NEURIPS 2026

Safe-Fair MACPO

Burden-Fair Constrained Policy Optimization for Safe Multi-Agent Reinforcement Learning

Safety can protect the team while repeatedly burdening the same agent.

Ankita Kushwaha Kiran Ravish Preeti Pawan Kumar

IIIT Hyderabad

A safe team can still distribute burden unfairly

Team-level safety view

A policy satisfies the global safety budget and preserves the shared team reward.

Agent-level operational view

The same agent may repeatedly wait, yield, detour, absorb intervention, or lose access to a scarce resource.

Repeated safe episodes

A **B yields** **C**

A **B yields** **C**

A **B yields** **C**

Research question

How can a safe MARL policy account for persistent disparities in operational burden?

Problem setup: constrained cooperative Markov game

Decentralized execution

$$\pi_\theta(\boldsymbol{a}_t \mid \boldsymbol{o}_t) = \prod_{i=1}^{n} \pi_{\theta_i}(a_{i,t} \mid o_{i,t})$$

Safety-only objective

$$\max_{\pi_\theta} \quad J_R(\pi_\theta)$$

$$\text{s.t.} \quad J_S(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_t \gamma^t c_t^s\right] \le d_s$$

What this objective omits

J_S controls the team's aggregate safety cost. It does not specify which agent carries the burden required to maintain safety.

Design requirement

Add an agent-wise, history-sensitive fairness cost while keeping safety as the first constraint.

Burden fairness requires memory

Agent-wise burden

$$b_{i,t} = \boldsymbol{w}^\top \boldsymbol{h}_{i,t}, \qquad u_{i,t} = b_{i,t}/q_i,$$
$$\bar{u}_t = \frac{1}{n}\sum_{j=1}^{n} u_{j,t}.$$

$\boldsymbol{h}_{i,t}$ can encode waiting, detours, hazard exposure, control effort, resource denial, or intervention. And, q_i is agent $i's$ target burden share.

Temporal fairness debt

$$z_{i,t+1} = \mathrm{clip}\big(\rho z_{i,t} + u_{i,t} - \bar{u}_t, -Z_{\max}, Z_{\max}\big)\,.$$

Equal-capability experiments use $q_i = 1/n$.

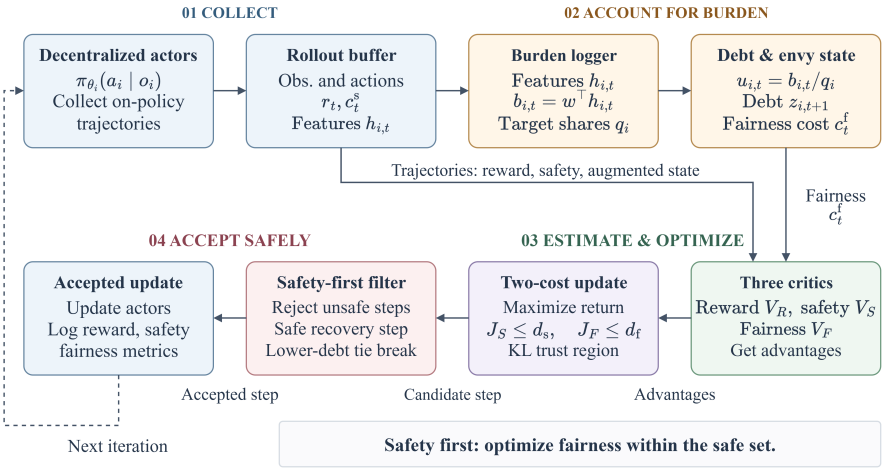
Fairness cost

$$c_t^f = \alpha_{\mathrm{debt}} \frac{1}{n}\sum_i |z_{i,t+1}| + \alpha_{\mathrm{envy}} \frac{1}{n(n-1)}\sum_{i\neq j}[u_{i,t} - u_{j,t} - \tau]_+.$$

Safe-fair constrained objective

$$\begin{aligned} \max_{\pi_\theta} \quad & J_R(\pi_\theta) \\ \text{s.t.} \quad & J_S(\pi_\theta) \le d_s, \\ & J_F(\pi_\theta) \le d_f. \end{aligned}$$

Safe-Fair MACPO architecture



Safety first: fairness influences optimization only inside the safety-feasible region.

Two-cost trust-region policy update

Agent i solves a local constrained step

$$
\begin{aligned}
\max_{\boldsymbol{x}_i} \quad & \boldsymbol{g}_i^\top \boldsymbol{x}_i \\
\text{s.t.} \quad & (\boldsymbol{b}_i^s)^\top \boldsymbol{x}_i + \hat{c}_i^s \le 0, \\
& (\boldsymbol{b}_i^f)^\top \boldsymbol{x}_i + \hat{c}_i^f \le 0, \\
& \tfrac{1}{2} \boldsymbol{x}_i^\top H_i \boldsymbol{x}_i \le \delta_i.
\end{aligned}
$$

KKT form

$$
\boldsymbol{x}_i^\star = \frac{1}{\lambda_0} H_i^{-1} \left(\boldsymbol{g}_i - \lambda_s \boldsymbol{b}_i^s - \lambda_f \boldsymbol{b}_i^f \right),
$$

with nonnegative multipliers.

- Reuses Fisher-vector products from MACPO.
- Low-dimensional dual search.
- Line search accepts predicted safety-feasible updates.
- Lower debt breaks safety-equivalent ties.

Theory: guarantees for the ideal constrained selector

Markovian debt

Appending Z_t to the environment state yields the augmented state $Y_t = (S_t, Z_t)$. Value functions can represent persistent burden without the full history.

Pareto efficiency

If the feasible set is nonempty and an exact selector exists, the reward-maximizing safe-fair solution is strongly Pareto efficient for $(J_R, -J_S, -J_F)$.

Conditional certificates

Under the stated dominance assumption, $J_S = 0$ implies zero hard violations. The fairness budget bounds discounted pairwise burden gaps, up to clipping residuals.

Scope of the theory

The results characterize the ideal constrained problem. They do not establish global convergence or automatic safety for finite-sample neural policy optimization.

Evaluation protocol and burden metrics

Task families

- Locomotion: HalfCheetah-2x3, Ant-2x4, ManyAgent-Ant
- Navigation: Safety-Gym MultiGoal, VMAS-SafeMultiGoal
- Resource sharing: Safety-Gym ChargeShare

Matched-budget comparison

Five independent training runs and 12 deterministic evaluation episodes per run. Most tasks use 180k environment steps.

Safety-first selection

1. Prioritize zero observed hard violations.
2. Prefer zero observed safety cost.
3. Compare return and burden balance jointly.

Complementary fairness metrics

Fairness cost, Jain burden index, Gini burden, envy gap, and debt AUC. No single metric fully characterizes fairness.

Matched budgets show favorable, task-dependent trade-offs

Ant-2x4 vs. MAPPO-Lag

Return

1789.29 → **1907.11**

Fairness cost

5.07 → **4.38**

Safe-Fair MACPO reports zero observed safety cost and hard violations.

HalfCheetah-2x3

Return

2503.33

Jain index

0.96

Zero observed safety cost and hard violations. The safety-feasible MACPO baseline reaches 1411.05 return and Jain 0.98.

ManyAgent-Ant vs. MAPPO-Lag

Return

3693.17 → **3852.73**

Fairness cost

15.55 → **10.53**

Fairness cost decreases by 32.3%, with zero observed safety cost and hard violations.

Evidence boundary

The method does not dominate every baseline on every task or metric. Matched recovery controls explain part of the observed safety and burden balance in some environments.

SAFE-FAIR MACPO

Thank you

Safety should protect the team without repeatedly assigning the cost of safety to the same agent.

Ankita Kushwaha Kiran Ravish Preeti Pawan Kumar

IIIT Hyderabad | NeurIPS 2026