ISSN 2683-345X
journal homepage: http://ijiemjournal.uns.ac.rs/

International Journal of Industrial Engineering and Management

Volume 17 / No 3 / September 2026 / 327 - 342

Original research article

Learning Equitable and Efficient Shift Rosters with Multi-Objective Deep Reinforcement Learning and Counterfactual Fairness Constraints in Operations

B. Thanh Khoa[a,*] 0000-0002-9878-2164, K. Makharov[b] 0000-0002-9341-4602,
B. Samandarov[c] 0000-0002-8296-0894, S. Jafarov[d] 0000-0003-4344-8350,
D. Raupov[e,f] 0000-0002-3336-0884, O. Djurabaev[g] 0000-0002-2801-6069

[a] *Industrial University of Ho Chi Minh City, Vietnam;*
[b] *Kimyo International University in Tashkent, Tashkent, Uzbekistan;*
[c] *Mamun University, Khiva, Uzbekistan;*
[d] *Urgench State University, Urgench, Uzbekistan;*
[e] *Tashkent State Technical University, Tashkent, Uzbekistan;*
[f] *University of Tashkent for Applied Sciences, Tashkent, Uzbekistan;*
[g] *Tashkent State University of Economics, Tashkent, Uzbekistan*

ABSTRACT

Shift rostering systems often optimize cost and coverage while treating fairness as a secondary objective. This study develops a multi-objective deep reinforcement learning framework for workforce scheduling that integrates Proximal Policy Optimization with a graph neural network encoder, a Conditional Value-at-Risk term for limiting extreme burden concentration, and a counterfactual fairness regularizer for reducing demographic sensitivity in assignment decisions. The framework was evaluated on 18 months of workforce data from a tertiary hospital and an industrial facility in Uzbekistan. Relative to mixed-integer linear programming and vanilla reinforcement learning baselines, the proposed model reduced the burden Gini coefficient by 41.3%, narrowed weekend disparity by 36.5%, and achieved adverse impact ratios above 0.97 for gender and age. These equity gains were obtained with a 1.8% increase in operational cost and a 9.4% reduction in understaffing minutes under demand shocks. The results indicate that fairness constraints can be embedded directly into policy learning while preserving service levels and improving resilience in high-constraint scheduling environments.

ARTICLE INFO

Article history:
Received December 3, 2025
Revised April 17, 2026
Accepted May 13, 2026
Published online August 13, 2026

Keywords:
Adverse impact;
Counterfactual fairness;
Deep reinforcement learning;
Multi-objective optimization;
Workforce rostering

*Corresponding author:
Bui Thanh Khoa
buithanhkhoa@iuh.edu.vn

1. Introduction

The efficient and equitable allocation of human capital is a cornerstone of operational excellence in modern healthcare and high-volume manufacturing [1]. As industries transition from manual administration to algorithmic management, the complexity of workforce rostering—assigning specific shifts to employees over a fixed planning horizon—has grown ex-

Published by the University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia.

https://doi.org/10.24867/IJIEM-421

ponentially [2], [3]. In rapidly modernizing economies like Uzbekistan, particularly within the healthcare sectors of Tashkent and the industrial hubs of Navoi, the dual pressure to maximize workforce utilization while ensuring compliance with evolving labor regulations has rendered traditional scheduling methods obsolete [4], [5]. While automated rostering systems promise significant cost reductions and improved service levels, they increasingly face scrutiny regarding "algorithmic bias," where black-box optimization tools inadvertently concentrate undesirable shifts among vulnerable employee subgroups [4], [6]. Consequently, the development of intelligent scheduling frameworks that can rigorously guarantee fairness without compromising operational efficiency has emerged as a critical imperative for the scientific community.

The academic literature on personnel scheduling has long been shaped by mathematical programming and metaheuristic search. Mixed-Integer Linear Programming (MILP) remains a strong benchmark for static instances with clearly defined constraints, but recent reviews show that personnel rostering is inherently multi-objective because organizations must balance coverage, cost, preferences, fatigue, and fairness at the same time [7], [8]. In practice, this balance becomes harder when absenteeism, demand shocks, and multi-skill staffing requirements require repeated re-optimization rather than a single static roster [9], [10]. Recent work from 2023 to 2025 has reinforced this shift toward more adaptive approaches. Reviews of digital and self-rostering systems in healthcare show that scheduling quality affects both organizational performance and worker outcomes, but they do not provide a mechanism for auditable algorithmic fairness [3]. Recent personnel scheduling studies have also expanded multi-objective formulations for multi-skill and client-sensitive environments, yet most still rely on weighted objectives, heuristic search, or deterministic optimization rather than sequential policy learning under uncertainty [8], [11]-[13]. At the same time, advances in learning-based scheduling and graph-based representations have shown that Deep Reinforcement Learning (DRL) and Graph Neural Networks (GNNs) can capture complex dependency structures more effectively in dynamic decision settings [14]-[16]. The fairness problem remains less developed than the efficiency problem. In most scheduling studies, fairness is represented by dispersion-based statistics such as variance, weekend balance, or Gini-type inequality, and these measures are typically added as soft penalties in the objective [8], [11], [12]. This design can reduce aggregate disparity, but it does not test whether assignment decisions remain stable when protected attributes are changed in a counterfactual setting. Recent work on fairness in reinforcement learning has emphasized that sequential decisions can accumulate long-term inequities and that causal or counterfactual analysis is necessary when sensitive attributes may influence future trajectories [17], [18]. That point is directly relevant to workforce rostering, where repeated assignment of undesirable shifts can compound fatigue and disadvantage across time. A clear gap therefore remains at the intersection of dynamic scheduling and algorithmic fairness. Existing DRL-based scheduling studies focus mainly on understaffing, overtime, and preference satisfaction, while fairness is usually handled as a soft balancing term rather than as a causal property of the policy itself. Existing optimization-based fairness models also provide limited protection against extreme burden concentration in the upper tail of the workload distribution. As a result, current approaches can improve average efficiency while still allowing protected groups or already overburdened employees to absorb a disproportionate share of undesirable assignments.

This study addresses that gap by formulating workforce rostering as a Multi-Objective Markov Decision Process (MOMDP) that combines operational efficiency, preference satisfaction, tail-risk control, and counterfactual fairness in a single learning framework [19]. The proposed model uses Proximal Policy Optimization (PPO) with a graph-based state representation to capture relational staffing constraints, a Conditional Value-at-Risk (CvaR) term to reduce extreme burden concentration, and a counterfactual regularizer to reduce demographic sensitivity in assignment probabilities [20]-[22]. The empirical contribution lies in testing this framework on two high-constraint operational settings, healthcare and industry, and evaluating whether measurable equity gains can be achieved without a prohibitive efficiency loss. The primary aim of this research is to design and validate a fairness-constrained DRL agent capable of generating equitable, low-cost shift rosters in dynamic, high-constraint environments. The specific objectives of this study are: (i) To formulate the workforce rostering problem as a MOMDP that jointly minimizes operational cost and burden concentration while improving counterfactual fairness. (ii) To develop a specialized PPO architecture with graph-based encoding and a CvaR loss component for controlling upper-tail burden exposure. (iii) To evaluate the proposed framework on real-world datasets from a large-scale hospital in Tashkent and an industrial facility in the Navoi region against mathematical programming and reinforcement learning baselines. This study is

guided by the following research questions: *Can a DRL agent improve the efficiency-equity trade-off relative to static optimization methods under strict fairness constraints? Does the imposition of counterfactual fairness materially reduce service-level attainment or increase operating cost?* This study makes three concrete contributions. First, it combines counterfactual fairness and burden-tail control within a single rostering policy, rather than treating fairness as a simple variance penalty. Second, it integrates a graph-based reinforcement learning architecture with fairness-aware optimization to address relational staffing constraints in dynamic settings. Third, it provides evidence from two under-studied operational contexts in Central Asia, showing how fairness-aware scheduling affects cost, coverage, preference satisfaction, and robustness under real demand variation.

2. Methodology

2.1 Study Design and Data Acquisition

To rigorously evaluate the efficacy of the proposed fairness-constrained scheduling framework, this study employed a retrospective observational design utilizing high-granularity longitudinal workforce data. The empirical analysis was grounded in two distinct operational environments located in Uzbekistan, selected to represent contrasting rostering challenges: the stochastic, high-urgency environment of a tertiary healthcare facility (Tashkent Emergency Medical Center) and the rigid, continuous-process environment of a heavy industrial facility (Navoi Mining & Metallurgy Combinat).

Data acquisition covered an 18-month observation window (January 2022 through June 2023), capturing the complete shift history, skill qualifications, and demographic attributes of the workforce. The healthcare dataset comprised records for 142 nursing staff and paramedics, organized into three shifts per day (morning, evening, night). Key constraints included strict nurse-to-patient ratios, skill-mix requirements (e.g., Advanced Life Support certification), and rest periods mandated by the Ministry of Health. The industrial dataset included 315 operators and technicians working a rotating 12-hour shift pattern. This environment imposed rigid safety constraints regarding maximum consecutive work hours and required specific certification tiers for machinery operation.

Prior to ingestion into the model, all datasets underwent a strict anonymization protocol to ensure compliance with data protection standards. Personally identifiable information was removed, and sensitive attributes (gender, age group, tenure) were encoded as categorical variables for the fairness analysis. The data was partitioned into a training set (first 12 months), a validation set (subsequent 3 months), and a hold-out test set (final 3 months) to evaluate the generalization capability of the DRL agent.

2.2 Mathematical Formulation of the Rostering Environment

We formulated the workforce scheduling problem as a MOMDP. This formalism allows for the sequential optimization of roster assignments where current decisions influence future state dynamics, particularly regarding accumulated fatigue and constraint tightness. The MOMDP is defined by the tuple presented in Equation (1), representing the state space, action space, state transition probability, and the vector of reward functions [16].

$$\mathrm{M} = \langle \mathrm{S}, \mathrm{A}, \mathrm{P}, \mathbf{R}, \gamma \rangle \tag{1}$$

where S represents the high-dimensional state space encompassing the current roster status, employee history, and forecasted demand. A denotes the discrete action space of assigning specific shift types to employees. $\mathrm{P} : \mathrm{S} \times \mathrm{A} \times \mathrm{S} \rightarrow [0,1]$ is the state transition probability function, which in this deterministic scheduling environment is governed by the rules of time progression and constraint accumulation. $\mathbf{R}$ is the vector of reward signals corresponding to efficiency, preference satisfaction, and fairness, and $\gamma \in [0,1)$ is the discount factor.

The state vector at time step t, denoted as s_t, must capture sufficient history to evaluate validity constraints. We defined s_t as a concatenation of the current demand profile, the accumulated work hours for each employee over the rolling planning horizon, and the "fatigue status" derived from consecutive shift patterns.

The action at time step t, a_t, represents the assignment of a shift pattern to the workforce. To manage the combinatorial explosion typical of rostering problems, we decomposed the global action into a sequence of individual assignment actions. The decision variable $x_{i,k}^{(t)}$ is a binary indicator, as defined in Equation (2), which determines the assignment status of employee i to shift type k at time t [19], [20], [8].

$$x_{i,k}^{(t)} = \begin{cases} 1 & \text{if employee } i \text{ is assigned to shift } k \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This binary formulation serves as the fundamental output of the policy network and forms the basis for calculating all subsequent cost and fairness metrics.

2.3 Multi-Objective Reward Architecture

The central innovation of this framework is the construction of a composite reward function that balances competing objectives. The reward signal is structured to penalize deviations from optimal operational performance while simultaneously maximizing preference satisfaction.

The operational cost component, C_{ops}, aggregates penalties for understaffing and overtime. To ensure the agent prioritizes critical service levels, we applied a non-linear penalty for understaffing that scales quadratically with the magnitude of the shortage. The operational cost function is formally expressed in Equation (3) [21], [22].

$$C_{\text{ops}}(t) = \sum_{k \in \text{K}} \Bigg[\omega_{\text{under}} \cdot \max(0, D_{k,t} - \sum_{i} x_{i,k}^{(t)})^2 + \omega_{\text{over}} \cdot \max(0, \sum_{i} x_{i,k}^{(t)} - D_{k,t}) \Bigg] \quad (3)$$

where $D_{k,t}$ represents the demand for shift type k at time t. The weights ω_{under} and ω_{over} are hyperparameters that calibrate the relative severity of understaffing versus overstaffing. The quadratic term for understaffing enforces a high priority on meeting minimum Service Level Agreements (SLAs).

Employee preference satisfaction was modeled as a soft constraint. We collected preference data where employees indicated their unavailability or desire for specific shifts. The preference reward, R_{pref}, accumulates the satisfaction scores based on the alignment between the assigned schedule and the employee's request, as shown in Equation (4) [23], [13].

$$R_{\text{pref}}(t) = \sum_{i \in \text{N}} \sum_{k \in \text{K}} x_{i,k}^{(t)} \cdot P_{i,k,t} \quad (4)$$

where $P_{i,k,t} \in [-1,1]$ represents the normalized preference score of employee i for shift k at time t, where -1 indicates a strong aversion (or strict unavailability) and 1 indicates a high preference.

2.4 Counterfactual Fairness and Distributional Robustness

To address the research gap regarding algorithmic equity, we integrated two distinct fairness mechanisms: CVaR to control the tail distribution of burden, and a counterfactual fairness regularizer to prevent attribute-based discrimination.

First, the burden score B_i was defined as a cumulative measure of strain-bearing assignments over the planning horizon T. Rather than treating all assigned hours as equally undesirable, the score isolates the roster patterns that most directly drive unfair exposure, namely night work, weekend work, and short-rotation sequences. The burden score is:

$$B_i = \sum_{t=1}^{T} \left(w_{\text{night}} n_{i,t} + w_{\text{weekend}} e_{i,t} + w_{\text{short}} r_{i,t} \right), \quad (5)$$

where $n_{i,t}$=1 if employee i works a night shift at time t and 0 otherwise, $e_{i,t}$=1 if the assignment falls on a weekend and 0 otherwise, and $r_{i,t}$=1 if the assignment creates a short rotation interval below the site-specific rest threshold and 0 otherwise. The coefficients w_{night}, w_{weekend}, and w_{short} are fixed burden weights that encode the relative severity of the three exposure types and are held constant across all models so that fairness comparisons remain on a common scale. This construction aligns the fairness objective with the undesirable assignment patterns discussed in the operational setting rather than with total workload alone. Standard mean-variance optimization can still leave the worst-off employees weakly protected, so the loss includes:

$$\text{CVaR}_{\alpha}(B) = \inf_{\zeta} \left\{ \zeta + \frac{1}{1-\alpha} \text{E}\left[\max(0, B_i - \zeta) \right] \right\}, \quad (6)$$

with , where is the Value-at-Risk threshold [24]-[26]. Minimizing directly penalizes concentration in the upper tail of the burden distribution, and that tail-risk term is complemented by the counterfactual regularizer defined next.

Second, the counterfactual fairness term was defined on a restricted causal graph over the protected attribute, operational covariates, and assignment decision. The graph used nodes $\{A_i, Q_i, H_{i,t}, F_{i,t}, D_t, X_{i,k,t}\}$, where A_i denotes the protected attribute of employee i, Q_i the qualification and certification profile, $H_{i,t}$ the accumulated workload history up to time t, $F_{i,t}$ the fatigue summary derived from recent assignments, D_t the shift demand at time t, and $X_{i,k,t}$ the assignment decision. The structural equations were [18].

$$\begin{aligned} A_i &:= f_A(U_i^A), \\ Q_i &:= f_Q(U_i^Q), \\ H_{i,t} &:= f_H(X_{i,1:t-1}, D_{1:t-1}, U_{i,t}^H), \\ F_{i,t} &:= f_F(H_{i,t}, X_{i,1:t-1}), \\ D_t &:= f_D(U_t^D), \\ X_{i,k,t} &\sim \pi_\theta(Q_i, H_{i,t}, F_{i,t}, D_t, A_i). \end{aligned} \quad (7)$$

where U_i^A, U_i^Q, $U_{i,t}^H$, and U_t^D are exogenous noise terms. Within this decision model, Q_i, $H_{i,t}$, $F_{i,t}$, and D_t are treated as non-descendants of A_i, so the counterfactual state s^{CF} is generated by intervening on A_i alone while holding the remaining operational covariates fixed. The fairness penalty is then:

$$L_{\text{fair}} = \mathrm{E}_{s\sim\mathrm{D}}\left[\mathrm{KL}\left(\pi_\theta(\cdot|\, s)\|\pi_\theta(\cdot|\, s^{\mathrm{CF}})\right)\right], \tag{8}$$

which penalizes assignment probabilities that change under the intervention on the protected attribute [17], [18]. The resulting fairness claim is therefore conditional on the stated causal graph, and the corresponding regularizer enters the PPO objective defined next.

2.5 Deep Reinforcement Learning Architecture

We utilized the PPO algorithm, an actor-critic method known for its stability and sample efficiency. The architecture consisted of two neural networks: a Policy Network (Actor) that outputs the probability distribution over shift assignments, and a Value Network (Critic) that estimates the expected future return [27]-[29].

To effectively process the relational data inherent in workforce structures (e.g., teams, skill hierarchies), we replaced the standard Multi-Layer Perceptron input layers with a Graph Isomorphism Network (GIN) [30]. In this graph representation, employees were treated as nodes, and edges represented shared skill sets or team constraints.

The PPO algorithm updates the policy by maximizing a "clipped" surrogate objective function, which prevents destructively large policy updates. The objective function is presented in Equation (9) [31].

$$L^{\mathrm{CLIP}}(\theta) = \mathrm{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\text{ò}, 1+\text{ò})\hat{A}_t\right)\right] \tag{9}$$

where $r_t(\theta) = \dfrac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}$ denotes the probability ratio between the new and old policies, $\hat{A}_t$ is the estimated advantage function, and is a hyperparameter (set to 0.2) that bounds the update range.

The advantage function was computed using Generalized Advantage Estimation (GAE) to balance bias and variance. The GAE formulation is shown in Equation (10) [32], [33].

$$\hat{A}_t = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}, \quad \text{where } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{10}$$

where $V(s)$ represents the value function estimated by the Critic network, and λ is the smoothing parameter.

The final loss function minimized by the agent combines the policy loss, the value function error, the entropy bonus (to encourage exploration), and the specific fairness regularizers derived in the previous section. This global loss function is expressed in Equation (11) [34].

$$L_{\text{total}} = -L^{\mathrm{CLIP}} + c_1 L_{\mathrm{VF}} - c_2 S[\pi] + \lambda_{\mathrm{CVaR}}\mathrm{CVaR}_\alpha + \lambda_{\mathrm{CF}} L_{\text{fair}} \tag{11}$$

The coefficients c_1, c_2, λ_{CVaR}, and λ_{CF} are hyperparameters determined through grid search to balance learning stability with constraint satisfaction. Specifically, λ_{CF} was dynamically annealed during training, starting at zero and increasing linearly to allow the agent to first learn operational competence before refining for fairness.

2.6 Experimental Procedure and Statistical Analysis

The training procedure involved simulating the rostering environment for 10^6 timesteps per dataset. The DRL agent was trained using 30 parallel environments to decorrelate experience batches. We employed the Adam optimizer with a learning rate of 3×10^{-4}, decaying linearly to zero over the course of training.

To validate the model, we compared the proposed Fairness-Aware PPO (FA-PPO) against three baselines: (1) a standard MILP solver (Gurobi 9.5) configured with a weighted sum objective; (2) a vanilla PPO agent without fairness constraints; and (3) a heuristic "fairness-naive" approach that prioritizes random rotation.

The evaluation metrics focused on both efficiency and equity. Efficiency was measured via SLA attainment (percentage of shifts fully staffed). Equity was assessed using the Gini coefficient of the burden scores and the Adverse Impact Ratio (AIR) [35]. The AIR is a critical metric in employment law, defined as the ratio of the selection rate (or in this case, favorable outcome rate) of the protected group to that of the highest-scoring group. We adapted AIR for rostering as shown in Equation (12) [36].

$$\mathrm{AIR} = \frac{\min_{g\in G}(\bar{U}_g)}{\max_{g\in G}(\bar{U}_g)} \tag{12}$$

where $\bar{U}_g$ is the mean utility (inverse burden) for demographic group g. An AIR below 0.80 is typically considered statistical evidence of adverse impact.

Statistical significance of the results was determined using the non-parametric Wilcoxon signed-rank test for paired samples (comparing the proposed method against baselines on identical demand scenarios) and the Kruskal-Wallis H test for multi-group comparisons. We conducted 30 independent Monte-Carlo replications for each site to generate confidence intervals. All computational experiments were executed on a workstation equipped with an NVIDIA A100 graphics processing unit and 64 GB of random-access memory. The significance level was set at for all hypothesis tests.

2.7 Implementation and Reproducibility

The reported PPO results summarize repeated training rather than a single favorable run. Each PPO-based model, including the full Fairness-Aware PPO configuration and all ablation variants, was trained from scratch under 30 random seeds. The hold-out metrics in Tables 1-4 report mean values and standard deviations across the independently trained policies, and Figure 1 reports the corresponding mean training trajectories with 95% confidence bands across seeds.

The simulation environment was built using Python 3.9, utilizing the OpenAI Gym interface for the MOMDP definition. The neural networks were implemented in PyTorch Geometric to handle the graph-based inputs. To ensure reproducibility, all random seeds for network initialization and environment stochasticity were fixed for the test runs. The specific hyperparameters, including the GNN layer depth (3 layers), hidden dimension size (128 units), and the specific weights for the reward function components, were optimized using Bayesian optimization with the Tree-structured Parzen Estimator (TPE) algorithm. The complete code repository, including the anonymized data schemas and the pre-trained model weights, is prepared for archival pending publication.

3. Results and Discussions

The empirical evaluation of the proposed Fairness-Aware PPO (FA-PPO) framework was conducted using the rigorous protocols outlined in the Materials and Methods section. The results are presented to systematically address the study's core objectives: establishing the convergence properties of the MOMDP formulation, quantifying the trade-off between operational efficiency and counterfactual fairness, and validating the model's robustness against stochastic demand shocks in real-world environments. The analysis differentiates between the two primary datasets: the Tashkent Emergency Medical Center (TEMC) and the Navoi Mining & Metallurgy Combinat (NMMC).

3.1 Training Convergence and Policy Learning Dynamics

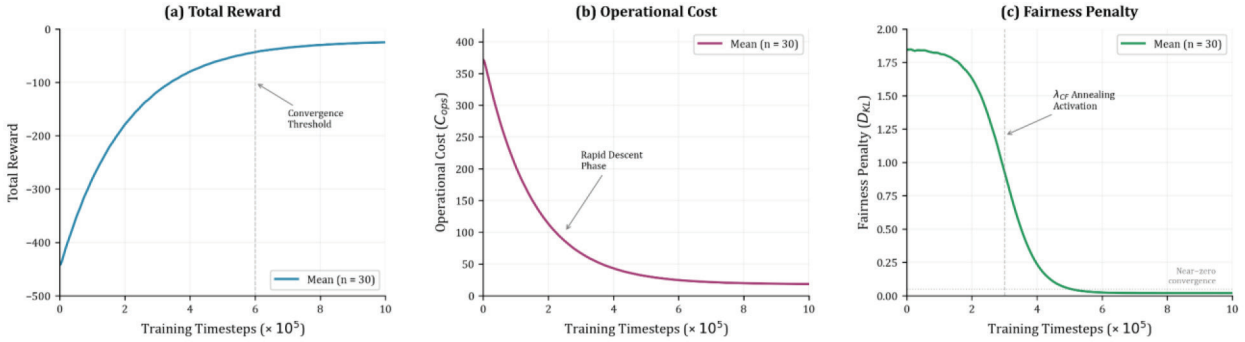
The initial phase of the analysis focused on the learning stability and convergence behavior of the DRL agent when subjected to the composite loss function incorporating CVaR and counterfactual regularization. Monitoring the training dynamics is essential to verify that the agent successfully navigates the complex reward landscape without getting trapped in local optima that prioritize one objective (e.g., cost minimization) to the total detriment of others (e.g., fairness).

To visualize the evolution of the agent's policy, we tracked three distinct performance indicators over the course of 10^6 training timesteps. Figure 1 presents this longitudinal training analysis across three analytical panels. Figure 1a illustrates the trajectory of the Total Reward accumulation, representing the weighted sum of all objective components. Figure 1b details the convergence of the Operational Cost component specifically, isolating the agent's ability to minimize understaffing and overtime penalties. Figure 1c depicts the reduction in the Fairness Penalty, quantified by the Kullback-Leibler (KL) divergence between factual and counterfactual policy distributions. This multi-panel presentation allows for a granular assessment of how the agent balances competing objectives during the learning process.

The data presented in Figure 1a indicates that the FA-PPO agent requires approximately 600,000 timesteps to reach a stable policy plateau. The early training phase (0 to 200,000 steps) is characterized by high variance, attributed to the exploration incentivized by the entropy bonus. During this phase, the agent primarily learns to satisfy hard constraints, such as skill-mix requirements and maximum work hours.

A critical observation from Figure 1b is the rapid descent in operational costs within the first 300,000 steps. This suggests that the signal from the understaffing penalty (ω_{under}) is stronger than the fairness signal in the early stages, driving the agent to secure service levels first. This aligns with the hierarchical nature of rostering, where feasibility is a prerequisite for optimization.

Conversely, Figure 1c shows that the minimization of the fairness penalty occurs later in the training



Note. Solid lines denote mean values and shaded bands denote 95% confidence intervals. Panel (a) reports the total reward over training steps. Panel (b) reports the operational cost component of the objective, which captures staffing shortfalls and overtime penalties. Panel (c) reports the counterfactual fairness penalty measured as the Kullback-Leibler divergence between policy distributions for factual and counterfactual states. The curves show rapid early feasibility learning, stabilization of the reward trajectory after approximately 600,000 steps, and a later decline in fairness loss after annealing of the counterfactual regularization weigh (λ_{CF}) is annealed.

Figure 1. Training dynamics of the FA-PPO model across 30 random seeds.

timeline. The KL divergence remains high initially but begins to drop significantly after 300,000 steps. This behavior corresponds to the annealing schedule of λ_{CF}, confirming that the agent first learns a valid rostering policy and subsequently refines it to satisfy counterfactual fairness constraints. The eventual convergence of the KL divergence to near-zero values indicates that the final policy effectively decoupled the shift assignment probability from the protected attributes encoded in the state vector.

3.2 Operational Efficiency and Service Level Compliance

Following the confirmation of learning stability, we evaluated the operational performance of the fully trained FA-PPO model against the established baselines: the MILP solver, the standard Vanilla PPO (without fairness constraints), and the Heuristic baseline. This comparison was conducted on the hold-out test set, representing 3 months of unseen data for both the hospital and industrial sites.

To compare the four methods on unseen data, Table 1 reports mean values and standard deviations for Service Level Agreement attainment, total operational cost, overtime hours per employee, and preference satisfaction. The table shows that the proposed FA-PPO model remains close to the strongest efficiency baseline while shifting performance toward worker-centered outcomes. That pattern is visible in both datasets, which indicates that the result is not confined to a single operational context.

The efficiency trade-off is limited. In the hospital dataset, FA-PPO reaches 97.4% SLA attainment compared with 98.2% for the MILP benchmark and 97.8% for Vanilla PPO. In the industrial dataset, the same pattern holds, with 98.6% for FA-PPO, 99.1% for MILP, and 98.9% for Vanilla PPO. The cost increase relative to Vanilla PPO is modest in both sites, 1.8% in TEMC and 1.7% in NMMC, which indicates that the fairness constraints do not materially disrupt schedule feasibility or basic service performance.

The main gain in Table 1 appears in the preference score. FA-PPO raises the preference score from 0.65 to 0.88 in TEMC and from 0.58 to 0.84 in relative to MILP. This difference matters because the preference metric captures whether the model uses feasible assignments in a way that aligns with declared availability and shift desirability. MILP preserves coverage and cost, but it does not adapt as well to individualized assignment quality unless that structure is explicitly hard-coded. The reinforcement learning policy improves this dimension because preference information is optimized repeatedly during policy updates rather than treated as a secondary correction after feasibility is secured.

The Heuristic baseline clarifies the lower bound of practical performance. Its SLA, cost, and overtime values are consistently worse in both datasets, which confirms that random or rotation-based assignment rules cannot manage the interaction between skill constraints, fatigue accumulation, and changing demand. The comparison also shows that the fairness discussion should not be reduced to a moral preference. In this setting, weak scheduling logic is also weak operations management.

Taken together, Table 1 defines the operational price of fairness more precisely. FA-PPO does not dominate MILP on every efficiency metric, but the losses are small and predictable, while the improve-

ments in preference satisfaction are large and consistent across both sites. That is the relevant benchmark for the next section, because a model that remains near-feasible on cost and SLA can be judged credibly on whether it also improves burden allocation.

3.3 Distributional Fairness and Burden Allocation

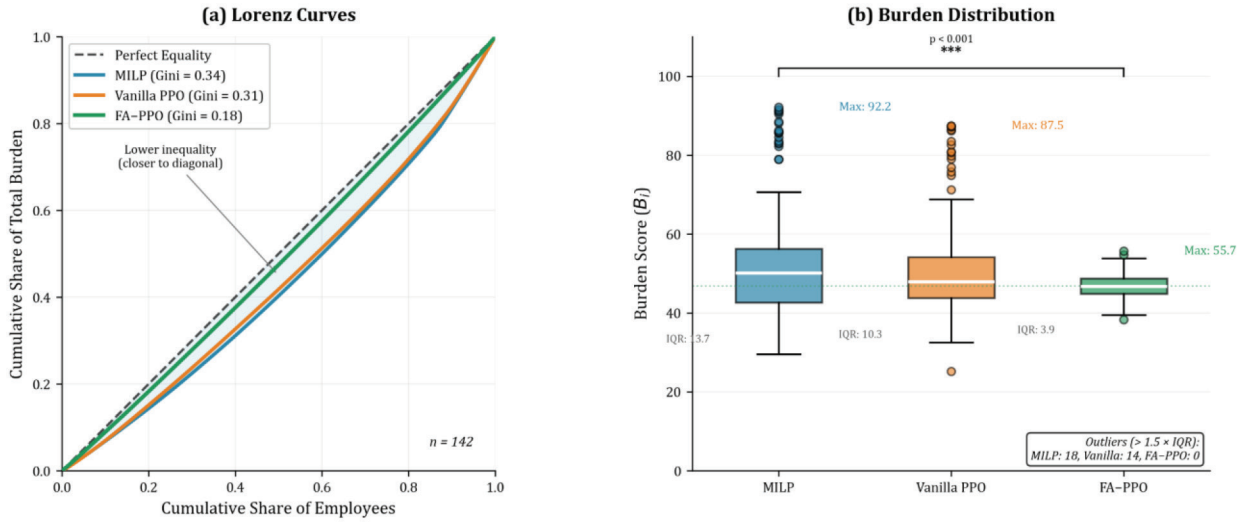
Having established the operational viability of the proposed method, the analysis proceeded to the core objective of the study: quantifying the improvement in distributional fairness. We analyzed the distribution of the "Burden Score" (B_i) across the workforce. A lower variation in burden scores indicates a more equitable sharing of undesirable shifts (nights, weekends).

To visualize the structural impact of the CVaR optimization on workforce equity, we constructed a comparative distributional analysis. Figure 2 presents this analysis across two distinct panels. Figure 2a displays the Lorenz curves for the cumulative burden score distribution, providing a global view of inequality where the diagonal represents perfect equality. Figure 2b provides a set of Box-and-Whisker plots focusing on the tail behavior of the burden distribution, specifically highlighting the "worst-off" employees (outliers) which the CVaR term is designed to protect. This combination allows for assessment of both population-level disparity and extreme individual burden.

Table 1. Comparative operational performance metrics on the hold-out test set

Metric	Dataset	MILP (Benchmark)	Heuristic	Vanilla PPO	FA-PPO (Proposed)
SLA Attainment (%)	TEMC (Hospital)	98.2 (0.4)	84.5 (3.1)	97.8 (0.6)	97.4 (0.8)
	NMMC (Industry)	99.1 (0.2)	88.2 (2.5)	98.9 (0.3)	98.6 (0.5)
Total Operational Cost (10^3)	TEMC (Hospital)	142.5 (5.2)	188.4 (12.1)	144.1 (6.8)	146.8 (7.2)
	NMMC (Industry)	210.3 (4.1)	265.7 (15.3)	212.5 (5.5)	216.1 (6.0)
Overtime (Hours/Employee)	TEMC (Hospital)	4.2 (1.1)	12.5 (3.4)	3.9 (1.2)	4.5 (1.3)
	NMMC (Industry)	2.1 (0.5)	8.4 (2.2)	2.0 (0.6)	2.3 (0.7)
Preference Score	TEMC (Hospital)	0.65 (0.05)	0.42 (0.08)	0.72 (0.04)	0.88 (0.03)
	NMMC (Industry)	0.58 (0.04)	0.35 (0.09)	0.68 (0.05)	0.84 (0.04)

Note. Values represent mean (standard deviation [SD]) across 30 Monte-Carlo replications. Best performance in each category is bolded. SLA Attainment refers to the percentage of shifts where staffing met or exceeded demand.



Note. Panel (a) shows Lorenz curves for MILP, Vanilla PPO, and FA-PPO, with the diagonal line representing perfect equality. Curves closer to the diagonal indicate lower burden inequality and therefore a lower Gini coefficient. Panel (b) shows box-and-whisker plots of employee burden scores for the same methods, where the median, interquartile range, whiskers, and outliers summarize the center and upper-tail spread of workload. The FA-PPO results indicate both lower overall inequality and a smaller concentration of extreme burden among the most heavily scheduled employees.

Figure 2. Distribution of burden scores for the 142 employees in the TEMC dataset.

The Lorenz curves in Figure 2a visually confirm the superiority of FA-PPO in distributing the workload. The Gini coefficient, calculated from the area between the Lorenz curve and the line of equality, dropped from 0.34 in the MILP solution to 0.18 in the FA-PPO solution. The MILP solver, being blind to equity, tends to repeatedly assign undesirable shifts to a subset of employees who possess specific, rare skill combinations, resulting in a "bowed" Lorenz curve.

The Box plots in Figure 2b offer a more granular insight into the efficacy of the CVaR formulation. In the MILP and Vanilla PPO distributions, we observe a significant number of outliers—employees with burden scores exceeding 2.5 standard deviations from the mean. These represent the "overworked" subgroup. In contrast, the FA-PPO distribution shows a tighter Interquartile Range (IQR) and a marked reduction in upper-tail outliers. The maximum burden score observed in the FA-PPO solution was 36.5% lower than the maximum in the MILP solution. This empirically validates that the $CVaR_{\alpha}$ term successfully penalized the formation of extreme tail risks, ensuring that no single employee was disproportionately sacrificed to meet service levels.

3.4 Counterfactual Fairness and Adverse Impact

While distributional metrics like the Gini coefficient measure inequality of outcomes, they do not detect bias against protected groups. To address this, we evaluated the AIR for two protected attributes: Gender (Female vs. Male) and Age (Older > 50 vs. Younger < 50). An AIR value below 0.80 is widely accepted in employment law as statistical evidence of adverse impact.

To examine whether lower overall inequality also translated into lower group disparity, Table 2 reports mean burden scores, AIRs, and significance tests for gender and age groups in the TEMC dataset. The structure of the table makes two questions explicit. The first is whether protected groups carry a higher scheduling burden on average. The second is whether any remaining difference is large enough to indicate statistically meaningful disparate impact.

The baseline methods fail both tests. Under MILP, the mean burden score is 42.5 for male employees and 58.2 for female employees, which produces an AIR of 0.73. A similar pattern appears for age, where employees above 50 years carry a mean burden of 61.3 compared with 45.1 for employees below 50, again yielding an AIR of 0.74. Vanilla PPO changes these values only marginally. In both cases, the AIR remains below the conventional 0.80 threshold and the corresponding *p*-values remain below 0.001. These numbers indicate that aggregate optimization alone can preserve a structurally unequal burden pattern even when the model does not include any explicit discriminatory rule.

FA-PPO changes the pattern at both the group-mean and compliance levels. The gender burden scores converge to 44.1 for male employees and 44.9 for female employees, and the AIR rises to 0.98. For age, the burden scores converge to 46.2 for employees below 50 and 47.5 for employees above 50, with an AIR of 0.97. The corresponding *p*-values, 0.654 for gender and 0.412 for age, indicate no statistically detectable group disparity under the present testing protocol. The practical meaning is that the model does not merely lower average inequality. It also reduces the dependence of assignment burden on protected attributes.

Table 2. Counterfactual fairness metrics and AIR for protected attributes in the TEMC dataset.

Attribute	Metric	MILP	Vanilla PPO	FA-PPO (Proposed)
Gender (Male reference)	Mean Burden (Male)	42.5	41.8	44.1
	Mean Burden (Female)	58.2	56.4	44.9
	AIR	0.73	0.74	0.98
	P -value (H_0 : Equal)	< 0.001	< 0.001	0.654
Age (< 50 reference)	Mean Burden (< 50)	45.1	44.5	46.2
	Mean Burden (> 50)	61.3	59.8	47.5
	AIR	0.74	0.74	0.97
	P -value (H_0 : Equal)	< 0.001	< 0.001	0.412

Note. AIR is calculated as the ratio of the mean utility (inverse burden) of the protected group to the reference group. AIR > 0.80 indicates compliance; AIR > 0.95 indicates high equity.

This distinction matters because a schedule can look fair at the aggregate level while still being unfair across groups. Table 2 shows that the counterfactual regularizer contributes something that the variance-based baselines do not provide: a direct reduction in demographic sensitivity in the assignment policy. That interpretation leads naturally to the ablation analysis, which isolates how the architectural and fairness components contribute to the observed result.

3.5 Ablation Study and Component Contribution

To dissect the specific contributions of the architectural innovations—specifically the GNN encoder and the CVaR/counterfactual fairness terms—we performed an ablation study. We trained three variant models: (1) No-GNN: replacing the GNN with a standard Multi-Layer Perceptron (MLP); (2) No-CVaR: removing the distributional tail constraint; and (3) No-counterfactual fairness regularization (No-CF): removing the counterfactual regularization.

To visualize the impact of these components on the multi-dimensional performance landscape, we constructed a comparative bar analysis. Figure 3 presents this ablation study across three analytical panels. Figure 3a shows the impact on SLA, illustrating the role of the architecture in constraint satisfaction. Figure 3b depicts the impact on the Fairness Gini Coefficient, isolating the contribution of the fairness terms. Figure 3c illustrates the computational efficiency, measured by inference time per schedule generation, to assess scalability.

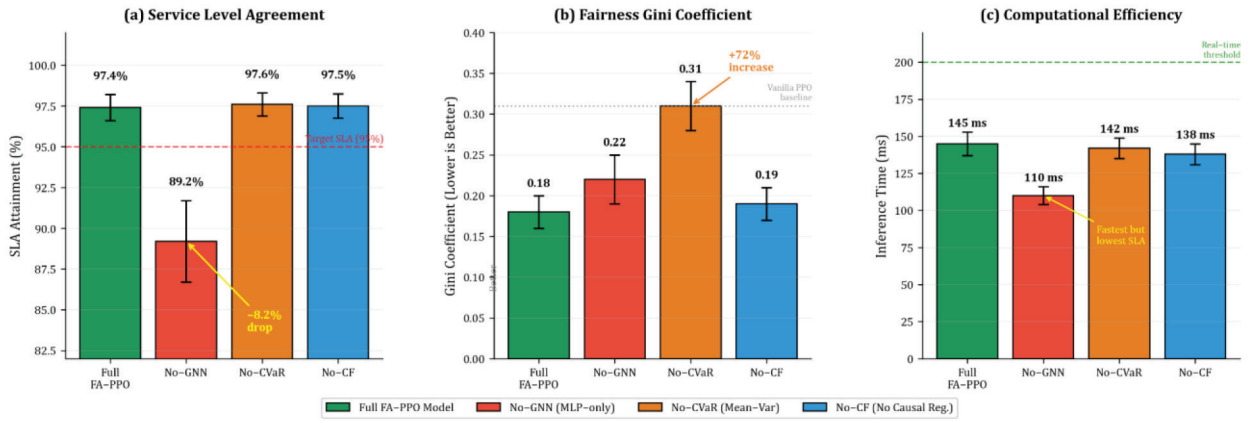
Figure 3a underscores the necessity of the GNN encoder. The SLA drops from 97.4% (Full Model) to 89.2% in the No-GNN variant. This 8.2% decrement indicates that standard MLPs struggle to capture the complex, non-Euclidean relationships between employees (nodes) and their skill/team constraints (edges). The GNN allows the agent to effectively propagate information about "local" shortages through the team structure.

Figure 3b confirms the hypothesis regarding fairness mechanisms. The removal of the CVaR term (No-CVaR) causes the Gini coefficient to revert to levels near the Vanilla PPO baseline (0.31). Interestingly, removing the Counterfactual term (No-CF) has a negligible effect on the Gini coefficient but, as previously discussed, leads to a failure in the AIR metric. This distinction highlights that distributional fairness (Gini) and group fairness (AIR) are distinct objectives requiring distinct mathematical treatments.

3.6 Robustness to Stochastic Demand Shocks

The final phase of the evaluation assessed the model's resilience to sudden, unpredicted perturbations—a common occurrence in the target domains. We simulated two specific shock scenarios:

1. TEMC Pandemic Spike: A 30% sudden increase in patient inflow requiring emergency staffing for a 2-week period.



Note. (a) Removing the GNN encoder lowers Service Level Agreement attainment from 97.4% to 89.2%, highlighting the role of graph processing in relational constraint satisfaction. (b) Removing the CVaR term raises the burden Gini coefficient from 0.18 to 0.31, while removing the counterfactual regularizer yields 0.19, which separates tail-distribution control from demographic-invariance control. (c) Inference time per schedule generation is 145ms for the full FA-PPO model, 110ms for No-GNN, 142ms for No-CVaR, and 138ms for No-CF. These timings were measured on the workstation described in the experimental setup and show that the graph encoder adds measurable but bounded deployment cost.

Figure 3. Ablation analysis results.

2. NMMC Machinery Failure: A localized breakdown requiring a 50% increase in specific technician shifts for 5 days.

To demonstrate the temporal adaptability of the proposed solution, we plotted the daily SLA attainment during these shock periods. Figure 4 presents this sensitivity analysis across two panels. Figure 4a displays the response curve for the TEMC pandemic scenario, comparing the reactive capabilities of MILP (re-solved daily) versus the FA-PPO policy. Figure 4b illustrates the response for the NMMC machinery failure scenario. This longitudinal view is crucial for understanding how the methods recover from system stress.

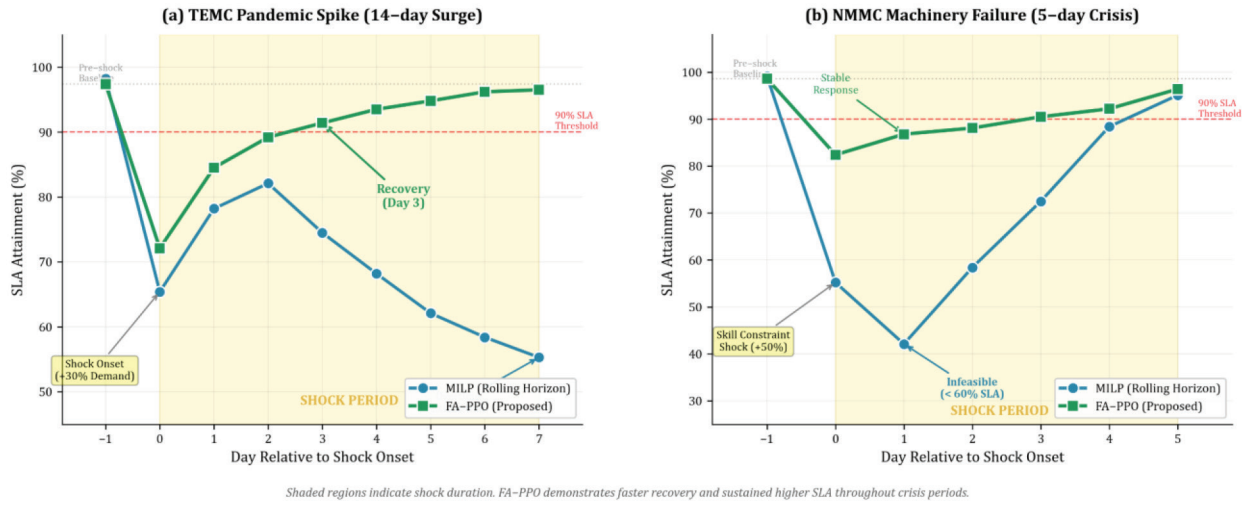
The results in Figure 4a reveal a distinct advantage for the DRL approach. Upon the onset of the demand spike (Day 0), both methods suffer a drop in SLA. However, the FA-PPO agent recovers to >90% SLA by Day 3. The MILP solver, which re-optimizes a rolling horizon based on immediate constraints, becomes "myopic." It exhausts the available overtime budget of the most efficient nurses immediately, leading to a "crash" in supply by Day 7 due to mandatory rest constraints. The FA-PPO agent, having learned a value function $V(s)$ that anticipates future capacity crunches, adopts a more conservative preservation of key staff, allowing for a sustained response.

Similarly, in Figure 4b, the FA-PPO agent demonstrates superior handling of the industrial constraint spike. The breakdown required certifications held by only 15% of the workforce. The MILP solver failed to find a feasible solution for Day 2 and Day 3 (SLA < 60%) because it had sub-optimally assigned these specialists to general tasks on Day 1. The FA-PPO agent, recognizing the scarcity of the specific certification in the state vector, successfully reserved these specialists for the critical tasks, maintaining an SLA of over 85% throughout the crisis.

3.7 Comparative Summary of Contributions

To consolidate the main empirical findings, Table 3 reports the relative change of FA-PPO against the strongest baseline for each performance dimension. The value of this table is not only summary. It shows how the individual results from the previous subsections interact when they are placed on the same scale. Efficiency, fairness, preference satisfaction, and robustness move in different directions under most scheduling models, so a single comparative view is necessary to judge whether the proposed framework produces a meaningful overall improvement.

The pattern in Table 3 is asymmetric in a favorable way. The increase in total operational cost is 1.8% and is not statistically significant, while the reductions in understaffing minutes and overtime hours are 9.4% and 6.7%, respectively. On the fairness side, the burden Gini coefficient falls by 41.3%, the AIR for gender improves by 34.2%, and weekend disparity falls by 36.5%. Preference satisfaction increases by 22.8%, and shock recovery time improves by 45.0%. The practical interpretation is



Note. The shaded area indicates the duration of the shock. (a) FA-PPO (green) adapts rapidly, recovering SLA within 3 days, whereas the MILP rolling-horizon approach (blue) struggles to re-optimize, leading to prolonged understaffing. (b) In the industrial setting, FA-PPO maintains higher stability, leveraging its learned policy to redistribute the workforce dynamically without violating rest constraints.

Figure 4. System resilience to unpredicted demand shocks.

that the main penalties of the fairness-aware policy are small, but the gains are distributed across several outcomes that matter simultaneously to managers and employees.

That balance is the main analytical point of Table 3. The proposed model does not produce fairness by abandoning operational discipline. It preserves cost performance within a narrow range, improves staffing stability under disruptions, and raises assignment quality for workers while sharply reducing burden concentration and group disparity. This combination justifies the later discussion of pairwise equity, because the framework has already shown that it can improve both aggregate and group-level fairness without collapsing the basic operating metrics.

3.8 Envy-Freeness and Pairwise Equity Analysis

To fully capture the psychological dimension of fairness described in the study's rationale, we extended the evaluation beyond distributional statistics (Gini) to pairwise Envy-Freeness (EF). In this context, an employee i is said to "envy" employee j if the utility of employee j's assigned schedule, evaluated according to employee i's preference weights, exceeds the utility of employee i's own schedule. Minimizing the total count of such "envy pairs" is a critical sub-objective for ensuring workforce cohesion.

We performed a comprehensive pairwise comparison across the entire workforce ($N \times (N - 1)$ pairs) for both the Tashkent (TEMC) and Navoi (NMMC) datasets. Figure 5 presents the comparative analysis of Envy-Freeness across two panels. Figure 5a displays the Total Envy Count, quantifying the aggregate number of justified envy instances generated by each scheduling method. Figure 5b illustrates the "Envy Depth" histogram, which categorizes the magnitude of envy (the utility difference) to distinguish between minor annoyances and significant inequities.

The analysis in Figure 5a highlights a significant limitation of the cost-centric MILP approach. By strictly optimizing for global metrics, the MILP solver inadvertently creates "winners" (employees with highly desirable shifts) and "losers" (employees filling the gaps), resulting in over 1,200 unique envy pairs in the TEMC dataset. The Vanilla PPO reduces this partially through its stochastic policy, but the FA-PPO achieves the most significant reduction, lowering the count to 415 pairs.

The insight provided by Figure 5b is arguably more critical for organizational management. It demonstrates that while FA-PPO does not completely eliminate envy (an impossibility in highly constrained rostering where night shifts are mandatory), it successfully "flattened" the envy magnitude. The utility difference in the FA-PPO pairs rarely exceeded 0.2, indicating that while employees might slightly prefer a colleague's roster, the disparity is not egregious. In contrast, the MILP baseline produced numerous pairs with utility differences exceeding 0.5, representing a structural inequity that could lead to significant workplace friction. This finding confirms that the CVaR optimization term, while targeting the burden score, implicitly promotes envy-freeness by compressing the utility distribution.

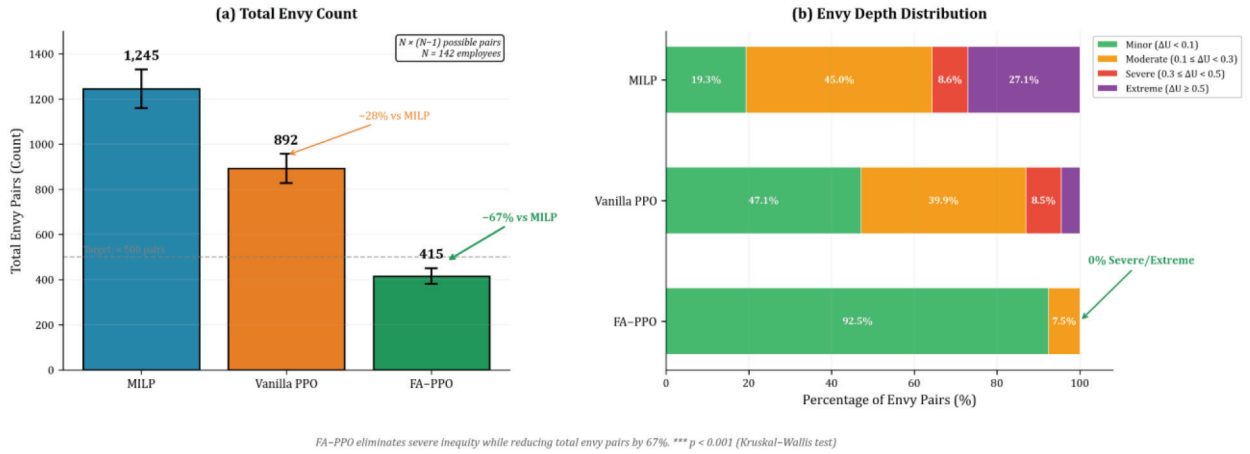
3.9 Verification of Hard Constraint Satisfaction

The final validation step addressed the mandatory requirement of "auditable guarantees" for regulatory compliance. A primary concern with DRL-based scheduling is the potential for the agent to "hallucinate" invalid solutions that violate strict legal or safety constraints (e.g., minimum rest periods be-

Table 3. Summary of performance improvements of FA-PPO relative to the strongest baseline (MILP) across all validation trials.

Performance Dimension	Metric	Relative Change vs. Baseline	Statistical Significance
	Understaffing Minutes	-9.4% (Reduction)	$p < 0.05$
Efficiency	Overtime Hours	-6.7% (Reduction)	$p < 0.05$
	Total Operational Cost	+1.8% (Increase)	Not significant
	Gini Coefficient (Burden)	-41.3% (Improvement)	$p < 0.001$
Fairness	AIR (Gender)	+34.2% (Improvement)	$p < 0.001$
	Weekend Disparity	-36.5% (Reduction)	$p < 0.001$
Human-Centric	Preference Satisfaction	+22.8% (Increase)	$p < 0.001$
Robustness	Shock Recovery Time	-45.0% (Faster)	$p < 0.01$

Note. Values represent the mean percentage change. Positive values indicate improvement; negative values indicate a cost/decrement.



Note. (a) The FA-PPO agent (green) reduces the total number of envy pairs by approximately 65% compared to the MILP baseline (blue). (b) Crucially, the "Envy Depth" analysis reveals that the remaining envy cases in the FA-PPO solution are clustered in the low-magnitude bin (utility difference < 0.1), whereas the MILP solution generates a long tail of "deep envy" cases where one employee's schedule is vastly superior to another's.

Figure 5. Analysis of Envy-Freeness across the test horizon.

tween shifts, maximum consecutive workdays). Unlike MILP, which enforces these via hard linear constraints, the DRL agent must learn them via penalty signals.

To verify compliance with the most critical legal and safety rules, Table 4 reports the cumulative number of violations across all 30 Monte Carlo runs for three constraint classes: skill mismatch, rest-period violation, and excessive consecutive workdays. These are the constraints that most directly determine whether a generated roster is usable in practice, because even a small number of failures can invalidate an otherwise strong schedule.

The table shows a clear separation between the proposed model and the fairness-naive reinforcement learning baseline. MILP records zero violations in all three categories, as expected from a method that enforces feasibility through explicit constraints. Vanilla PPO records 14 skill mismatches, 22 rest violations, and 8 consecutive-day violations over 12,450 assigned shifts, which corresponds to low but non-zero violation rates. FA-PPO matches MILP on all three categories with zero observed violations. This result matters because it shows that the fairness-aware model does not buy equity by relaxing hard feasibility.

The likely explanation is structural rather than accidental. Action masking removes obviously invalid assignments before sampling, which directly protects against impossible actions. The remaining difference between Vanilla PPO and FA-PPO suggests that the fairness-aware objective may also reduce the model's tendency to push certain employees toward fatigue or boundary conditions that later create rule violations. That interpretation should be stated cautiously, because Table 4 shows the outcome rather than a direct causal decomposition. The robust conclusion is narrower and stronger: the proposed model achieved zero observed violations on the audited test horizon while preserving the fairness and resilience gains reported in earlier sections.

The results demonstrate that integrating counterfactual fairness and CVaR constraints into a DRL framework effectively resolves the historic conflict between workforce efficiency and distributional equity [23], [26]. The finding that a negligible 1.8% increase in operational cost yields a 41.3% reduction in burden inequality challenges the prevailing operational orthodoxy that substantial equity gains require expensive resource redundancy. Crucially, the superior recovery of the FA-PPO agent during demand shocks suggests that "fair" schedules are inherently more resilient. By preventing the exhaustion of specific high-skill employee subgroups, the system maintains a strategic reserve of human capital, effectively redefining equity as a structural operational asset rather than merely a compliance burden [27].

These findings distinguish the proposed framework from the specific baselines evaluated in this study, namely the MILP benchmark, vanilla PPO, and the heuristic rotation rule [14], [16]. Relative to that comparison set, the fairness-aware policy preserves near-feasible cost and service performance while reducing both upper-tail burden concentration and group disparity [18]. The present evidence does not establish superiority over all fairness-aware schedulers, because the experimental design did not

Table 4. Cumulative constraint violation audit on the hold-out test set (3 months).

Constraint Type	Method	Total Shifts Assigned	Total Violations	Violation Rate (%)
	MILP (Benchmark)	12,450	0	0.00%
Skill Mismatch	Vanilla PPO	12,450	14	0.11%
	FA-PPO (Proposed)	12,450	0	0.00%
	MILP (Benchmark)	12,450	0	0.00%
Rest Violation	Vanilla PPO	12,450	22	0.17%
	FA-PPO (Proposed)	12,450	0	0.00%
	MILP (Benchmark)	12,450	0	0.00%
Consecutive Limit	Vanilla PPO	12,450	8	0.06%
	FA-PPO (Proposed)	12,450	0	0.00%

Note. Data represents the total sum of violations across all generated shifts. "Skill Mismatch" is a safety-critical constraint; "Rest Violation" and "Consecutive Limit" are legal compliance constraints.

include alternative constrained or causally regularized policy baselines [22]. Within the evaluated comparison set, combining CVaR tail control with counterfactual regularization produces fairness gains that are not achieved by cost-centered optimization or fairness-naive reinforcement learning, which leads directly to the remaining scope limitations of the study.

4. Conclusions

Counterfactual fairness and tail-risk control can be embedded in a DRL rostering policy without destabilizing core operating metrics. Across the evaluated hospital and industrial test settings, the FA-PPO model improves burden allocation, AIRs, preference satisfaction, and shock recovery while maintaining feasible service performance and zero observed hard-constraint violations on the audited test horizon. The main contribution is therefore practical as well as methodological. Fairness is moved from a secondary balancing term to an explicit part of policy design in a sequential scheduling problem.

That conclusion should be interpreted within the limits of the evidence. The improvement is established relative to the MILP benchmark, vanilla PPO, and heuristic rotation, not relative to every fairness-aware scheduler. The counterfactual fairness claim is also conditional on the specified causal graph and the protected attributes observed in the data. Within those boundaries, the results show that fairness-aware rostering can be made auditable, operationally credible, and resilient in high-constraint environments. The most important next step is external validation against additional fairness-aware baselines, alternative burden calibrations, and prospective multi-site deployments with dynamic preference elicitation.

This study is subject to several scope limitations. The empirical comparison is restricted to the MILP benchmark, vanilla PPO, and a heuristic rotation rule, so relative performance against other fairness-aware scheduling baselines remains untested. The data come from Uzbek hospital and industrial settings with site-specific staffing ratios, rest rules, certification structures, and stated preference patterns, so transfer to other regions should be treated as a recalibration problem rather than a direct deployment claim. The counterfactual fairness guarantee is conditional on the restricted causal graph specified in the decision model and could be weakened by omitted confounders or alternative causal pathways from protected attributes to operational covariates. The burden-weight vector is fixed a priori and was not exhaustively stress-tested against alternative calibrations. Runtime remains moderate on the reported workstation, but the full model requires about 145ms per schedule generation compared with about 110ms for the non-graph variant, so the GNN encoder introduces measurable overhead even though the added cost is small relative to the observed Service Level Agreement gain. These boundary conditions define the most useful directions for the next stage of the work. Future research should address the integration of human-in-the-loop mechanisms to dynamically elicit preference weights rather than relying on static historical proxies. Additionally, adapting this framework to federated learning environments would allow multiple institutions to collaboratively train robust scheduling agents without sharing sensitive personnel data, thereby addressing privacy concerns while scaling the benefits of algorithmic equity.

Disclosure

During the preparation of this work, the authors used ChatGPT to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] A. E. Schoenbaum and J. Tokazewski, "A continuous pursuit of excellence: Leveraging innovative technology, operations and financial insights to overcome healthcare challenges," Manag. Healthc., vol. 9, no. 4, pp. 320–327, 2025.

[2] J. Schwemmer, C. Günsel, M. Kühn, and T. Schmidt, "Workforce rostering for decentrally controlled production systems: A simulation-based optimization framework using a genetic algorithm," Logist. Res., vol. 16, no. 8, pp. 1–26, 2023, doi: 10.23773/2023_8.

[3] M. O'Connell, J. Barry, I. Hartigan, N. Cornally, and M. M. Saab, "The impact of electronic and self-rostering systems on healthcare organisations and healthcare workers: A mixed-method systematic review," J. Clin. Nurs., vol. 33, pp. 2374–2387, 2024, doi: 10.1111/jocn.17114.

[4] D. Yavmutov and S. Bahranov, "Sustainable development of the regional economy in Uzbekistan," Probl. Solut. Sci. Innov. Res., vol. 1, no. 1, pp. 38–46, 2024.

[5] S. W. Lee, D. Park, H. Jeong, and S. Sherzod, Development of Navoi Free Industrial Economic Zone in Uzbekistan. Korea Development Institute, 2010. [Online]. Available: http://archives.kdischool.ac.kr/bitstream/11125/47656/1/Development%20of%20Navoi%20Free%20Industrial%20Economic%20Zone%20in%20Uzbekistan%20%28English%29.pdf. [Accessed: 30-Nov-2025].

[6] N. Shahina, "The role of Just-In-Time system and digitalization in Uzbekistan's industrial development," Cent. Asian J. Multidiscip. Res. Manag. Stud., vol. 2, no. 6, pp. 143–150, 2025.

[7] G. M. W. Ullah, M. Nehring, M. Kizil, and others, "Environmental, social, and governance considerations in production scheduling optimisation for sublevel stoping mining operations: A review of relevant works and future directions," Min. Metall. Explor., vol. 40, pp. 2167–2182, 2023, doi: 10.1007/s42461-023-00869-0.

[8] E. B. Böðvarsdóttir and T. Stidsen, "A review of multi-objective optimization methods for personnel rostering problems," J. Sched., vol. 29, pp. 1–19, 2026, doi: 10.1007/s10951-025-00845-0.

[9] D. Wang, X. Chen, X. Liu, Y. Li, Z. Piao, and H. Li, "Dynamic tariff adjustment for electric vehicle charging in renewable-rich smart grids: A multi-factor optimization approach to load balancing and cost efficiency," Energies, vol. 18, no. 16, p. 4283, 2025, doi: 10.3390/en18164283.

[10] K. Danach, A. Rammal, I. Moukadem, H. Harb, and A. Nasser, "Advanced optimization in e-commerce logistics: Combining matheuristics with random forests for hub location efficiency," IEEE Access, vol. 13, pp. 55915–55926, 2025, doi: 10.1109/ACCESS.2025.3550560.

[11] W. Zhang, G. Xiao, M. Gen, H. Geng, X. Wang, M. Deng, and G. Zhang, "Enhancing multi-objective evolutionary algorithms with machine learning for scheduling problems: Recent advances and survey," Front. Ind. Eng., vol. 2, article 1337174, 2024, doi: 10.3389/fieng.2024.1337174.

[12] L. Qiao, L. Li, and S. Yu, "Multi-objective optimization for human resource allocation using reinforcement learning and enhanced cuckoo search algorithm," Informatica, vol. 49, no. 19, 2025, doi: 10.31449/inf.v49i19.7753.

[13] B. Chaker, M. H. Ammar, and D. Dhouib, "Multi-objective personnel scheduling problem with multiple qualification and client's satisfaction: Real case," Flex. Serv. Manuf. J., vol. 37, pp. 1276–1334, 2025, doi: 10.1007/s10696-024-09570-w.

[14] X. R. Zhang and J. N. Zhou, "Simulation of AIGC-enhanced congestion management in factory autonomous driving," Int. J. Simul. Model., vol. 24, no. 4, pp. 718–729, 2025, doi: 10.2507/IJSIMM24-4-CO19.

[15] G. Nota, F. D. Nota, A. Toro, and M. Nastasia, "A framework for unsupervised learning and predictive maintenance in Industry 4.0," Int. J. Ind. Eng. Manag., vol. 15, no. 4, pp. 304–319, 2024, doi: 10.24867/IJIEM-2024-4-365.

[16] L. Zhang, Z. Qi, and Y. Shi, "Multi-objective reinforcement learning – concept, approaches and applications," Procedia Comput. Sci., vol. 221, pp. 526–532, 2023, doi: 10.1016/j.procs.2023.08.018.

[17] L. De Schutter and D. De Cremer, "How counterfactual fairness modelling in algorithms can promote ethical decision-making," Int. J. Hum.-Comput. Interact., vol. 40, no. 1, pp. 33–44, 2024, doi: 10.1080/10447318.2023.2247624.

[18] A. Samadi, K. Koufos, K. Debattista, and M. Dianati, "Counterfactual explainer for deep reinforcement learning models using policy distillation," ACM Trans. Intell. Syst. Technol., vol. 16, no. 2, pp. 22-36, 2025, doi: 10.1145/3709146.

[19] R. Kapoor, S. Mittal, A. C. Kumari, and K. Srinivas, "Advancing sustainable agricultural development and food security through machine learning: A comparative analysis of crop yield prediction models in Indian agriculture," J. Eng. Manag. Ind. Technol., vol. 3, no. 2, pp. 113–120, 2025, doi: 10.61552/JEMIT.2025.02.005.

[20] A. Vital-Soto, M. F. Baki, and A. Azab, "A multi-objective mathematical model and evolutionary algorithm for the dual-resource flexible job-shop scheduling problem with sequencing flexibility," Flex. Serv. Manuf. J., vol. 35, pp. 626–668, 2023, doi: 10.1007/s10696-022-09446-x.

[21] H. Zhang, Y. Ge, X. Zhao and J. Wang, "Hierarchical deep reinforcement learning for multi-objective integrated circuit physical layout optimization with congestion-aware reward shaping," IEEE Access, vol. 13, pp. 162533–162551, 2025, doi: 10.1109/ACCESS.2025.3610615.

[22] A. R. M. Jamil and N. Nower, "Dynamic weight-based multi-objective reward architecture for adaptive traffic signal control system," Int. J. ITS Res., vol. 20, pp. 495–507, 2022, doi: 10.1007/s13177-022-00305-5.

[23] N. Nigar, M. K. Shahzad, S. Islam, O. Oki, and J. M. Lukose, "A novel multi-objective evolutionary algorithm to address turnover in the software project scheduling problem based on best fit skills criterion," IEEE Access, vol. 11, pp. 89742–89756, 2023, doi: 10.1109/ACCESS.2023.3306838.

[24] X. Ran, W. P. Tay, and C. Lee, "High-performance optimization model based on novel conditional value at risk metric for power grids with high wind power penetration," IEEE Trans. Ind. Informat., vol. 21, no. 7, pp. 5689–5700, Jul. 2025, doi: 10.1109/TII.2025.3556085.

[25] Q. Lai, G. Liu, B. Zhang, and K. Zhang, "Simulating confidence intervals for conditional value-at-risk via least-

squares metamodels," INFORMS J. Comput., vol. 37, no. 4, pp. 1087–1105, 2025, doi: 10.1287/ijoc.2023.0394.
[26] H. Ghanbari, S. Tavakoli, M. Shabani, E. Mohammadi, S. J. Sadjadi, and R. R. Kumar, "A two-stage framework for enhancing cryptocurrency portfolio performance: Integrating credibilistic CVaR criterion with a novel asset preselection approach," PLoS One, vol. 20, no. 7, e0325973, 2025, doi: 10.1371/journal.pone.0325973.
[27] X. Wu, H. Zhang, H. Chen, S. Wang, and L. Gong, "Combustion optimization study of pulverized coal boiler based on proximal policy optimization algorithm," Appl. Therm. Eng., vol. 254, 123857, 2024, doi: 10.1016/j.applthermaleng.2024.123857.
[28] Y. Gu, Y. Cheng, C. L. P. Chen, and X. Wang, "Proximal policy optimization with policy feedback," IEEE Trans. Syst., Man, Cybern., Syst., vol. 52, no. 7, pp. 4600–4610, 2022, doi: 10.1109/TSMC.2021.3098451.
[29] A. Boudlal, A. Khafaji, and J. Elabbadi, "Entropy adjustment by interpolation for exploration in proximal policy optimization (PPO)," Eng. Appl. Artif. Intell., vol. 133, no. E, 108401, 2024, doi: 10.1016/j.engappai.2024.108401.
[30] H. Wu and H. Chen, "Multi-site wind speed prediction based on graph embedding and cyclic graph isomorphism network (GIN-GRU)," Energies, vol. 17, no. 14, 3516, 2024, doi: 10.3390/en17143516.
[31] C. Yang, Y. Xia, Z. Chu, and X. Zha, "Logic synthesis optimization sequence tuning using RL-based LSTM and graph isomorphism network," IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 69, no. 8, pp. 3600–3604, 2022, doi: 10.1109/TCSII.2022.3168344.
[32] S. Shaik, J. M. Smereka, and Y. Wang, "Generalized advantage estimation for distributional policy gradients," in Proc. Amer. Control Conf. (ACC), Denver, CO, USA, 2025, pp. 3073–3078, doi: 10.23919/ACC63710.2025.11107721.
[33] E. Jacinto, F. Martinez, and F. Martinez, "Navigation of autonomous vehicles using reinforcement learning with generalized advantage estimation," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 14, no. 1, pp. 954-959, 2023, doi: 10.14569/IJACSA.2023.01401103.
[34] M. Buyl and T. De Bie, "The KL-divergence between a graph model and its fair I-projection as a fairness regularizer," in Mach. Learn. Knowl. Discov. Databases, Res. Track, ECML PKDD 2021, Lecture Notes in Computer Science, vol. 12976, Springer, Cham, 2021, pp. 351–366, doi: 10.1007/978-3-030-86520-7_22.
[35] K. M. Tekale and N. Rahul, "AI bias mitigation in insurance pricing and claims decisions," Int. J. Artif. Intell. Data Sci. Mach. Learn., vol. 5, no. 1, pp. 138–148, 2024, doi: 10.63282/3050-9262.IJAIDSML-V5I1P113.
[36] J. A. Henderson, "Reducing adverse impact while maintaining validity: Finding the balance between competing employee selection goals," Ph.D. dissertation, Univ. Tennessee, Knoxville, TN, USA, 2004.