

# Beyond Scalar Welfare: Enforcing Identity-Aware Equity in Multi-Agent Reinforcement Learning

Nhat-Hoang P. Nguyen  
Air Traffic Management Research  
Institute  
Nanyang Technological University  
Singapore, Singapore  
nguyenph001@e.ntu.edu.sg

Duc-Thinh Pham  
Center for AI Research, VinUniversity  
Hanoi, Vietnam  
Air Traffic Management Research  
Institute  
Nanyang Technological University  
Singapore, Singapore  
thinh.pd@vinuni.edu.vn

Trung-Hoang Le  
Faculty of Information Technology,  
University of Science  
Ho Chi Minh City, Vietnam  
Vietnam National University  
Ho Chi Minh City, Vietnam  
lthoang@hcmus.edu.vn

Sameer Alam  
Air Traffic Management Research  
Institute  
Nanyang Technological University  
Singapore, Singapore  
sameeralam@ntu.edu.sg

Vu N. Duong  
Center for AI Research, VinUniversity  
Hanoi, Vietnam  
vu.dn@vinuni.edu.vn

## ABSTRACT

Equity in cooperative multi-agent reinforcement learning (MARL) is imposed via a single scalar, such as an inequality-averse social welfare function or a dispersion index (e.g., Gini, coefficient of variation, or Jain). These proxies conflate two dimensions of inequity, including who and how much, offering no guarantee of identity-level equity. We introduce an identity-aware notion of equity for otherwise heterogeneous agents, requiring each agent’s outcome to remain within a relative tolerance band around the mean. From this premise, we propose the Minimum Reallocation of Excess (MRE), separating incidence (who fall outside the band) from displacement (how much outside-band mass they carry), measuring the least reassignment needed to restore equity. We prove convexity, piecewise linearity, Lipschitz continuity, and Pigou-Dalton compliance, providing a stable, aligned metric for optimizing equity-improving transfers. We further design the Equity-Lagrangian Dual (ELD) update, enforcing an expected-MRE constraint while maximizing return. The update is backbone-agnostic and preserves decentralized execution. Formal case analyses show that dispersion and welfare proxies can misrank allocations or fail to enforce identity-level equity, whereas ELD attains near-equal outcomes at comparable efficiency. Illustrative experiments in two cooperative games with state-of-the-art MARL frameworks show that ELD reduces MRE relative to proxy-based constraints while maintaining aggregate returns, proving that identity-aware equity can be enforced without architectural changes. This framework replaces proxy objectives with a measurable equity target and a plug-in primal-dual rule, enabling identity-level equity during MARL training.

## KEYWORDS

Multi-agent Reinforcement Learning; Fairness; Equity; Cooperative

### ACM Reference Format:

Nhat-Hoang P. Nguyen, Duc-Thinh Pham, Trung-Hoang Le, Sameer Alam, and Vu N. Duong. 2026. Beyond Scalar Welfare: Enforcing Identity-Aware Equity in Multi-Agent Reinforcement Learning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 9 pages. <https://doi.org/10.65109/LWVB6211>

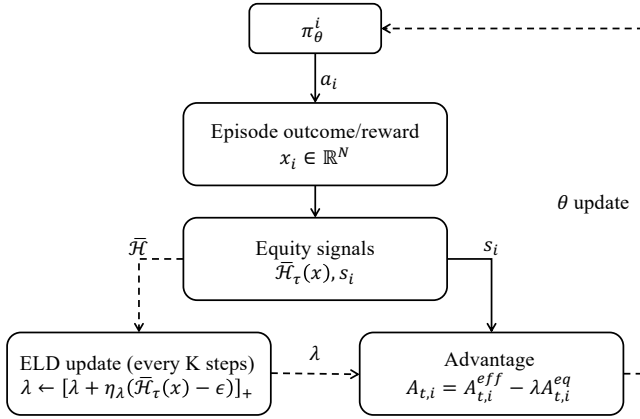
## 1 INTRODUCTION

Multi-agent reinforcement learning (MARL) enables distributed decision-making where learners coordinate under shared, tightly coupled resources, such as wireless scheduling [24], traffic signal control [38], data-center and network congestion management [15], and multi-robot teams [27, 29]. In such settings, the practice of maximizing a single aggregate return often induces *winner-take-more* dynamics, where early advantages compound, a few agents capture outsized utility, and the remainder shoulder persistent burden-outcomes that may be efficient in aggregate yet brittle, difficult to justify, and misaligned with stakeholder expectations [20].

A principled response draws on distributive justice traditions that treat impartiality (equal treatment of equals), equity (transfers that reduce unjust disparities), and efficiency (respect for Pareto improvements) as coequal aims [1]. Practically, equity has often been measured via scalar dispersion or balance indices (e.g., coefficient of variation (CV), Jain’s index [14], and Gini [12]), even as fairness is enforced by optimizing inequality-averse social-welfare functions (SWFs), e.g., generalized Gini,  $\alpha$ -fairness, Nash/leximin with gradient-based reinforcement learning [9, 41]. These tools are powerful yet imperfect, as scalar summaries compress identity and temporal structure. They can misreport distinct treatment profiles as equally “fair”. At the same time, SWFs blend equity and efficiency, so improved welfare need not imply that indistinguishable agents are treated equally throughout learning and execution.



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



**Figure 1: Equity-Lagrangian Dual (ELD) update using identity-aware equity (MRE) signals; shown per agent  $i$**

Recent advancements in MARL promise several benefits, including greater feasibility, reduced dispersion, and a better balance between equity and efficiency. However, challenges remain, including low efficiency and a lack of clarity regarding who deviates from desired equity [16, 42]. Complementary theories and algorithms indicate that both lexicographic and min-max fairness are attainable through learning [11, 28, 36]. Additionally, Proximal Policy Optimization (PPO) updates can be directly used to optimize welfare objectives [17, 32]. These developments do not fully address guarantees at the identity level.

This work addresses equity in cooperative MARL through two main contributions. First, we propose the Minimum Reallocation of Excess (MRE), an identity-aware equity metric, which treats near-equality as a specification rather than a proxy. Operationally, equity is not only about how dispersed outcomes look, but about how much outcome mass must be shifted from over-served identities to under-served identities to restore near-equality. We therefore interpret our equity signal as the per-agent average units that must be returned to disadvantaged identities to bring all agents back within a tolerance band. This makes equity violations actionable, as it reports the repair cost in the same units as the outcomes being allocated, rather than in a unitless dispersion score.

Second, we introduce the Equity-Lagrangian Dual (ELD) update, a plug-in constrained policy-gradient scheme that enforces an expected-MRE budget while maximizing return. ELD is backbone-agnostic and requires no architectural changes, with a periodic two-timescale dual ascent tracks feasibility, an identity-aware episode reducer supplies per-agent signals, and an advantage construction with action-independent baselines provides an unbiased estimator of the Equity-Lagrangian gradient. Figure 1 provides an at-a-glance overview of the proposed identity-aware equity metric and its ELD enforcement mechanism. With a case analysis (section 3.3) and experiments on cooperative environments (section 5), ELD consistently shifts the efficiency-equity frontier, reducing MRE (i.e., less reallocation needed to achieve equity) at comparable aggregate returns, demonstrating that identity-level equity can be measured and enforced within standard MARL training loops.

## 2 LITERATURE REVIEW

A growing strand of MARL casts equity as optimization of a social welfare function (SWF) that aggregates individual utilities while embedding three classical axioms of distributive justice—impartiality (equal treatment of equals), equity (Pigou-Dalton improvements), and efficiency (Pareto monotonicity) [1, 22]. By replacing utilitarian sums with inequality-averse welfare (e.g., generalized Gini,  $\alpha$ -fairness, Nash social welfare, leximin), one makes the equity-efficiency trade-off explicit and tunable, with optimization amenable to policy-gradient machinery [4, 21, 23, 26, 33]. In machine learning, this SWF perspective complements constraint/penalty approaches by providing axiomatic guarantees and a single scalar objective for end-to-end learning in sequential settings [13, 37].

Dispersion metrics such as CV and Jain’s index [14] summarize distributional spread or balance, whereas welfare indices (e.g., generalized Gini, Nash social welfare) treat fairness as an objective to be optimized [23, 33]. Despite their popularity, these scalars can misreport equity as they compress identity-level structure into a single number. Under Adler’s and Aristotle’s “equal treatment of equals”, impartiality requires that heterogeneous agents not be favored [1]. Yet CV/Jain/Gini may assign the same score to distributions with different incidence (who is above/below the mean) and concentration (how excess is clustered), obscuring unequal treatment patterns that matter in practice. In short, these measures are excellent for diagnostics but incomplete for identity-aware equity, and they do not, by themselves, enforce equity; they must be paired with learning rules or constraints that operationalize the target.

Within MARL, early identity-aware equity was posed as an optimality criterion. [40] introduced a regularized max-min formulation for equitable policies in multi-agent MDPs. The first influential deep MARL solution, FEN [16], used a hierarchical controller with consensus to interleave self-utility and a global fairness signal, showing promising reductions in dispersion on benchmark tasks, but fairness was operationalized via CV-style surrogates, which may conflate low variance with low efficiency and cannot guarantee identity-level equalization. These limitations prompted a pivot toward axiomatic SWFs as primary objectives. This shift clarified trade-offs but did not ensure identity-level near-equality.

Two consolidated lines followed. **Formal SWF-RL**: [35] treated fair policy learning as maximizing a strictly concave, symmetric SWF (e.g., generalized Gini) over long-run returns and provided discounted-to-average reward justification, together with algorithmic adaptations for deep RL [7, 35]. **Decentralized fairness architectures**: SOTO [42] separated self-oriented and team-oriented heads per agent and optimized differentiable SWFs (generalized Gini,  $\alpha$ -fairness, leximin-style limits). SOTO consistently improved the equity-efficiency frontier on canonical toy environments (Matthew Effect, SUMO-like traffic) under decentralized execution. However, because SOTO still optimizes a single scalar welfare, fairness fidelity ultimately depends on the chosen scalarization; near-equality among indistinguishable agents is not guaranteed under non-stationarity or partial observability.

Theory has increasingly treated fairness as first-class in unknown environments. [17] optimized lexicographic/min-max fairness without Bellman decomposability, proving sublinear regret and guarantees via optimistic online optimization and policy-gradient

**Table 1: Notations**

Symbol	Description
$N$	number of elements/agents
$\tau$	tolerance band
$[\cdot]_+$	maximum of $\cdot$ and 0
$e_i$	outside-band excess for agent $i$
$E$	total excess
$\mathbb{1}(\cdot)$	indicator function
$u$	uniform baseline $(\frac{1}{N}, \dots, \frac{1}{N})$
$\mathcal{H}$	Minimum Reallocation of Excess
$I$	Identity matrix
$\mathbf{1}$	all-ones vector
$\varepsilon$	equity budget
$\lambda$	dual variable
$\psi$	constraint violation signal
$\eta_\lambda$	updating rate of $\lambda$

variants, thereby formalizing when fair-optimal policies are learnable with finite data. At the algorithmic level, fairness has been fused into mainstream deep RL updates, which directly optimizes generalized-Gini welfare with independent learners, providing higher minimum-agent returns and lower inequality than FEN/SOTO baselines on grid world and Matthew-effect tasks [34]. These works reinforce that fairness can be embedded in scalable MARL without catastrophic efficiency loss, especially when combined with PPO and careful credit assignment [31, 32].

Despite advancements in MARL, most methods still rely on a single metric that blends equity and efficiency, failing to ensure near-equality among indistinguishable agents. The field needs a clear equity-focused metric to identify which agents exceed the mean and a learning rule to enforce near-equality within a specified tolerance, while remaining compatible with decentralized MARL and maintaining high aggregate welfare.

### 3 EQUITY IN COOPERATIVE MARL

We study cooperative multi-agent reinforcement learning (MARL) in domains where heterogeneous agents must coordinate to achieve high overall return and near-equal outcomes. Prior MARL work often encodes “fairness” by optimizing a single scalar, such as an inequality-averse social-welfare function (SWF) or dispersion index—thereby blending efficiency with a proxy for equity and potentially obscuring who is treated unequally during learning and execution. This motivates an equity-enforce formulation that keeps identities explicit and treats near-equality as a specification to be enforced, not just a preference to be traded off.

We use equity to mean identity-level near-equality of outcomes among agents with equal entitlement. We reserve fairness for broader normative objectives (e.g., welfare optimization) used in prior MARL work. Table 1 summarizes the main notations.

#### 3.1 Cooperative MARL

We consider cooperative resource-allocation games in MARL where a team of otherwise indistinguishable agents jointly “collects” limited items/jobs/rewards over an episode. Decisions are decentralized at execution but can be coordinated during training. Each episode

outputs an outcome vector  $x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N$ , where  $x_i$  is agent  $i$ ’s realized share/utility/income (e.g., tasks completed, items secured, or per-agent return). Efficiency is the total team return, defined as

$$J_{eff} = \sum_{i=1}^N x_i \quad (1)$$

Beyond efficiency, we require equity, where agents with equal entitlement should achieve near-equal outcomes. To formalize “near-equal”, we introduce a tolerance band  $\tau \in [0, \infty)$  around the mean outcome  $\bar{x}$ , where  $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ . Outcomes within the range of  $[(\tau - 1)\bar{x}, (\tau + 1)\bar{x}]$  are treated as acceptably equal; only outside-band deviations signal inequity. Let  $e_{i,\tau}$  denote agent  $i$ ’s outside-band excess (zero if inside the band), specifically

$$e_{i,\tau}(x) = [|x_i - \bar{x}| - \tau\bar{x}]_+ \quad (2)$$

This identity-level signal lets us ask (i) who violates equity (incidence) and (ii) by how much (displacement), rather than relying on anonymous dispersion summaries. Two aspects of identity are explicit in this construction. First, excess is recorded per index, with no sorting or anonymization, thereby preserving who is implicated. Second, equity is defined at the identity level, meaning equal sharing means  $e_i = \bar{e} = \frac{1}{N} \sum_i e_i$  for every  $i$ .

#### 3.2 Identity-aware Equity Metrics

The tolerance band turns equity into a question of outside-band violations, where in-band differences are acceptable and ignored. This produces hinge-like behavior, hence convex and piecewise linear form. MRE then asks a concrete repair question, what is the smallest amount of outcome mass that must move from over-served to under-served identities to make violations uniformly shared.

**3.2.1 Incidence (who).** The incident rate

$$\text{IncRate}_\tau(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(e_{i,\tau} > 0) \quad (3)$$

where  $\mathbb{1}(\cdot)$  is the indicator function. Equation (3) reports the fraction of identities with nonzero excess. Conditional on at least one identity being flagged, we form the selection profile

$$q_i^{inc} = \frac{\mathbb{1}(e_i > 0)}{\sum_{j=1}^N \mathbb{1}(e_{j,\tau} > 0)} \quad (4)$$

a probability vector over indices that records which identities are implicated. Its deviation from the uniform baseline  $u = (\frac{1}{N}, \dots, \frac{1}{N})$  is measured by total variation (TV):

$$\mathcal{H}_\tau^{inc}(x) = \frac{1}{2} \sum_i |q_i^{inc} - u_i| \in [0, 1 - u] \quad (5)$$

This distinguishes episodes that flag the same number of identities but not the same people.

**3.2.2 Displacement (how).** To assess how excess mass is distributed across identities, we normalize by its total when  $E = \sum_i e_i > 0$  and define the magnitude profile as  $q_i^{mag} = \frac{e_i}{E}$ . Concentration relative to  $u$  is again captured by TV:

$$\mathcal{H}_\tau^{mag}(x) = \frac{1}{2} \sum_i |q_i^{mag} - u_i| \in [0, 1 - u] \quad (6)$$

3.2.3 *Minimum Reallocation of Excess.* The Minimum Reallocation of Excess (MRE) is defined as  $\mathcal{H}_\tau(x) = \frac{1}{2} \|e_\tau(x) - Eu\|_1$ , and the normalized form is  $\overline{\mathcal{H}}_\tau(x) = \frac{1}{N} \mathcal{H}_\tau(x)$ .

Intuitively,  $\mathcal{H}$  is the least amount of outcome mass that must be reassigned among identities to convert the observed excess  $e$  into the uniform target  $Eu$ . The value is zero if and only if excess is already uniform (or  $E = 0$ ). The normalized index reads as a concentration coefficient between perfectly uniform and maximally concentrated (all excess on a single identity).

Crucially,  $\mathcal{H}$  preserves identity awareness through the components it aggregates, as different episodes can return the same scalar yet differ in who is affected, which is visible in  $q^{inc}$  and  $q^{mag}$ . Thus, the suite  $(\text{IncRate}, q^{inc}, E, q^{mag}, \mathcal{H})$  jointly answers who is affected, by how much, how concentrated the effects are, and how much correction is required.

### 3.3 Case analysis: Different dispersions, same equity values

We use a small synthetic example to illustrate a specific failure mode of standard inequality indices in identity-aware equity, as they summarize overall dispersion, but they do not reveal whether inequity is concentrated on particular identities outside the tolerance band. The goal here is to isolate the pattern that drives equity interventions.

Alongside CV (Equation 7a), Gini (Equation 7b), and Jain (Equation 7c), we further include Theil's T (Equation 7d) and Atkinson's index ( $\epsilon = \frac{1}{2}$ ) (Equation 7e). These indices are anonymous summaries, as CV and Jain are fully determined by the first two moments under fixed  $N$ , and Gini, Theil, and Atkinson aggregate symmetric functions of pairwise gaps or normalized ratios. As a result, they respond to dispersion but do not distinguish whether disparities are (i) mostly inside the tolerance band versus outside it, or (ii) distributed broadly versus concentrated on a few identities, which is the operational pattern that drives equity interventions.

$$CV = \frac{\sigma}{\bar{x}}, \quad (7a)$$

$$G = \frac{\sum_i^N \sum_j^N |x_i - x_j|}{2N^2 \bar{x}}, \quad (7b)$$

$$J = \frac{\left(\sum_i^N x_i\right)^2}{N \sum_i^N x_i^2}, \quad (7c)$$

$$T = \frac{1}{N} \sum_i^N \frac{x_i}{\bar{x}} \log\left(\frac{x_i}{\bar{x}}\right), \quad (7d)$$

$$A_{\epsilon=\frac{1}{2}} = 1 - \left[ \frac{1}{N} \sum_i^N \left(\frac{x_i}{\bar{x}}\right)^{1-\epsilon} \right]^{\frac{1}{1-\epsilon}} = 1 - \frac{\left(\sum_i^N \sqrt{x_i}\right)^2}{N^2 \bar{x}} \quad (7e)$$

With allocations shown in Figure 2, the classical indices cluster tightly because the constructions match the global spread. However, these allocations differ sharply in *outside-band structure*. Some cases place many identities near the mean with only mild, scattered band violations, while others concentrate the same overall dispersion into a small set of identities that fall far outside the band. From an equity standpoint, these are not equivalent episodes, since concentrated

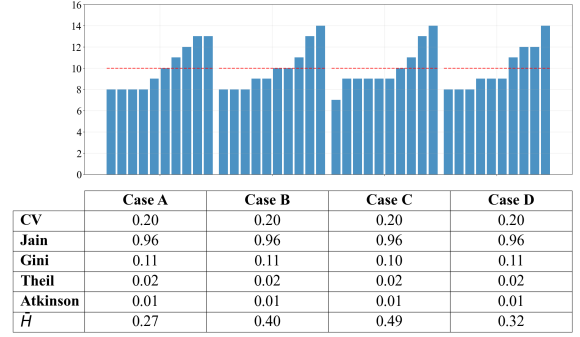


Figure 2: Illustrative examples with 4 cases with different distributions and their corresponding equity scores

violations imply repeated disadvantage for a few identities and call for larger corrective transfers.

MRE separates these patterns by design. It first discards in-band variation and keeps only outside-band excess, then measures the minimal outcome mass that must be transferred so that outside-band burden is shared uniformly. As a result, cases that look similar under dispersion indices can provide different MRE values, because MRE tracks the repair cost of restoring tolerance-band equity, and is sensitive to who bears the violations.

This case treats deviation as non-directional (large vs. small, without “higher is better”). In applications, direction often matters; splitting  $MRE$  into above-mean and below-mean components (or using a signed variant) can improve interpretability and align the metric with domain semantics.

### 3.4 Theoretical properties

MRE is an  $L_1$  distance to a uniform target after a piecewise-linear excess mapping, which explains convexity and piecewise linearity. This also connects directly to Pigou–Dalton style equalizing transfers, where any mean-preserving transfer from a more-burdened identity toward a less-burdened one moves the excess vector closer to uniform sharing and reduces the repair cost.

We collect the key properties that render  $\mathcal{H}$  both behaviorally aligned with equity and amenable to optimization.

PROPOSITION 1 (MINIMAL REASSIGNMENT).  $\mathcal{H}(x) = \frac{1}{2} \sum_{i=1}^N |e - \bar{e}|$ .

Among nonnegative vectors with equal sums, the minimal mass that must be moved to transform one into the other equals half their  $L_1$  distance. Instantiating the target as the uniform vector gives the stated form.

PROPOSITION 2 (CONVEXITY AND PIECEWISE LINEARITY). Let  $e \in \mathbb{R}^N$  and  $\bar{e} = \frac{1}{N} \sum_{j=1}^N e_j$ . Define  $\mathcal{H}(e) = \frac{1}{2} \sum_{i=1}^N |e_i - \bar{e}|$ . Then  $\mathcal{H}$  is convex and piecewise linear. Moreover, a subgradient  $g \in \partial \mathcal{H}(e)$  can be written component-wise as

$$g_i = \frac{1}{2} (\sigma_i - \bar{\sigma}), \quad \bar{\sigma} = \frac{1}{N} \sum_{j=1}^N \sigma_j,$$

where  $\sigma_i \in \partial |e_i - \bar{e}|$ , i.e.,  $\sigma_i = +1$  if  $e_i > \bar{e}$ ,  $\sigma_i = -1$  if  $e_i < \bar{e}$ , and  $\sigma_i \in [-1, 1]$  if  $e_i = \bar{e}$ .

Convexity in  $e$  supports stable constrained optimization of expected equity (e.g., primal–dual updates) under stochastic learning dynamics.

**PROPOSITION 3 (SCHUR-CONVEXITY / PIGOU-DALTON).** *If  $x$  majorizes  $x'$  and  $\sum_i x_i = \sum_i x'_i$ , then  $\mathcal{H}(x) \geq \mathcal{H}(x')$ ; any mean-preserving Pigou-Dalton transfer strictly reduces  $\mathcal{H}$  unless  $e$  is already uniform. Then, the metric aligns with equity, as it rewards spreading excess more evenly across identities.*

## 4 EQUITY-LAGRANGIAN DUAL (ELD) UPDATE

We maximize efficiency while enforcing identity-aware equity. Equity is measured by the Minimum Reallocation of Excess (MRE), which compares outside-band deviations to a uniform allocation across identities. A normalized form lets it pair cleanly with constraints and dual updates.

The dual multiplier is embedded directly in the objective, rather than treating equity as an unweighted second goal, so gradients track the active equity budget. This yields a backbone-agnostic primal–dual update that remains compatible with decentralized execution.

### 4.1 Learning objective

Let there be  $N$  agents and an episodic horizon  $T$ . Each episode produces per-step task rewards  $r_t^{\text{eff}}$  and a per-agent share vector  $x \in \mathbb{R}^N$ . Fix tolerance  $\tau \geq 0$  and a equity budget  $\varepsilon \geq 0$ .

We maximize task return subject to an equity constraint

$$\max_{\theta} J_{\text{eff}}(\theta) \quad \text{s.t.} \quad J_{\text{eq}}(\theta) = \mathbb{E}_{\theta}[\mathcal{H}(x)] \leq \varepsilon \quad (8)$$

In other words, fairness is treated as a specification (feasibility) rather than a soft preference mixed into a single scalar objective.

Equation (8) is a standard constrained MDP (CMDP) formulation [2] solved via Lagrangian relaxation. We convert the inequality constraint into an unconstrained saddle-point objective; the dual acts as a shadow price of unfairness, tuning the trade-off. Then, the associated Lagrangian is

$$\mathcal{L}(\theta, \lambda) = J_{\text{eff}}(\theta) - \lambda (J_{\text{eq}}(\theta) - \varepsilon), \quad \lambda \geq 0 \quad (9)$$

Optimizing  $\mathcal{L}$  by ascending in  $\theta$  (policy improvement) and ascending in  $\lambda$  (constraint enforcement) implements a standard primal-dual procedure that targets a Karush-Kuhn-Tucker (KKT) condition, i.e., a stationary policy–multiplier pair satisfying primal feasibility (expected MRE within the budget), dual feasibility ( $\lambda \geq 0$ )– for the constrained problem [19]. At a stationary point  $(\theta^*, \lambda^*)$  we satisfy KKT:  $J_{\text{eq}}(\theta^*) \leq \varepsilon$ ,  $\lambda^* \geq 0$ , and  $\lambda^* (J_{\text{eq}}(\theta^*) - \varepsilon) = 0$ .

**Interpretation.** The constraint enforces *identity-aware equity* (equal sharing of the excess beyond a tolerance window). Unlike scalarized objectives, Equation (9) provides a *specification-level guarantee* on inequity.

**LEMMA 4.1 (SCALE-SHAPE FACTORIZATION).** *If  $E > 0$  and  $p_i = \frac{e_i}{E}$ ,  $u_i = \frac{1}{N}$ , then  $\mathcal{H}(x) = \frac{1}{2}E(x) \|p(x) - u\|_1$ .*

**PROOF.**  $\sum_i |e_i - \frac{E}{N}| = E \sum_i |\frac{e_i}{E} - \frac{1}{N}| = E \|p - u\|_1$ .  $\square$

**PROPOSITION 4 (CONVEXITY  $\dot{\varphi}$  PIGOU-DALTON).**  *$e_i = [x_i - \bar{x} - \tau]_+$  is convex in  $x$ ; hence  $E$  is convex. As an  $L_1$  norm of an affine map of  $e$ ,  $\mathcal{H}$  is convex and piecewise linear. For fixed  $E$ , any mean-preserving*

*transfer of excess from a more- to a less-shared agent decreases  $\|p - u\|_1$  and thus decreases  $\mathcal{H}$ .*

### 4.2 Solution method

The ELD update integrates a Lagrangian dual into standard on-policy MARL training to enforce an expected normalized equity budget while preserving the backbone’s optimization steps. The method relies on an episode-level equity reducer and a periodic dual ascent, and does not prescribe a particular actor-critic architecture.

We use two forms of the equity signal depending on the backbone. The episode-level scalar  $\overline{\mathcal{H}}_{\tau}(x)$  is used for the dual update and as a team-level cost in CTDE-style learners. The identity-level signal  $s_i$  is used for per-agent shaping or per-agent cost returns in fully decentralized learners. For on-policy training, signals are computed from the current rollout; for replay-based training, signals are computed at collection time and stored with the episode for consistent off-policy updates.

**4.2.1 Periodic Dual Ascent.** At the end of each episode, the reducer forms the identity-aware excess vector  $e$  from the outcome vector  $x$  using the relative band  $\tau$  and computes the MRE  $\mathcal{H}_{\tau}(x)$ , and its normalized counterpart  $\overline{\mathcal{H}}_{\tau}(x) = \frac{1}{N}\mathcal{H}_{\tau}(x)$ . The *constraint violation* signal is  $\psi = \overline{\mathcal{H}}_{\tau}(x) - \varepsilon$ . Using the violation signal, we update  $\lambda$  on a slower time-scale to stabilize feasibility tracking [6, 18]. Every  $K$  episodes, take a projected subgradient step [3, 25] for inequality constraints

$$\lambda \leftarrow [\lambda + \eta_{\lambda}\psi]_+ \quad (10)$$

with a step size  $\eta_{\lambda}$  chosen on a slower time scale than the actor update (classical two-timescale stochastic approximation), which promotes stability near feasibility and reduces oscillations [5].

**4.2.2 Identity Equity-aware signals.** To expose who is outside the band, define the incidence profile  $q^{\text{inc}}$  and the per-identity signal

$$s_i = e_i q_i^{\text{inc}} = \frac{e_i \mathbb{1}(e_i > 0)}{\sum_j \mathbb{1}(e_j > 0)} \quad (11)$$

which evenly spreads the episode’s equity pressure across implicated identities while remaining scale-insensitive to  $E$ . We use  $s_i$  in two places: (i) a terminal shaping term (diagnostics/curriculum) that subtracts  $\lambda s_i$  from agent  $i$ ’s episode score, and (ii) an optional centered control variate inside the advantage (variance reduction) that does not bias the gradient. Baseline/control-variate usage follows standard policy-gradient invariance [39]. The shaped episode reward for reporting is  $R_i = \sum_{t=0}^{T-1} \gamma^t r_{t,i}^{\text{eff}} - \lambda s_i$ .

**4.2.3 Unbiased policy-gradient estimator.** Let  $G_{t,i}^{\text{eff}}$  be the per-time return tied to efficiency, and define two equity returns:

$$G_{t,i}^{\text{eff}} = \sum_{k=t}^{T-1} \gamma^{k-t} r_{k,i}^{\text{eff}}, \quad (12a)$$

$$G_t^{\text{eq}} = \gamma^{T-1-t} \overline{\mathcal{H}}_{\tau}(x), \quad (12b)$$

$$G_{t,i}^{\text{id}} = \gamma^{T-1-t} s_i \quad (12c)$$

---

**Algorithm 1** Equity-Lagrangian Dual (ELD): periodic dual ascent with equity shaping

---

**Require:** tolerance  $\tau$ , budget  $\varepsilon$ , discount  $\gamma$ , stepsizes  $(\eta_\theta, \eta_\lambda)$ , update period  $K$

- 1: Initialize policies  $\pi_{\theta_i}$  ( $i = 1 \dots N$ ), dual  $\lambda \leftarrow 0$ , baselines  $b_i^{\text{eff}}, b_i^{\text{eq}}$
  - 2: **for**  $n = 1, 2, \dots$  **do**
  - 3:   **Rollouts (backbone):** collect on-policy episodes with the current trainer
  - 4:   **Equity reducer:** for each episode compute  $\bar{x}, e, E, \mathcal{H}_\tau(x)$  and set  $\bar{\mathcal{H}}(x) = \frac{1}{N} \mathcal{H}_\tau(x)$
  - 5:   **Dual update (every  $K$ ):** if  $n \bmod K = 0$  then  $\lambda \leftarrow [\lambda + \eta_\lambda (\bar{\mathcal{H}}(x) - \varepsilon)]_+$ ; broadcast  $\lambda$
  - 6:   **Identity signals:**  $c \leftarrow \sum_i \mathbb{1}(e_i > 0)$ ;  $q_i^{\text{inc}} \leftarrow \mathbb{1}(e_i > 0) / \max\{c, 1\}$ ;  $s_i \leftarrow e_i q_i^{\text{inc}}$ ; broadcast  $s_i$
  - 7:   **Returns/advantages:** for each  $(i, t)$  compute  $G_{t,i}^{\text{eff}}, G_{t,i}^{\text{eq}} = \gamma^{T-1-t} s_i$  and  $A_{t,i} = (G_{t,i}^{\text{eff}} - b_i^{\text{eff}}) - \lambda (G_{t,i}^{\text{eq}} - b_i^{\text{eq}}) - \lambda \alpha (s_i - \bar{s}_t)$
  - 8:   **Policy step (backbone):** replace the backbone’s advantage with  $A_{t,i}$  and perform its policy update
  - 9: **end for**
- 

With action-independent baselines  $b_i^{\text{eff}}, b_i^{\text{eq}}$  (running/batch means or GAE baselines), the combined advantage uses  $G_{t,i}^{\text{id}}$  for FD (per-agent credit assignment) and can use  $G_t^{\text{eq}}$  for CTDE as team-level cost, namely

$$A_{t,i} = (G_{t,i}^{\text{eff}} - b_i^{\text{eff}}) - \lambda (G_{t,i}^{\text{id}} - b_i^{\text{eq}}) - \lambda \alpha (s_i - \bar{s}_t) \quad (13)$$

and for CTDE backbones that consume a team-level cost, replace  $G_{t,i}^{\text{id}}$  by  $G_t^{\text{eq}}$  (with the same baseline invariance). The last term is optional,  $\alpha \in [0, 1]$  is a fixed coefficient, and  $\bar{s}_t$  is an action-independent mean (e.g., batch mean at time  $t$ ).

**4.2.4 Training method.** The training procedure proceeds by alternating policy improvement with a periodic dual ascent driven by an episode-level reducer. After each rollout, the reducer computes once per episode—the constraint-excess vector, total excess, and  $\bar{\mathcal{H}}$  (Section 3.2), which together define the violation signal  $\psi$ . The dual update (Equation 10) is executed every  $K$  episodes (Algorithm 1, line 5), keeping  $\lambda$  on a slower timescale than the actor. This two-timescale scheme damps oscillations near feasibility and enables stable tracking of the equity budget.

Policy improvement then uses an Equity-Lagrangian learning signal. The equity return is an episode scalar proportional to  $\bar{\mathcal{H}}$ ; with action-independent baselines (line 7), the resulting combined advantage is an unbiased estimator of  $\nabla_\theta \mathcal{L}(\theta, \lambda)$ . This preserves gradient correctness, reduces variance, and steers updates toward identities responsible for the observed inequity without changing the objective tied to  $\bar{\mathcal{H}}$ .

During the “Policy step” (line 8), we substitute the backbone’s advantage with the combined advantage and apply the usual update (e.g., PPO/TRPO surrogate ascent; actor-critic such as COMA/WQMIX). No architectural changes are required beyond exposing the episode reducer and the dual variable. When the equity budget is slack,  $\lambda \approx 0$  and the method collapses to the base trainer; when sustained excess is detected,  $\lambda$  increases, attenuating out-of-band advantages until the normalized MRE returns to the target.

## 5 EXPERIMENTS

In this section, we evaluate our proposed ELD against baselines using identity-aware equity metrics. Overall, ELD consistently shifts the efficiency–equity frontier outward, achieving near-equality within a specified tolerance while maintaining high aggregate performance and stable learning.

### 5.1 Setup

**Environments.** We evaluate two cooperative grid-world scenarios that benchmark equity–efficiency trade-offs. In the Matthew Effect [16, 42], 10 Pac-Man-like agents collect respawning resources; the more an agent collects, the larger/faster it becomes, inducing cumulative advantage. We use the canonical instantiation with 10 agents and 3 moving “ghost” resources; size/speed scale with prior collection, and resources respawn after capture. Efficiency is measured as the total items collected per episode (higher is better); equity is our identity-aware Minimum Reallocation of Excess (MRE; lower is better), and we also report CV for comparability. In Job Scheduling [16, 42], 4 agents share a single permanent server in a  $5 \times 5$  grid and must coordinate access so that utilization remains high while per-agent completions are equitable; we measure server utilization (higher is better), MRE (lower is better), CV, and the min/max per-agent utility.

**Baselines.** The proposed ELD is compared to Independent PPO [8], COMA (centralized critic with counterfactual advantage) [10], WQMIX (weighted value factorization) [30], FEN (hierarchical fair-efficient learners) [16], and SOTO (two-head SWF-optimizing architecture; generalized Gini and  $\alpha$ -fairness variants) [42].

**ELD Integration.** In addition, each baseline is upgraded into a principled, constraint-driven equity learner by embedding the ELD architecture–periodic dual ascent, identity-aware signals, and unbiased equity advantages–within baselines COMA, WQMIX, FEN, and SOTO. This yields ELD<sub>COMA</sub>, ELD<sub>WQMIX</sub>, ELD<sub>FEN</sub>, and ELD<sub>SOTO</sub>, which enforce an expected-MRE budget and shift the efficiency–equity frontier.

**Configurations.** The evaluation considers both fully decentralized (FD) and centralized training with decentralized execution (CTDE) paradigms. Results are reported as mean  $\pm$  standard deviation over 100 independent runs; we equalize environment steps and configurations across methods. We tuned ELD tolerance/budget on a validation split and reused PPO backbones as their baselines.

### 5.2 Results and Analysis

**Matthew effect – Fully decentralized** (Figure 3). Across FD baselines, ELD improves identity-level equity without sacrificing efficiency because the penalty activates only on outside-band violations and attributes the burden to the implicated identities. This provides a consistent leftward shift of the efficiency–equity frontier, where normalized MRE  $\bar{H}$  decreases while episode return is maintained or improved, and the minimum agent outcome rises, indicating

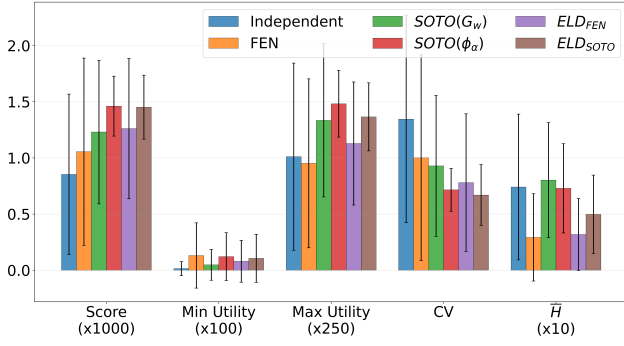


Figure 3: Algorithm comparison on Matthew Effect (FD)

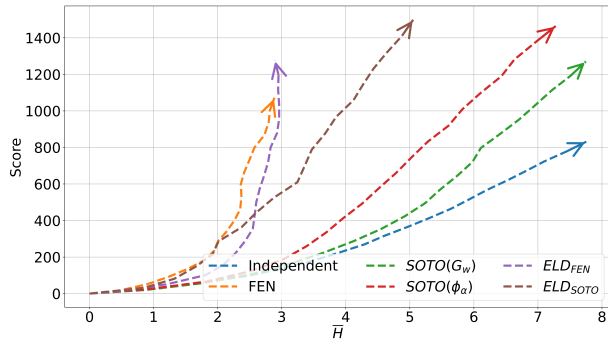


Figure 4: Solution trajectories in Matthew Effect (FD)

fewer identities remain persistently underserved.  $ELD_{SOTO}$  benefits most, suggesting that optimizing a single welfare scalar can still concentrate violations on a few identities even when dispersion metrics look acceptable. In contrast,  $ELD$  forces corrective pressure to scale with outside-band mass, which directly targets the operational reallocation needed to restore tolerance-band equity.  $ELD_{FEN}$  primarily reduces dispersion (e.g., CV) with smaller MRE gains, consistent with  $FEN$  already compressing spread while leaving residual outside-band structure that is not fully identity-equalized.

The training trajectories in Figure 4 explain the mechanism, as when  $\bar{H}$  rises,  $\lambda$  increases, slowing the growth of inequity while reward continues to accumulate, producing a more stable learning path with fewer runaway-advantage episodes.

**Matthew effect – CTDE** (Figure 5). Under CTDE, centralized critics improve credit assignment;  $ELD$  then steers gradients away from policies that repeatedly place the same identities outside the tolerance band. Across  $COMA$ ,  $WQMIX$ ,  $FEN$ , and  $SOTO$ ,  $ELD$  variants achieve comparable scores with materially lower MRE and higher worst-case outcomes, indicating reduced identity-level monopolization. A key observation is that dispersion indices can move little even as MRE improves, because  $ELD$  targets only outside-band mass rather than pushing all outcomes toward exact equality. This separation is important in cooperative MARL, since in-band variation can be efficiency-preserving, while outside-band violations signal the inequity that stakeholders would operationally correct. Overall, the results support that an interpretable equity constraint

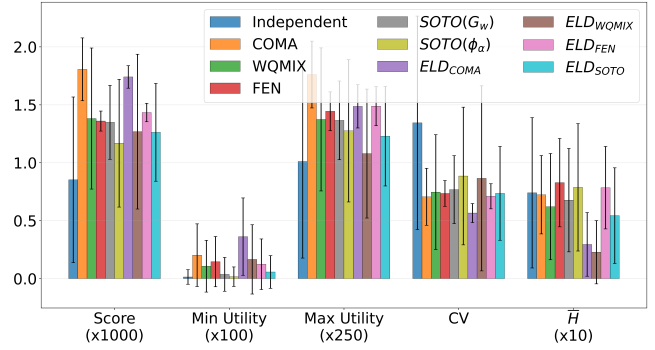


Figure 5: Algorithm comparison on Matthew Effect (CTDE)

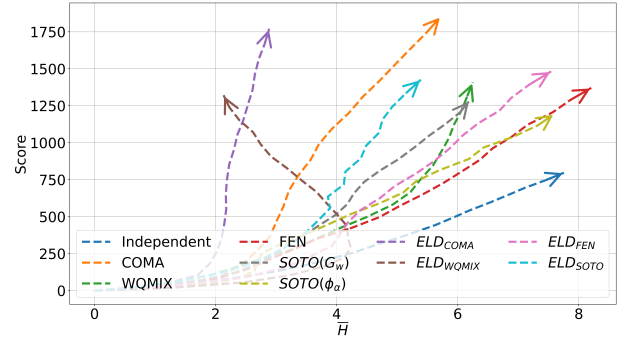


Figure 6: Solution trajectories in Matthew Effect (CTDE)

Table 2: Comparison on Job Scheduling (CTDE)

	Utilization	Min utility	Max utility	CV	$\bar{H}$	
Baseline	Random	$0.16 \pm 0.02$	$0.03 \pm 0.01$	$0.05 \pm 0.01$	$0.26 \pm 0.11$	$0.06 \pm 0.13$
	Independent	$0.99 \pm 0.02$	$0.00 \pm 0.00$	$0.90 \pm 0.11$	$1.56 \pm 0.20$	$20.17 \pm 4.32$
	COMA	$0.98 \pm 0.03$	$0.00 \pm 0.01$	$0.83 \pm 0.17$	$1.43 \pm 0.33$	$17.58 \pm 6.59$
	WQMIX	$0.35 \pm 0.04$	$0.06 \pm 0.01$	$0.11 \pm 0.02$	$0.22 \pm 0.09$	$0.05 \pm 0.22$
	FEN	$0.93 \pm 0.02$	$0.03 \pm 0.04$	$0.62 \pm 0.15$	$1.04 \pm 0.32$	$10.90 \pm 5.50$
	$SOTO(G_w)$	$0.99 \pm 0.02$	$0.00 \pm 0.00$	$0.98 \pm 0.04$	$1.71 \pm 0.05$	$22.87 \pm 0.99$
	$SOTO(\phi_\alpha)$	$0.98 \pm 0.10$	$0.00 \pm 0.00$	$0.93 \pm 0.13$	$1.61 \pm 0.23$	$21.19 \pm 4.03$
Proposed	$ELD_{COMA}$	$0.95 \pm 0.02$	$0.03 \pm 0.04$	$0.61 \pm 0.17$	$0.99 \pm 0.35$	$10.46 \pm 6.10$
	$ELD_{WQMIX}$	$0.78 \pm 0.02$	$0.03 \pm 0.02$	$0.51 \pm 0.09$	$0.98 \pm 0.22$	$9.08 \pm 3.06$
	$ELD_{FEN}$	$0.94 \pm 0.02$	$0.05 \pm 0.05$	$0.52 \pm 0.15$	$0.81 \pm 0.32$	$7.43 \pm 5.22$
	$ELD_{SOTO}$	$0.99 \pm 0.00$	$0.00 \pm 0.02$	$0.85 \pm 0.15$	$1.44 \pm 0.29$	$17.90 \pm 6.05$

can be enforced via a plug-in primal-dual update, without modifying the backbone architecture.

Figure 6 shows the same effect dynamically, where once violations emerge,  $ELD$  maintains score growth while reducing the rate of increase in  $\bar{H}$ , consistent with a dual price that becomes active only under constraint pressure.

**Job Scheduling – CTDE** (Table 2). The single-server setting highlights why dispersion-focused fairness can be misleading, as low inequality can be achieved trivially by low throughput, while high utilization can conceal severe monopolization where some identities are repeatedly starved. This is reflected in baselines that attain strong utilization yet keep the minimum utility near zero, implying persistent exclusion.  $ELD$  improves the utilization–equity trade-off by penalizing only outside-band episodes that indicate

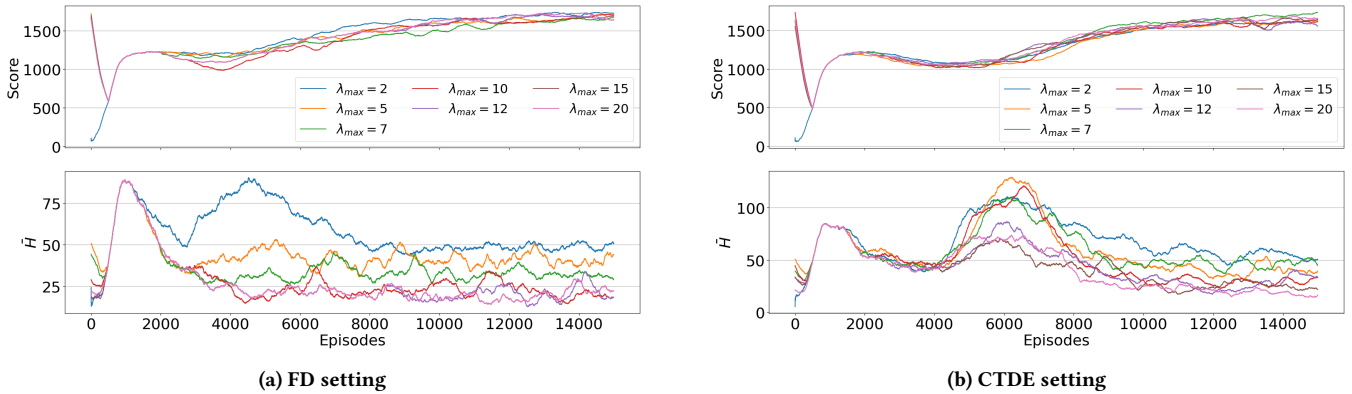


Figure 7: Performance of  $\text{ELD}_{\text{SOTO}}$  when varying  $\lambda_{\max}$

sustained disadvantage, thus discouraging monopolization while preserving productive scheduling. For COMA and FEN, ELD keeps utilization competitive while increasing the floor and reducing MRE, meaning fewer completion units must be reassigned to restore tolerance-band equity. For WQMIX, ELD moves learning into a more productive regime where utilization rises and inequity becomes measurable, but the repair cost remains lower than high-utilization baselines. Taken together, ELD reduces identity-level monopolization under contention while maintaining efficiency, and MRE provides the operational lens that dispersion indices cannot.

Overall, ELD improves the efficiency-equity frontier via a plug-in primal-dual rule that penalizes only outside-tolerance excess and focuses correction on implicated identities, preserving efficiency gradients when the equity budget is slack. This design yields stable decentralized learning and tunable equity guarantees without sacrificing aggregate performance across diverse cooperative MARL architectures.

**Ablation.** In Figure 7, we test whether ELD depends on saturating the multiplier by training  $\text{ELD}_{\text{SOTO}}$  with identical seeds while capping the dual variable at  $\lambda_{\max} \in \{2, 5, 7, 10, 12, 15, 20\}$ . The efficiency curves are largely insensitive to the cap once  $\lambda_{\max}$  is moderate, as final scores converge to similar plateaus, indicating that allowing stronger constraint pressure does not inherently suppress return. Equity, however, exhibits a clear cap-dependent regime. With small caps, the dual variable cannot respond sufficiently when outside-band mass accumulates, so MRE remains elevated and the minimum outcome stays low. As the cap increases to mid-range values, MRE drops substantially and the floor improves, showing that the learned policy reduces persistent violations rather than merely compressing dispersion. Beyond about  $\lambda_{\max} \approx 12$  to 15, gains saturate and changes become marginal, suggesting the optimal multiplier  $\lambda^*$  is interior for most training and the bound mainly prevents rare runaway episodes. CTDE runs show a transient early MRE spike, but higher caps accelerate recovery and return lower steady-state inequity, consistent with stronger corrective pressure when the centralized critic first discovers high-return but inequitable behaviors. Overall, the ablation indicates that ELD does not require extreme multipliers to work; a practical default is  $\lambda_{\max} \in \{12, 15\}$ , which is large enough to enforce the constraint when needed while avoiding unnecessary sensitivity to the cap.

## 6 CONCLUSION

In this study, we introduced an identity-aware notion of equity for cooperative MARL called Minimum Reallocation of Excess (MRE). It quantifies how much redistribution is needed to bring every agent within a tolerance band around the mean outcome, separating those who violate equity from how much. From this, we proposed an Equity-Lagrangian Dual (ELD) update that plugs into state-of-the-art frameworks as a lightweight, primal-dual shaping of advantages. The update activates only when out-of-band violations occur and vanishes otherwise, preserving efficiency-seeking gradients. The integration requires no architectural overhaul and is compatible with fully decentralized and CTDE protocols.

Across evaluations, ELD consistently improved the efficiency-equity frontier, as in Matthew (FD/CTDE) and Job Scheduling (CTDE), ELD integrated variants reduced MRE by about 20% to 60% and lowered dispersion (CV), while maintaining or modestly improving return and raising the worst-case agent outcome. Our work has two main implications. Conceptually, we move from scalar welfare measures to identity-aware equalization. Methodologically, we introduce a drop-in mechanism that improves the fairness of existing multi-agent reinforcement learning (MARL) algorithms without sacrificing performance. Future research will focus on personalizing equity for diverse agents and ensuring convergence under partial observability and communication constraints.

## ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore, and the Civil Aviation Authority of Singapore, under the Aviation Transformation Programme. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore and the Civil Aviation Authority of Singapore.

This work is carried out at the Air Traffic Management Research Institute (ATMRI), Nanyang Technological University (NTU), Singapore. The first author, Nhat-Hoang P. Nguyen, would like to acknowledge the Vingroup Science and Technology Scholarship Program for Overseas Study for Master’s and Doctorate Degrees managed by VinUniversity for the graduate research scholarship.

## REFERENCES

- [1] Matthew Adler. 2012. *Well-being and fair distribution: beyond cost-benefit analysis*. OUP USA.
- [2] Eitan Altman. 1999. *Constrained Markov Decision Processes*. Chapman & Hall/CRC.
- [3] Dimitri P. Bertsekas. 1999. *Nonlinear Programming* (2nd ed.). Athena Scientific.
- [4] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. 2011. The Price of Fairness. *Operations research* 59, 1 (Jan. 2011), 17–31. <https://doi.org/10.1287/opre.1100.0865>
- [5] Vivek S Borkar. 1997. Stochastic approximation with two time scales. *Systems & Control Letters* 29, 5 (1997), 291–294.
- [6] Vivek S. Borkar. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency.
- [7] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. 2017. Multi-objective Bandits: Optimizing the Generalized Gini Index. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 625–634. <https://proceedings.mlr.press/v70/busa-fekete17a.html>
- [8] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviy-chuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533* (2020).
- [9] Zimeng Fan, Nianli Peng, Muhang Tian, and Brandon Fain. 2023. Welfare and Fairness in Multi-objective Reinforcement Learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1991–1999. <https://doi.org/10.48550/arXiv.2212.01382> Also available as arXiv:2212.01382.
- [10] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32-1.
- [11] Zoltán Gábor, Zsolt Kalmár, and Csaba Szepesvári. 1998. Multi-criteria Reinforcement Learning. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 197–205.
- [12] Corrado W Gini. 1912. Variability and mutability, contribution to the study of statistical distributions and relations. *Studi Economico-Giuridici della R. Università de Cagliari* (1912).
- [13] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/be3159ad04564bfb90db9e32851ebf9c-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/be3159ad04564bfb90db9e32851ebf9c-Paper.pdf)
- [14] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. 1984. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA* 21, 1 (1984), 2022–2023.
- [15] Nathan Jay, Noga Rotman, Brighten Godfrey, Michael Schapira, and Aviv Tamar. 2019. A deep reinforcement learning perspective on internet congestion control. In *International Conference on Machine Learning*. PMLR, 3050–3059.
- [16] Jiechuan Jiang and Zongqing Lu. 2019. Learning Fairness in Multi-Agent Systems. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/10493aa88605cad5ab4752b04a63d172-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/10493aa88605cad5ab4752b04a63d172-Paper.pdf)
- [17] Peizhong Ju, Arnob Ghosh, and Ness Shroff. 2024. Achieving Fairness in Multi-Agent MDP Using Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=yoVq2BGQdP>
- [18] Vijay Konda and John Tsitsiklis. 1999. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems*, S.olla, T. Leen, and K. Müller (Eds.), Vol. 12. MIT Press. [https://proceedings.neurips.cc/paper\\_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)
- [19] Harold W. Kuhn and Albert W. Tucker. 1951. Nonlinear Programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Jerzy Neyman (Ed.). University of California Press, Berkeley, CA, 481–492. Symposium held in 1950; volume published in 1951.
- [20] Robert K Merton. 1968. The Matthew effect in science: The reward and communication systems of science are considered. *Science* 159, 3810 (1968), 56–63.
- [21] Jeonghoon Mo and Jean Walrand. 2002. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on networking* 8, 5 (2002), 556–567.
- [22] Hervé Moulin. 2004. *Fair division and collective welfare*. MIT press.
- [23] John F Nash et al. 1950. The bargaining problem. *Econometrica* 18, 2 (1950), 155–162.
- [24] Yasar Sinan Nasir and Dongning Guo. 2019. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. *IEEE Journal on selected areas in communications* 37, 10 (2019), 2239–2250.
- [25] Angelia Nedić and Asuman Ozdaglar. 2009. Subgradient Methods for Saddle-Point Problems. *Journal of Optimization Theory and Applications* 142, 1 (2009), 205–228.
- [26] Włodzimir Ogryczak, Hanan Luss, Michał Pióro, Dritan Nace, and Artur Tomaszewski. 2014. Fair optimization and networks: A survey. *Journal of Applied Mathematics* 2014, 1 (2014), 612018.
- [27] James Orr and Ayan Dutta. 2023. Multi-agent deep reinforcement learning for multi-robot applications: A survey. *Sensors* 23, 7 (2023), 3625.
- [28] Giseung Park, Woohyeon Byeon, Seongmin Kim, Elad Havakuk, Amir Leshem, and Youngchul Sung. 2024. The max-min formulation of multi-objective reinforcement learning: from theory to a model-free algorithm. In *Proceedings of the 41st International Conference on Machine Learning (Vienna, Austria) (ICML '24)*. JMLR.org, Article 1604, 27 pages.
- [29] David Portugal and Rui P Rocha. 2013. Distributed multi-robot patrol: A scalable and fault-tolerant framework. *Robotics and Autonomous Systems* 61, 12 (2013), 1572–1587.
- [30] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. 2020. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems* 33 (2020), 10199–10210.
- [31] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- [32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [33] Anthony F Shorrocks. 1980. The class of additively decomposable inequality measures. *Econometrica: Journal of the Econometric Society* (1980), 613–625.
- [34] Umer Siddique, Peilang Li, and Yongcan Cao. 2025. 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*. 2744–2746.
- [35] Umer Siddique, Paul Weng, and Matthieu Zimmer. 2020. Learning Fair Policies in Multi-Objective (Deep) Reinforcement Learning with Average and Discounted Rewards. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.), PMLR, 8905–8915. <https://proceedings.mlr.press/v119/siddique20a.html>
- [36] Joar Skalse, Lewis Hammond, Charlie Griffin, and Alessandro Abate. 2022. Lexicographic Multi-Objective Reinforcement Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3430–3436. <https://doi.org/10.24963/ijcai.2022/476> Main Track.
- [37] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (London, United Kingdom) (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2239–2248. <https://doi.org/10.1145/3219819.3220046>
- [38] Elise Van der Pol and Frans A Oliehoek. 2016. Coordinated deep reinforcement learners for traffic light control. *Proceedings of learning, inference and control of multi-agent systems (at NIPS 2016)* 8 (2016), 21–38.
- [39] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.
- [40] Chongjie Zhang and Julie A Shah. 2014. Fairness in multi-agent sequential decision-making. *Advances in Neural Information Processing Systems* 27 (2014).
- [41] Ruida Zhou, Tao Liu, Dileep Kalathil, P. R. Kumar, and Chao Tian. 2022. Anchor-Changing Regularized Natural Policy Gradient for Multi-Objective Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. [https://proceedings.neurips.cc/paper\\_files/paper/2022/hash/57fbc68cb318cad62c4ae4c91c83cba3-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2022/hash/57fbc68cb318cad62c4ae4c91c83cba3-Abstract-Conference.html)
- [42] Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. 2021. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12967–12978. <https://proceedings.mlr.press/v139/zimmer21a.html>