fairness performance overall, it is explicitly trained to optimize the GGF objective. On the other hand, our proposed methods consistently improve fairness relative to the original policies.

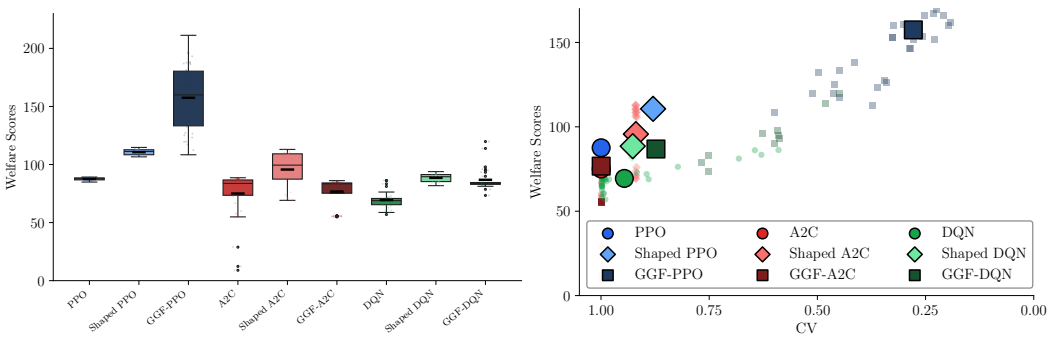
### C.2 Harvest Regrow

We also present welfare scores and Pareto front plot of our methods against standard deep RL methods and their GGF counterparts in the harvest regrow environment. Figure 7a shows welfare values computed from 100 evaluation trajectories per trained agent. Once again, the results show that inference-time shaped variants consistently achieve higher welfare than their standard counterparts and, in some cases, are comparable to GGF-trained policies. Interestingly, Shaped-A2C achieves higher welfare scores than GGF-A2C, even though GGF-A2C is trained to explicitly optimize the GGF objective. As expected, GGF-PPO and GGF-DQN achieve higher welfare scores than their standard and inference-time counterparts. However, our proposed inference-time methods achieve comparable welfare scores to GGF methods while maintaining lower inequality. This is further validated by the Pareto front plot (see Figure 7b). These results suggest that post-training alignment can effectively recover fairness without requiring retraining under a specific fairness objective.

## D Hyperparameters

In this section, we detail the hyperparameters and training configurations used across all experiments. For the base RL agents, we utilized the implementations provided by the Stable Baselines3 library [1]. To ensure robust performance and a fair baseline comparison, we tuned the hyperparameters for these agents using the Optuna optimization framework. For the multi-objective baselines that optimize the welfare function directly during training (i.e., GGF-PPO, GGF-A2C, and GGF-DQN), we adopted the author-provided hyperparameters detailed in (Siddique et al., 2020), which originally proposed these algorithms. For our inference-time shaped agents (Shaped PPO, Shaped A2C, Shaped DQN), the parameters of the base policies remain completely frozen. The hyperparameters listed for the Shaped agents therefore correspond to the offline training of the QFair critic

[1] https://github.com/DLR-RM/stable-baselines3




(a) Welfare scores.

(b) Pareto front plot.

Figure 7: Performance comparison of PPO, A2C, DQN, their inference-time shaped variants, and GGF-trained counterparts in the Harvest Regrow.

and the inference-time temperature scaling ($\lambda$). The tables below summarize the hyperparameter values. To denote environment-specific tuning, we use the following subscripts: $sc$ for the species conservation environment, $fr$ for the four-room environment, and $hr$ for the harvest regrow crops environment.

Table 3: Hyperparameters for PPO and Shaped PPO

| **Hyperparameter** | **PPO (Base)** | **Shaped PPO** |
|---|---|---|
| Discount factor ($\gamma$) | $0.99_{sc,fr}, 0.95_{hr}$ | Base Frozen |
| Learning rate | $3 \times 10^{-4}_{sc,fr,hr}$ | $1 \times 10^{-3}_{sc,fr,hr}$ (QFair) |
| Optimizer | $\text{Adam}_{sc,fr,hr}$ | $\text{Adam}_{sc,fr,hr}$ (QFair) |
| Network size | $[128, 128]_{sc,fr,hr}$ (Actor/Critic) | $[256, 256, 256]_{sc,fr,hr}$ (QFair) |
| Batch size | $64_{sc,fr}, 128_{hr}$ | $256_{sc,fr,hr}$ (QFair) |
| n_steps | $2048_{sc,fr,hr}$ | N/A |
| Clip range | $0.2_{sc,fr,hr}$ | N/A |
| Entropy coefficient | $0.0_{sc,fr}, 0.01_{hr}$ | N/A |
| Value function coefficient | $0.5_{sc,fr,hr}$ | N/A |
| Training timesteps | $1 \times 10^{6}_{sc,fr,hr}$ | $5 \times 10^{5}_{sc,fr,hr}$ (QFair) |

Table 4: Hyperparameters for A2C and Shaped A2C

| **Hyperparameter** | **A2C (Base)** | **Shaped A2C** |
|---|---|---|
| Discount factor ($\gamma$) | $0.99_{sc,fr}, 0.95_{hr}$ | Base Frozen |
| Learning rate | $7 \times 10^{-4}_{sc,fr,hr}$ | $1 \times 10^{-3}_{sc,fr,hr}$ (QFair) |
| Optimizer | $\text{RMSprop}_{sc,fr,hr}$ | $\text{Adam}_{sc,fr,hr}$ (QFair) |
| Network size | $[128, 128]_{sc,fr,hr}$ (Actor/Critic) | $[256, 256, 256]_{sc,fr,hr}$ (QFair) |
| n_steps | $5_{sc,fr,hr}$ | N/A |
| Entropy coefficient | $0.0_{sc}, 0.01_{fr,hr}$ | N/A |
| Value function coefficient | $0.5_{sc,fr,hr}$ | N/A |
| Training timesteps | $1 \times 10^{6}_{sc,fr,hr}$ | $5 \times 10^{5}_{sc,fr,hr}$ (QFair) |

preserving strong task performance.

Table 5: Hyperparameters for DQN and Shaped DQN

| **Hyperparameter** | **DQN (Base)** | **Shaped DQN** |
|---|---|---|
| Discount factor ($\gamma$) | $0.99_{sc,fr}, 0.95_{hr}$ | Base Frozen |
| Learning rate | $1 \times 10^{-4}_{sc,fr,hr}$ | $1 \times 10^{-3}_{sc,fr,hr}$ (QFair) |
| Optimizer | $\text{Adam}_{sc,fr,hr}$ | $\text{Adam}_{sc,fr,hr}$ (QFair) |
| Network size | $[64, 64]_{sc,fr,hr}$ (Q-Net) | $[256, 256, 256]_{sc,fr,hr}$ (QFair) |
| Buffer size | $1 \times 10^{5}_{sc,fr,hr}$ | $1 \times 10^{5}_{sc,fr,hr}$ (QFair) |
| Exploration fraction | $0.1_{sc,fr,hr}$ | $\beta_{\text{explore}} = 0.1_{sc,fr,hr}$ |
| Double DQN | $\text{True}_{sc,fr,hr}$ | N/A |
| Dueling DQN | $\text{True}_{sc,fr,hr}$ | N/A |
| Training timesteps | $1 \times 10^{6}_{sc,fr,hr}$ | $5 \times 10^{5}_{sc,fr,hr}$ (QFair) |