

# Fairness Aware Reinforcement Learning via Proximal Policy Optimization

Gabriele La Malfa<sup>1</sup>, Jie M. Zhang<sup>1</sup>, Michael Luck<sup>2</sup>, Elizabeth Black<sup>1</sup>

<sup>1</sup>King’s College London

<sup>2</sup>University of Sussex

gabriele.la\_malfa@kcl.ac.uk, jie.zhang@kcl.ac.uk, Michael.Luck@sussex.ac.uk, elizabeth.black@kcl.ac.uk

## Abstract

Fairness in multi-agent systems (MAS) focuses on equitable reward distribution among agents in scenarios involving sensitive attributes such as race, gender, or socioeconomic status. This paper introduces fairness in Proximal Policy Optimization (PPO) with a penalty term derived from a fairness definition such as demographic parity, counterfactual fairness, or conditional statistical parity. The proposed method, which we call Fair-PPO, balances reward maximisation with fairness by integrating two penalty components: a retrospective component that minimises disparities in past outcomes and a prospective component that ensures fairness in future decision-making. We evaluate our approach in two games: the Allelopathic Harvest, a cooperative and competitive MAS focused on resource collection, where some agents possess a sensitive attribute, and HospitalSim, a hospital simulation, in which agents coordinate the operations of hospital patients with different mobility and priority needs. Experiments show that Fair-PPO achieves fairer policies than PPO across the fairness metrics and, through the retrospective and prospective penalty components, reveals a wide spectrum of strategies to improve fairness; at the same time, its performance pairs with that of state-of-the-art fair reinforcement-learning algorithms. Fairness comes at the cost of reduced efficiency, but does not compromise equality among the overall population (Gini index). These findings underscore the potential of Fair-PPO to address fairness challenges in MAS.

**Code** — <https://github.com/gabrielelamalfakcl/fairness-aware-ppo>

**Extended version** — <https://arxiv.org/pdf/2502.03953>

## Introduction

In multi-agent systems (MAS), agents interact in an environment to pursue individual or shared goals. Fairness in MAS focuses on whether the reward distribution mechanisms, driven by agent decisions or other processes, treat agents fairly. For instance, fair reinforcement learning explores methods to promote fairness by enabling agents to learn a fair policy (Reuel and Ma 2024); fair division addresses fair resource allocation (Lindner and Rothe 2016; Amanatidis et al. 2023); and negotiation designs methods

for fair bargaining resolution (Güth and Kocher 2014; Debove, Baumard, and André 2016).

On the other hand, in human society, fairness is framed in terms of inequality or discrimination between privileged and disadvantaged groups. Sensitive attributes, such as race, gender and socioeconomic status, define subgroups historically marginalised in workplaces, healthcare, education, and politics.<sup>1</sup> To enhance fairness, individuals (are often nudged to) adjust their behaviour towards those holding sensitive attributes. For example, giving up a seat on public transport for an elderly person illustrates a behavioural adjustment to promote fairness. For this reason, integrating fairness into agents’ policies has been an area of growing investigation (Reuel and Ma 2024).

Foundational works in social sciences (Griesinger and Livingston Jr. 1973; Liebrand 1984) have identified agents’ attributes as a crucial factor influencing fairness outcomes in games. In this sense, inspired by algorithmic fairness (Mitchell et al. 2021; Castelnovo et al. 2022), we propose sensitive attributes as characteristics that should not affect an agent’s expected reward. We apply metrics from the algorithmic fairness literature, specifically demographic parity, counterfactual fairness, and conditional statistical parity, to the MAS context and use them to constrain agent behaviour and obtain fair policies. Building on gradient-based algorithms in reinforcement learning, and inspired by the work of Zhang et al. (Zhang et al. 2022), we propose a fairness-aware Proximal Policy Optimization (PPO) (Schulman et al. 2017b) method, which we call Fair-PPO, that improves policy fairness. We modify the PPO objective function to include a penalty term derived from a fairness metric, allowing multi-objective optimisation of the policy that accounts for both performance and fairness.

Our proposed penalty has two components. The first component penalises total reward disparities between agents that differ by a sensitive attribute by looking at past outcomes. The second component penalises disparities in the expected rewards as per the estimate of the value function of each agent. In other words, the first component is retrospective, addressing disparities in past outcomes, while the second is

<sup>1</sup>Throughout the paper, we use the term ‘sensitive attribute’ instead of ‘protected characteristic’ to avoid any confusion with the legal meaning reported, for example, in the UK Equality Act.

prospective, encouraging fairness in the agent’s future expectations and decision-making.

In summary, the main contribution of this work is the novel Fair-PPO reinforcement learning algorithm, which extends PPO with the retrospective and prospective penalty components. We perform experiments in a version of the Allelopathic Harvest (AH) game (Leibo et al. 2019), a MAS that combines cooperation and competition in resource collection, where two groups of agents with different preferences regarding available resources navigate a dynamic environment. Agents are distinguished according to whether they hold some sensitive attribute: agents with this attribute move more slowly and so are potentially disadvantaged in resource collection. Our evaluation of Fair-PPO extends to HospitalSim (HS), a novel hospital simulation. This MAS integrates classification, matching, and resource allocation as agents manage patients with different mobility and priority needs moving through various hospital areas.

Our experiments show that: (i) Fair-PPO produces fairer policies than PPO across the fairness metrics we propose and, through the retrospective and prospective penalty components, reveals a wide spectrum of strategies to improve fairness; (ii) fairness comes at the cost of efficiency but does not compromise equality among the overall population of agents, as measured by the Gini index; and (iii) Fair-PPO is well-suited for finding fair policies in a range of simulation types, from cooperative and competitive to coordination-based scenarios.

The paper is structured as follows: in Section Related Work, we review the literature on fairness in MAS, fairness in reinforcement learning and algorithmic fairness. Section Preliminaries introduces preliminary concepts of MAS and PPO. In Section Fair-PPO, we detail the fairness metrics and the penalty component integration into PPO. Sections Experiments and Results focus on evaluating our approach, presenting experimental results using AH and HS.

## Related Work

Our work is inspired by algorithmic fairness in seeking to contribute to fair reinforcement learning literature in MAS. We review recent prominent works in these fields.

### Fairness Measures in MAS

In MAS, fairness has been evaluated through methods tailored to the system design and objectives. A widely accepted and intuitive notion is *proportionality*: fairness is determined by the *proportion* of reward allocated to an agent or group compared to others. Proportionality is applied in many MAS where a resource is allocated to multiple agents, such as the ultimatum game (Güth and Kocher 2014; Debove, Baumard, and André 2016) or fair division games (Lindner and Rothe 2016; Amanatidis et al. 2023; Murhekar 2024). Beyond proportionality, *envy-freeness* ensures that no agent prefers another’s allocation, and *maximin share fairness* guarantees that each agent receives a share at least as good as they could secure by dividing resources themselves (Lipton et al. 2004; Budish 2011; Caragiannis, Kurokawa, and et al. 2019).

Alternative fairness measures propose to allocate rewards based on an agent’s merit (Joseph et al. 2016), enhancing *treatment equality* by equalising error rates across groups (Liu et al. 2017), or minimising *regret*, which captures the cost of deviating from the optimal trade-off between fairness and efficiency (Li, Liu, and Ji 2020; Patil et al. 2021; Jones, Nguyen, and Nguyen 2023; Barman et al. 2023). Last but not least are wealth distribution metrics such as *social welfare* (Kaplow and Shavell 2003; Brânzei, Gkatzelis, and Mehta 2017; Høgsgaard et al. 2023), which measure the total utility of all individuals based on a chosen outcome, and measures of inequality such as the Gini index (Farris 2010).

Our approach is based on the principle of ensuring a fair distribution of rewards among agents, similar to proportionality and equitable wealth distribution. However, the fundamental distinction lies in the basis of group formation through the sensitive attributes that permit the use of specialised fairness metrics from the algorithmic fairness literature to confront precise biases.

### Algorithmic Fairness

Algorithmic fairness addresses bias and discrimination in decision-making systems across domains such as justice (Berk 2019), education (Baker and Hawn 2021), credit scoring (Kozodoi, Jacob, and Lessmann 2022), and healthcare (Vyas, Eisenstein, and Jones 2020), (Giovanola and Tiribelli 2022), with a focus on protected attributes characterising discriminated groups. Fairness metrics are classified into group and individual categories. Group fairness metrics include *demographic parity* (Kamishima et al. 2012) and *equalised odds* (Hardt, Price, and Srebro 2016), which use confusion matrix rates, while *calibration-based metrics* evaluate prediction accuracy relative to group membership (Chouldechova 2016). Individual fairness, such as *counterfactual fairness* (Kusner et al. 2018), assesses consistency across factual and counterfactual scenarios.

### Fairness and Reinforcement Learning

Reinforcement learning traditionally focuses on learning policies that maximise expected rewards (Sutton and Barto 2018). However, this raises fairness concerns, as it can perpetuate biases and violate fairness and legal principles (Jabari et al. 2017). To address these issues, fairness constraints can be added to the optimisation process. For example, Siddique et al. (Siddique, Weng, and Zimmer 2020) and Zimmer et al. (Zimmer et al. 2021) define fairness as finding solutions that are *efficient* (benefiting everyone without waste), *impartial* (treating identical agents equally), and *equitable* (helping those who are worse off).

Chen et al. (2021) propose adjusting rewards through a multiplicative weight to achieve  $\alpha$ -fairness, while Zhang et al. (2014) implement *maximin fairness* to optimise the worst-performing agent’s outcome. Other works explore fairness across agent groups, including *demographic parity* (Jiang and Lu 2019; Wen, Bastani, and Topcu 2021; Chi et al. 2022). Some contributions address real-world complexities, such as agents with differing characteristics or

preferences, necessitating tailored fairness mechanisms (Yu, Siddique, and Weng 2023; Ju, Ghosh, and Shroff 2024).

To contextualise Fair-PPO as an algorithm to enhance fairness in decentralised multi-agent reinforcement learning, we benchmark it to FEN (Jiang and Lu 2019) and SOTO (Zimmer et al. 2021). While Fair-PPO, FEN and SOTO all provide solutions to balance efficiency and fairness, they diverge significantly in their architectural and optimisation strategies. Fair-PPO incorporates a penalty term based on tailored fairness definitions, such as demographic parity, into its optimisation objective. FEN adopts a hierarchical learning system and defines a fair-efficient reward to guide agents toward fair outcomes. SOTO, instead, optimises a social welfare function, and it is flexible to fairness metrics adaptation.

## Preliminaries

In this section, we first define the elements composing a MAS and then define gradient-based policies and PPO.

### Multi-Agent Systems

A MAS consists of multiple agents acting in an environment to achieve their goals. Let  $\mathcal{N} = \{1, \dots, n\}$  be the set of  $n$  agents. We denote a MAS as  $\mathcal{M} = (S, s_0, A, \mathcal{N}, At, P)$ , where  $S$  is the set of possible environment states,  $s_0 \in S$  is the initial state,  $A = A_1 \times \dots \times A_n$  is the joint action space, where  $A_i$  is the set of actions available to agent  $i \in \mathcal{N}$ ;  $At$  is the set of attributes available to the agents and we designate a single binary sensitive attribute  $z \in At$  for fairness considerations;  $P : S \times A \rightarrow \Delta(S)$  is the non-deterministic state transition function, which maps a state-action pair to a probability distribution over the next states.

We define an agent  $i \in \mathcal{N}$  as a tuple of attributes, policy and individual reward function. For the designated sensitive attribute  $z \in At$ , we denote its value for agent  $i$  as  $z_i \in \{0, 1\}$ . This allows us to partition the agents into a sensitive group  $\mathcal{N}_1 = \{i \in \mathcal{N} \mid z_i = 1\}$  and a non-sensitive group  $\mathcal{N}_0 = \{i \in \mathcal{N} \mid z_i = 0\}$ . The stochastic policy  $\pi_i : S \rightarrow \Delta(A_i)$  of agent  $i$  is a function mapping any given state  $s \in S$  to a probability distribution over the agent’s actions  $A_i$ . We denote the probability of taking an action  $a_i$  in state  $s$  as  $\pi_i(a_i \mid s)$  and the joint policy as  $\pi = \{\pi_1, \dots, \pi_n\}$ . The reward function for agent  $i$  from state  $s_t$  to  $s_{t+1}$ , due to the joint action  $a_t$ , is  $R_i : S \times A \times S \rightarrow \mathbb{R}$ . In this way, an agent  $i$  receives a reward  $r_{t+1,i} = R_i(s_t, a_t, s_{t+1})$ .

We define a trajectory or episode as a sequence of states, actions and rewards  $\tau = (s_0, a_0, r_1, s_1, a_1, r_2, \dots)$ , where  $a_t = (a_{t,1}, \dots, a_{t,n})$  is the joint action at time  $t$  and  $r_{t+1} = (r_{t+1,1}, \dots, r_{t+1,n})$  is the resulting joint rewards. The total reward achieved by an agent  $i$  over a trajectory  $\tau$  of length  $T$  is  $G_i(\tau) = \sum_{t=1}^T r_{t,i}$ .

The probability of a trajectory  $\tau = (s_0, a_0, \dots, s_T)$  occurring under a joint policy  $\pi$  is  $p(\tau \mid \pi) = p(s_0) \prod_{t=0}^{T-1} P(s_{t+1} \mid s_t, a_t) \prod_{i=1}^n \pi_i(a_{t,i} \mid s_t)$ , which is the product of the initial state probability, the environment transition probability and the policy probabilities at each step. The expected reward of an agent  $i$  is  $J_i(\pi) = \sum_{\tau} p(\tau \mid$

$\pi)G_i(\tau)$ , which is the sum of the total rewards  $G_i(\tau)$  from every possible trajectory multiplied by the probability of that trajectory.

### Proximal Policy Optimization

In reinforcement learning, policy-gradient methods optimise the parameters  $\theta$  of a policy  $\pi_{\theta}$  to maximise the agent’s objective function. These parameters are updated through gradient ascent. To improve the stability of the learning process, modern algorithms implement methods that constrain the policy update, like Trust Region Policy Optimization (Schulman et al. 2017a). PPO works with Clipped Surrogate Objective (CLIP), limiting, within a small range, the change in the probability ratio of actions between the old and new policies.

PPO integrates policy optimisation and value function accuracy into a single objective function as follows:

$$L_t(\theta_i) = \hat{\mathbb{E}}_t [L_t^{\text{CLIP}}(\theta_i) - c_1 L_t^{\text{VF}}(\theta_i) + c_2 H[\pi_{\theta_i}](s_t)] \quad (1)$$

where  $t$  is a timestep,  $i$  is the agent and  $c_1$  and  $c_2$  are weighting coefficient. The objective is composed of the following three components.

The Clipped Surrogate Objective  $L^{\text{CLIP}}$  constrains the policy update to improve the stability:

$$L_t^{\text{CLIP}}(\theta_i) = \min \left( \psi_t(\theta_i) \hat{A}_{t,i}, \text{clip}(\psi_t(\theta_i), 1 - \epsilon, 1 + \epsilon) \hat{A}_{t,i} \right)$$

where  $\psi_t(\theta_i)$  is the ratio of the probability of taking action  $a_t$  under the new policy to the probability of taking it under the old policy;  $\hat{A}_{(t,i)}$  is the advantage function for agent  $i$  at step  $t$ , which estimates how much better the action  $a_t$  is compared to the average value of the state.

The Value Function Loss improves the accuracy of the policy’s value estimation:

$$L_t^{\text{VF}}(\theta_i) = (V_{\theta_i}(s_t) - G_{t,i})^2$$

where  $V_{\theta_i}(s_t)$  is the estimate of the return for state  $s_t$ , and  $G_{t,i}$  is the empirical discounted rewards for agent  $i$  from state  $s_t$  onwards.

Finally, the Entropy Bonus encourages exploration by promoting more diverse action selection:

$$H[\pi_{\theta_i}](s_t) = - \sum_{a_i \in A_i} \pi_{\theta_i}(a_i \mid s_t) \log \pi_{\theta_i}(a_i \mid s_t).$$

### Fair-PPO

In this section, we formalise demographic parity, counterfactual fairness and conditional statistical parity in MAS when sensitive attributes are involved; then, we detail the demographic parity-based penalty and integrate it as a constraint into the PPO objective function. The objective function formulations incorporating counterfactual fairness and conditional statistical parity penalties can be found in Section A of the supplementary material.

## Fairness Metrics in MAS

Fair-PPO incorporates the notions of demographic parity, counterfactual fairness and conditional statistical parity as a penalty term into the gradient optimisation. These definitions compare the average expected returns of agent groups partitioned by the sensitive attribute  $z$ .

Let  $\pi$  be the joint policy. We define the average expected rewards for any subgroup of agents  $\mathcal{G} \subseteq \mathcal{N}$  as:  $\bar{J}_{\mathcal{G}}(\pi) = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} J_i(\pi)$ .

**Definition 0.1 (Demographic Parity).** In a MAS  $\mathcal{M}$ , demographic parity is satisfied if the average expected rewards are the same for both sensitive and non-sensitive groups  $\mathcal{N}_1$  and  $\mathcal{N}_0$ , i.e.,  $\bar{J}_{\mathcal{N}_1}(\pi) = \bar{J}_{\mathcal{N}_0}(\pi)$ . When demographic parity does not hold, we quantify the disparity as the absolute difference between these averages:

$$\Delta DP(\pi) = |\bar{J}_{\mathcal{N}_1}(\pi) - \bar{J}_{\mathcal{N}_0}(\pi)| \quad (2)$$

**Definition 0.2 (Counterfactual Fairness).** Consider a factual system  $\mathcal{M}$  with joint policy  $\pi$ , and a counterfactual system  $\mathcal{M}'$  which is identical to  $\mathcal{M}$  except that the sensitive attribute  $z$  is flipped for all agents. Let  $\pi'$  be the joint policy in  $\mathcal{M}'$ . Counterfactual fairness is satisfied if the expected return for every agent remains unchanged, i.e.,  $J_i(\pi) = J_i(\pi')$   $\forall i \in \mathcal{N}$ . When counterfactual fairness does not hold, we quantify the total disparity by summing the individual differences:

$$\Delta CF(\pi, \pi') = \sum_{i \in \mathcal{N}} |J_i(\pi) - J_i(\pi')| \quad (3)$$

**Definition 0.3 (Conditional Statistical Parity).** Let  $LF \in (At \setminus \{z\})$  be a non-sensitive, legitimate attribute, whose value for agent  $i$  is denoted by  $LF_i$ . Let  $\mathcal{L}_{LF}$  be the set of possible values for  $LF$ . Conditional statistical parity is satisfied if demographic parity holds across all subgroups of agents that share the same value for the attribute  $LF$ .

Formally, for each value  $v \in \mathcal{L}_{LF}$ , let  $\mathcal{N}_v = \{i \in \mathcal{N} \mid LF_i = v\}$ . Conditional statistical parity holds if, for all  $v \in \mathcal{L}_{LF}$ :  $\bar{J}_{\mathcal{N}_1 \cap \mathcal{N}_v}(\pi) = \bar{J}_{\mathcal{N}_0 \cap \mathcal{N}_v}(\pi)$ .

The total disparity is the sum of the disparities across each legitimate factor value:

$$\Delta CSP(\pi, LF) = \sum_{v \in \mathcal{L}_{LF}} |\bar{J}_{\mathcal{N}_1 \cap \mathcal{N}_v}(\pi) - \bar{J}_{\mathcal{N}_0 \cap \mathcal{N}_v}(\pi)| \quad (4)$$

## Fairness Metrics for Fair PPO

PPO focuses on maximising agents' rewards. This section extends PPO by incorporating fairness constraints directly into the optimisation process. Specifically, we add a penalty term to the PPO objective function, discouraging behaviours that diverge from any of the selected disparities above. However, designing the penalty accounting only for collected rewards can limit the effectiveness of learning policies, particularly in stochastic environments and the early training process stage. Thus, we compose the penalty of two parts: a retrospective component based on the total rewards collected in the past, and a prospective component based on the

critic's real-time value estimates. We modify Eq. 1 using the  $\Delta DP$  metric as follows:

$$L_t^{\text{Fair-PPO}}(\theta_i) = \hat{\mathbb{E}}_t [L_t^{\text{CLIP}}(\theta_i) - c_1 L_t^{\text{VF}}(\theta_i) + c_2 H[\pi_{\theta_i}](s_t) - \lambda \cdot L_t^{\text{fair}}]$$

where  $L_t^{\text{fair}}$  is one of the fairness penalties, and the hyperparameter  $\lambda$  controls the magnitude of the fairness penalty.

To formulate the group-based penalties, we first define the sample average return  $\bar{G}$  and sample average value  $\bar{V}$  for any agent subgroup  $\mathcal{G} \subseteq \mathcal{N}$ :

$$\bar{G}_{\mathcal{G}}(\tau) = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} G_i(\tau) \quad \text{and} \quad \bar{V}_{\mathcal{G}}(s_t) = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} V_{\theta_i}(s_t) \quad (5)$$

**Demographic Parity Penalty.** This penalty is the sample-based equivalent of the  $\Delta DP$  metric, penalising the difference in average outcomes between the groups  $\mathcal{N}_1$  and  $\mathcal{N}_0$ .

$$L_t^{\text{fair-DP}} = \alpha \cdot |\bar{G}_{\mathcal{N}_1}(\tau) - \bar{G}_{\mathcal{N}_0}(\tau)| + \beta \cdot |\bar{V}_{\mathcal{N}_1}(s_t) - \bar{V}_{\mathcal{N}_0}(s_t)| \quad (6)$$

where the first term, the retrospective component (weighted by  $\alpha$ ), compares the average total return of the two groups over the last trajectory  $\tau$ . The second term, the prospective component (weighted by  $\beta$ ), compares the groups' average value estimates for the current state  $s_t$ . The parameters  $\alpha$  and  $\beta$  balance their contributions.

## Experimental Setup

We test Fair-PPO in two MAS: a version of the Allelopathic Harvest (AH) (Leibo et al. 2019) and HospitalSim (HS), an original hospital simulation.<sup>2</sup>

**Allelopathic Harvest.** In this setup, two groups of agents with distinct preferences, one for red berries and the other for blue berries, move in a grid and engage in cooperative dynamics within their respective groups, i.e., they plant and ripen berry bushes of their favourite colour and compete against the opposing group by blocking agents with opposing preferences. The objective for each group is to ensure the proliferation of their preferred berry, thereby maximising their rewards. The allelopathic element of AH is reflected in the bushes' spontaneous growth, which follows a linear function based on the number of red and blue bushes already on the ground. Thus, actions such as changing the colour of berries on a bush play a fundamental role in enhancing the growth of favoured bushes and ensuring future rewards for the group. Within each group, half of the agents can move every turn, while others are limited to moving only every two turns. This difference in mobility is a sensitive attribute which can be interpreted as an impairment.

**HospitalSim.** HospitalSim (HS) simulates a hospital where three agents, an escort dispatcher, a triage router, and a doctor manager, coordinate to maximise patient care efficiency while ensuring treatment fairness between sensitive and non-sensitive patients.

<sup>2</sup>More details of the game in the supplementary material.

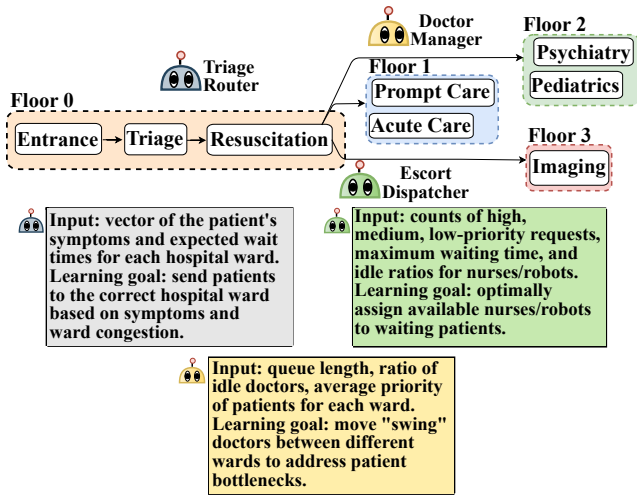


Figure 1: HospitalSim workflow and input/task descriptions for each agent.

The triage router assigns patients to six wards based on a multi-level mapping of symptoms/wards, and the expected wait time for each ward. Thus, the agent should balance a trade-off between sending patients to their specialised ward, which may be busy, versus a faster backup ward that offers a reduced reward. The escort dispatcher assigns available nurses and robots to impaired patients who need help moving in the hospital. Such an assignment is based on patients’ priority, how long they have been waiting for an escort and the proportion of nurses and robots available. The doctor manager mitigates overcrowding by assigning swing doctors to different wards. This decision considers each ward’s queue length, doctor availability, and the average patient priority.

Patients arrive throughout the day with defined priorities, illnesses, and physical impairments (sensitive attribute); impaired patients have slow movement and need assistance, leading to potential delays. The waiting time is penalised based on the priority of the patient. The system’s learning agents integrate with rule-based agents, handling static functions like patient intake and diagnosis. A representation of HS is shown in Figure 1.

## Train

In AH, we train two separate Fair-PPO models: one for agents with the sensitive attribute and another for those without. Training is conducted over 1000 episodes, each representing a new game, with 3000 time steps per episode. In HS, we train one model for each agent: the escort dispatcher, triage router, and doctor manager. Training spans 2000 episodes, with each episode simulating a new hospital day of twelve hours and a stream of 300 patients per day drawn from a generated pool. The simulation is event-driven, with parallel agents’ actions that sequentially drive the game (thus, there are no time steps). Therefore, Fair-PPO employs a decentralised training approach, with a centralised fairness penalty calculation to add to the PPO objec-

tive function.

In both games, after each episode, two penalty fairness components are calculated: the reward disparity is the difference in the total rewards accumulated by the non-sensitive group versus the sensitive group; the state-value disparity is the difference in the predicted state values between the two groups. Both components are normalised and weighted by  $\alpha$  and  $\beta$  parameters. However, while in AH the fairness penalty is calculated based on the rewards experienced by the learning agents (red and blue agents), in HS the penalty is based on the total and expected rewards of the patients who are not the learning agents.

We train Fair-PPO using a penalty weighted by  $\alpha$  and  $\beta$ , addressing demographic parity, counterfactual fairness, and conditional statistical parity separately. The parameters  $\alpha$  and  $\beta$  range from 0 to 1, taking discrete values with a step of 0.25, for a total of 25 combinations. PPO is the specific instance of Fair-PPO with  $\alpha = \beta = 0$ .

## Test and evaluation

We tested the policy on 500 episodes of 1500 time steps for AH and 500 episodes (days) of twelve hours with a stream of 300 patients per day drawn from a generated pool for HS (no time steps as explained in the previous section). Each metric is averaged over the time steps and episodes. In both AH and HS, demographic parity is the difference between the rewards collected by agents with and without a sensitive attribute (impairment). For conditional statistical parity, the legitimate factor is agents’ preference for red or blue berries in AH, and patients’ priority (high, medium and low) in HS.

## Benchmarks

We benchmark Fair-PPO against PPO ( $\alpha = 0$  and  $\beta = 0$ ), FEN (Jiang and Lu 2019) and SOTO (Zimmer et al. 2021). The other hyperparameters of each algorithm are reported in the supplementary material.

In both AH and HS, FEN is implemented following a hierarchical policy structure where a controller selects from several sub-policies. One sub-policy optimises for efficiency while the others explore fair strategies. For example, in AH, an efficient action is eating a preferred berry, while a fair one is ripening a berry for another agent. In HS, the triage router’s efficiency policy learns to send a patient to the correct and least congested ward. In contrast, a fairness-oriented policy might prioritise patients with an impairment to better balance the utility across the entire system. The agent can change the policy every  $T_{macro}$  macro-timesteps. The controller learns to maximise a fair-efficient reward that encourages actions oriented to the utility of all agents while penalising any deviation from fairness. In AH, FEN shares the parameters among all agents, employing a gossip protocol to average utility, while in HS, three independent FEN instances are employed for each learning agent.

In AH and HS, SOTO is implemented by providing each learning agent with a dual PPO-based policy that balances individual efficiency with collective fairness. In both cases, a self-oriented stream maximises local rewards. In contrast, a team-oriented stream optimises a social welfare function using an  $\alpha$ -fairness mechanism that weights an agent’s advan-

tage based on the utilities of others. The core difference of SOTO implementation in AH and HS is that in AH, the coordination of agents happens within each group (red and blue agents), and the fairness is calculated between non-sensitive and sensitive agents; in HS, the learning agents maximise their utility through the self-oriented policy, but the fairness-oriented policy minimises disparity between patients with and without impairment that are not learning agents.

## Results

In this section, we analyse the main findings from experiments in AH and HS.

**Fairness and efficiency measures.** Our focus is on demographic parity and conditional statistical parity calculated on agents’ rewards as the core fairness metrics. In AH, we also quantify fairness using the Gini index to assess reward distribution among agents, while overall efficiency is measured by the agents’ average accumulated rewards throughout the simulation. For HS, we evaluate hospital efficiency using several key metrics: the average number of patients treated daily, patients’ average waiting time for an escort, and the escorts’ average travel time. We specifically evaluate the triage router policy by analysing the percentage of patients directed to the primary ward, backup ward, or misclassified. The doctor manager is assessed by the average number of trips made by swing doctors.

### Fair-PPO generates fairer policies than PPO

Fair-PPO, through different combinations of  $\alpha$  and  $\beta$ , produces several policies with lower demographic disparity than PPO in both AH and HS. Figure 2 shows three clusters of strategies (from light pink to dark purple) obtainable through adjusting the retrospective and prospective components of Fair-PPO in AH. Higher demographic parity generally corresponds to higher Price of Fairness, computed as the reward percentage agents renounce in favour of fairness compared to PPO. A similar result can be observed in HS, see the first (patients treated for efficiency) and the second column (demographic disparity for fairness) of Table 1. A lower disparity is also registered within the subgroups of agents (conditional statistical disparity), in both AH and HS: for example, Table 2 first column, shows the improvement within the priority groups of patients (high, medium and low priority).

### Fair-PPO: comparison with FEN and SOTO

In AH, Fair-PPO matches SOTO on fairness-reward trade-offs, performing similarly in terms of equal reward distribution among agents. On the other hand, FEN records the lowest disparity but highly underperforms all algorithms in terms of inequality (Gini index) and average rewards. These results are reported in Figure 3.

In HS, Fair-PPO finds the fairest solutions as reported in Table 1, revealing a trade-off between lower efficiency, in terms of daily average treated patients, and lower demographic disparity. On the other hand, SOTO and FEN produce very prolific policies in terms of daily average treated patients. Looking at Figure 5, where the training phase is

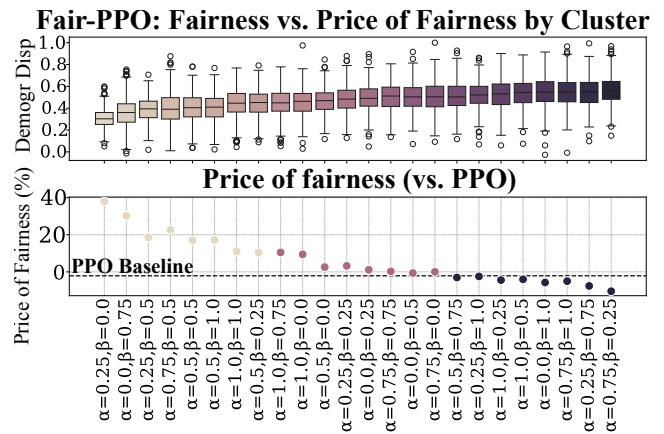


Figure 2: Allelopathic Harvest: Fair-PPO ( $\alpha$ ,  $\beta$ ) demographic disparity boxplots clusters (light pink, purple and dark purple) and the Price of Fairness vs. PPO.

reported, SOTO learns an efficient and fair policy, without succeeding in reaching the level of parity of Fair-PPO. The same trend is confirmed by the test in Table 1 and 2 throughout demographic disparity and conditional statistical disparity columns.

### Interpreting the strategies

By varying  $\alpha$  and  $\beta$ , Fair-PPO explores a wide variety of strategies in AH. For example, Figure 4, left chart, shows the action frequencies of a representative of each cluster of Figure 2 (light pink, purple and dark purple). Low demographic disparity (light pink) corresponds to a strategy that heavily prioritises ripening the bush, which is the action most beneficial for other agents collecting rewards. The highest disparity (dark purple) focuses more on berry consumption, the selfish action. SOTO reaches similar fairness results to Fair-PPO with lower disparity through a more uniformly distributed strategy, while FEN agents learns a berry-eating policy, which, later in the game, exhausts resources and dramatically decreases the rewards to collect.

In HS, Fair-PPO reduces the disparity between impaired and unimpaired patients by alleviating overcrowding. Since the hospital must admit all 300 patients daily, Fair-PPO optimises the triage router to heavily direct patients to backup wards. This strategy is evidenced by the data in the last three columns of Table 1 and further illustrated in the training graph in Figure 5. We also observe that a reduction in demographic disparity aligns with decreased patients’ waiting times for escorts and lower escort travel time. Both metrics underscore improved access to supportive resources for impaired patients. In conclusion, although SOTO manages to optimise the triage router better than other algorithms to send the patients to the specialised ward and also to reduce the swing doctor moves (see Table 1, fifth column), it does not achieve the same parity level of Fair-PPO and PPO.

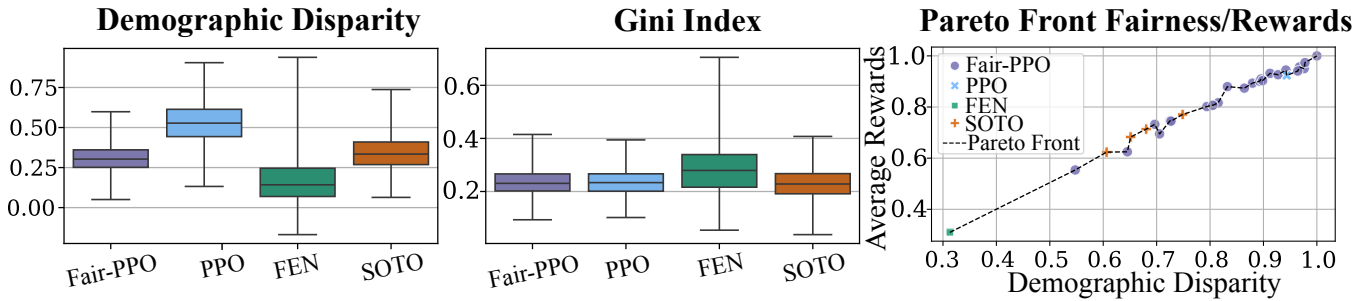


Figure 3: Allelopathic Harvest. Fair algorithms’ performance (average demographic disparity and Gini index) for each architecture (Fair-PPO, PPO, FEN and SOTO) and disparity/rewards Pareto frontier plotted for all algorithms.

Algorithm	Patients Treated (Daily Average)	Dem Disparity (Rewards)	Patient Wait Escort (Minutes)	Escort Travel Time (Minutes)	Swing Doctor Moves (Average)	Perfect Routing (% of Patients)	Backup Routing (% of Patients)	Incorrect Routing (% of Patients)
Fair-PPO $\alpha = 0.25, \beta = 0.25$	105.74	6.98	4.01	3.67	1923.93	6.04	44.93	49.03
Fair-PPO $\alpha = 0.25, \beta = 0.5$	105.84	6.62	3.94	3.55	1925.52	6.00	37.13	56.86
Fair-PPO $\alpha = 0.5, \beta = 0.25$	68.69	5.31	4.07	3.77	1924.72	3.83	43.78	52.39
PPO	105.92	6.83	3.94	3.56	1923.99	6.01	38.17	55.82
FEN	223.17	11.76	4.54	4.09	1924.25	19.65	18.42	61.93
SOTO $\alpha$ -fair = 1.0	240.03	9.10	4.51	4.06	1167.36	27.64	21.17	51.19

Table 1: HospitalSim. Average efficiency (patients treated) and fairness performance (demographic disparity) across Fair-PPO and benchmark, sorted according to the lowest demographic disparity. Further metrics are reported to complete the analysis.

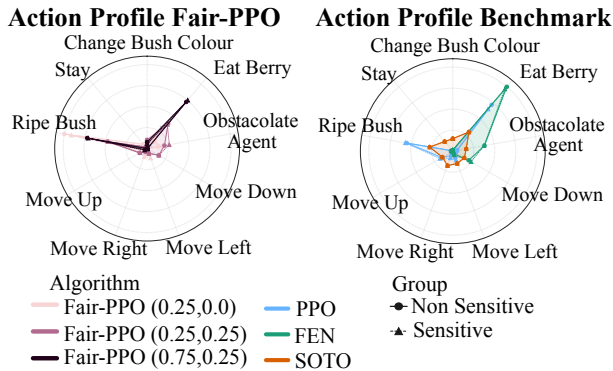


Figure 4: Allelopathic Harvest. Action frequencies of Fair-PPO and benchmarks. For Fair-PPO, we show one algorithm for each cluster of Figure 2.

Algorithm	Cond Stat Disp			Patients Treated (Daily Average)
	High	Med	Low	
Fair-PPO $\alpha=1.0, \beta=0.5$	5.43	5.45	7.63	63.75
Fair-PPO $\alpha=0.25, \beta=1.0$	5.43	5.43	7.54	64.05
Fair-PPO $\alpha=1.0, \beta=0.0$	5.19	5.39	7.44	64.15
PPO	5.32	5.30	8.10	63.85
FEN	11.77	11.80	11.63	223.76
SOTO $\alpha$ -fair=2.0	9.17	9.14	9.11	239.95

Table 2: HospitalSim. Average conditional statistical disparity (high, medium, low patients’ priority) and efficiency (patients treated) performance across Fair-PPO and benchmark, sorted for best demographic disparity.

## Conclusion

This paper extends PPO by incorporating a penalty based on fairness metric violations in the objective function. We design two penalty components: a retrospective component that addresses fairness violations based on past rewards and a prospective component that anticipates future fairness violations by leveraging the value function to estimate upcoming rewards. We refer to this algorithm as Fair-PPO. We found that Fair-PPO can discover various strategies to reduce unfairness across metrics for different levels of efficiency in both cooperative and competitive games (AH) and coordination-based simulations (HS). This work represents a step towards a fairness-aware PPO based on metrics in MAS involving agents with and without sensitive attributes. The implementation of the HS aims to stimulate fair reinforcement learning research towards real-world applications.

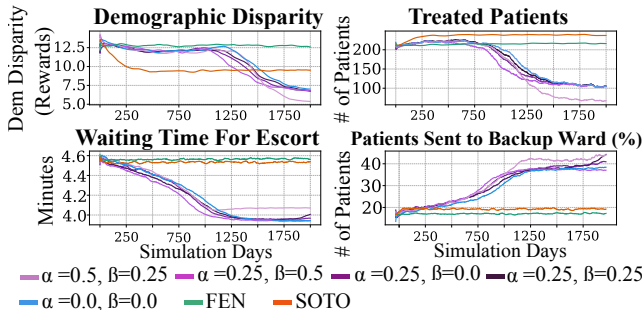


Figure 5: HospitalSim. Fairness (demographic disparity) and efficiency (daily treated patients and patients’ waiting time for escort) metrics convergence over the training for Fair-PPO and benchmarks.

## Ethical Statement

This work shows how a penalty-based approach can make a multi-agent policy converge towards fairer behaviours. The penalty is based on metrics involving agents with and without sensitive attributes. Thus, this work requires careful ethical considerations. First, the fairness metrics used and the sensitive attribute are an inevitable simplification of fairness in reality. In particular, we recognise that the sensitive attribute cannot capture the complexity of real-world scenarios. Second, our results show a trade-off between fairness and efficiency of a system. The deployment of a model requires transparency and awareness in the presence of this trade-off, in particular when the applications are intended to imitate a hospital, as in the HS. In conclusion, we stress that the findings of the paper should be only a stimulus for discussion on computational simulations and fairness, rather than a direct basis for policy.

## Acknowledgments

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence ([www.safeandtrustedai.org](http://www.safeandtrustedai.org)).

## References

- Amanatidis, G.; Aziz, H.; Birmpas, G.; and et al. 2023. Fair division of indivisible goods: Recent progress and open questions. *Artificial Intelligence*, 322: 103965.
- Baker, R.; and Hawn, A. 2021. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, 32.
- Barman, S.; Khan, A.; Maiti, A.; and Sawarni, A. 2023. Fairness and welfare quantification for regret in multi-armed bandits. In *AAAI'23/IAAI'23/EAAI'23, AAAI'23/IAAI'23/EAAI'23*. ISBN 978-1-57735-880-0.
- Berk, R. 2019. Accuracy and Fairness for Juvenile Justice Risk Assessments. *Journal of Empirical Legal Studies*, 16(1): 175–194.
- Brânzei, S.; Gkatzelis, V.; and Mehta, R. 2017. Nash Social Welfare Approximation for Strategic Agents. [arXiv:1607.01569](https://arxiv.org/abs/1607.01569).
- Budish, E. 2011. The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes. *Journal of Political Economy*, 119(6): 1061–1103.
- Caragiannis, I.; Kurokawa, D.; and et al., M. 2019. The Unreasonable Fairness of Maximum Nash Welfare. *ACM Trans. Econ. Comput.*, 7(3).
- Castelnovo, A.; Crupi, R.; Greco, G.; and et al. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1).
- Chen, J.; Wang, Y.; and Lan, T. 2021. Bringing Fairness to Actor-Critic Reinforcement Learning for Network Utility Optimization. In *IEEE INFOCOM 2021*, 1–10.
- Chi, J.; Shen, J.; Dai, X.; and et al. 2022. Towards Return Parity in Markov Decision Processes. volume 151, 1161–1178. PMLR.
- Chouldechova, A. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. [arXiv:1610.07524](https://arxiv.org/abs/1610.07524).
- Debove, S.; Baumard, N.; and André, J.-B. 2016. Models of the evolution of fairness in the ultimatum game: a review and classification. *Evolution and Human Behavior*, 37(3): 245–254.
- Farris, F. A. 2010. The Gini Index and Measures of Inequality. *The American Mathematical Monthly*, 117(10): pp. 851–864.
- Giovanola, B.; and Tiribelli, S. 2022. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI Soc.*, 38(2): 549–563.
- Griesinger, D. W.; and Livingston Jr., J. W. 1973. Toward a model of interpersonal motivation in experimental games. *Behavioral Science*, 18(3): 173–188.
- Güth, W.; and Kocher, M. G. 2014. More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, 108: 396–409.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. NIPS'16, 3323–3331. Red Hook, NY, USA. ISBN 9781510838819.
- Høgsgaard, M. M.; Karras, P.; Ma, W.; and et al. 2023. Optimally Interpolating between Ex-Ante Fairness and Welfare. [arXiv:2302.03071](https://arxiv.org/abs/2302.03071).
- Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2017. Fairness in Reinforcement Learning. volume 70, 1617–1626. PMLR.
- Jiang, J.; and Lu, Z. 2019. *Learning fairness in multi-agent systems*. Red Hook, NY, USA: Curran Associates Inc.
- Jones, M.; Nguyen, H.; and Nguyen, T. 2023. An Efficient Algorithm for Fair Multi-Agent Multi-Armed Bandit with Low Regret. *AAAI'23*, 37(7): 8159–8167.
- Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in Learning: Classic and Contextual Bandits. In *NIPS'16*, volume 29.
- Ju, P.; Ghosh, A.; and Shroff, N. 2024. Achieving Fairness in Multi-Agent MDP Using Reinforcement Learning. In *ICLR*.
- Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *ECML PKDD*, 35–50. ISBN 978-3-642-33486-3.
- Kaplow, L.; and Shavell, S. 2003. Fairness versus Welfare: Notes on the Pareto Principle, Preferences, and Distributive Justice. *The Journal of Legal Studies*, 32(1): 331–362.
- Kozodoi, N.; Jacob, J.; and Lessmann, S. 2022. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3): 1083–1094.
- Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2018. Counterfactual Fairness. [arXiv:1703.06856](https://arxiv.org/abs/1703.06856).
- Leibo, J. Z.; Perolat, J.; Hughes, E.; and et al. 2019. Malthusian Reinforcement Learning. *AAMAS '19*, 1099–1107. Richland, SC. ISBN 9781450363099.

- Li, F.; Liu, J.; and Ji, B. 2020. Combinatorial Sleeping Bandits With Fairness Constraints. *IEEE Transactions on Network Science and Engineering*, 7(3): 1799–1813.
- Liebrand, W. B. G. 1984. The effect of social motives, communication and group size on behaviour in an N-person multi-stage mixed-motive game. *European Journal of Social Psychology*, 14(3): 239–264.
- Lindner, C.; and Rothe, J. 2016. *Cake-Cutting: Fair Division of Divisible Goods*, 395–491. Springer Berlin Heidelberg. ISBN 978-3-662-47904-9.
- Lipton, R. J.; Markakis, E.; Mossel, E.; and Saberi, A. 2004. On approximately fair allocations of indivisible goods. EC '04, 125–131. ISBN 1581137710.
- Liu, Y.; Radanovic, G.; Dimitrakakis, C.; Mandal, D.; and Parkes, D. C. 2017. Calibrated Fairness in Bandits. arXiv:1707.01875.
- Mitchell, S.; Potash, E.; Barocas, S.; and et al. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021): 141–163.
- Murhekar, A. 2024. Fair and Efficient Chore Allocation: Existence and Computation. In Larson, K., ed., *IJCAI-24*, 8500–8501. Doctoral Consortium.
- Patil, V.; Ghalme, G.; Nair, V.; and Narahari, Y. 2021. Achieving Fairness in the Stochastic Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 22(174): 1–31.
- Reuel, A.; and Ma, D. 2024. Fairness in Reinforcement Learning: A Survey. arXiv:2405.06909.
- Schulman, J.; Levine, S.; Moritz, P.; Jordan, M. I.; and Abbeel, P. 2017a. Trust Region Policy Optimization. arXiv:1502.05477.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017b. Proximal Policy Optimization Algorithms. arXiv:1707.06347.
- Siddique, U.; Weng, P.; and Zimmer, M. 2020. Learning fair policies in multiobjective (deep) reinforcement learning with average and discounted rewards. ICML'20.
- Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA. ISBN 0262039249.
- Vyas, D. A.; Eisenstein, L. G.; and Jones, D. S. 2020. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, 383(9): 874–882.
- Wen, M.; Bastani, O.; and Topcu, U. 2021. Algorithms for Fairness in Sequential Decision Making. arXiv:1901.08568.
- Yu, G.; Siddique, U.; and Weng, P. 2023. Fair Deep Reinforcement Learning with Preferential Treatment. In *ECAI*, 2922–2929.
- Zhang, C.; and Shah, J. A. 2014. Fairness in multi-agent sequential decision-making. NIPS'14, 2636–2644.
- Zhang, L.; Shen, L.; Yang, L.; and et al. 2022. Penalized Proximal Policy Optimization for Safe Reinforcement Learning. arXiv:2205.11814.
- Zimmer, M.; Glanois, C.; Siddique, U.; and Weng, P. 2021. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. arXiv:2012.09421.