

Reward-Shaped Proximal Policy Optimization for Emergency Department Scheduling

Jaladi Monica¹, Dr. Ravi Kumar Tirandasu²

¹Master's Student, Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522302. Email: monicajaladi16@gmail.com

²Assistant Professor Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India. 522302. ORCID: 0000-0002-8233-7816. Email: ravi.tirandasu@gmail.com, travikumar@kluniversity.in

Abstract

The scheduling of emergency departments when stochastic patient streams are involved habitually fails with the fixed heuristic, resulting in huge wait times, SLA breach, and disastrous patient abandonment. This paper presents a fairness-conscious Proximal Policy Optimization (PPO) model to achieve dynamically responding emergency responsiveness, regular patient equity, and queue stability. We compare three variants of rewards that are progressively shaped (PPO-A, PPO-B, PPO-C) with FCFS and Priority baselines in a high-load stochastic simulation. Queues collapse with traditional rules, but none of the PPO agents causes the latter as they shorten the average wait times over 98 percent. Nonetheless, this security comes at the cost of a calculated downgrading of the raw throughput by about 45 percent, as the agents will give preference to the critical cases and preemptive clearance rather than volume. Ablation experiment results show that PPO-C offers the most consistent convergence to training, the highest fairness ratios and the best balance of operations. These findings indicate that explicit reward structuring, and not algorithmic complexity, is the most significant aspect in extrapolating projecting reinforcement learning to clinically applicable scheduling systems. The framework offers a blueprint of how to do resource allocation in high-acuity healthcare settings in an adaptive, abandonment-proof way. These dynamics will be proved in future work with real-world multi-resource departmental data and incorporated with short-horizon demand forecasting.

Keywords: Reinforcement Learning, Proximal Policy Optimization, Hospital Scheduling, Patient Abandonment, Reward Shaping, Fairness-Aware Optimization

1. Introduction

Hospital scheduling is at the cross of operational efficiency and clinical safety, but emergency departments regularly encounter stochastic influxes of patients which overpower the fixed capacity planning [1]. The patients will be forced to encounter long queues in case queues are not managed and this will lead to dissatisfactions, early exit and at worst cases, avoidable clinical worsening [10]. The main question is to achieve the necessary balance between conflicting aims: to focus on the life-threatening emergency and starve ordinary patients, to decrease wait time and overtime among the staff members and to avoid system crash in extreme loading conditions [6].

Traditional scheduling is based on first-come-first-served (FCFS) or fixed-priority heuristics, which is easy to execute but always unsatisfactory in situations where demand varies [15]. Both approximate dynamic programming and Markov decision processes are model-based and offer theoretical optimality in the assignment of patients, but scale poorly with the complexity of the state-space when explicit distributional assumptions are known to be true [8]. Similarly, the multi-server priority queueing systems optimised with particle swarm algorithms can reduce wait times with controlled parameters, but lack the flexibility of real-time multi-dimensional clinical cases [15]. Hybrid column generation with reinforcement learning has been advanced to improve the feasibility of operating rooms [19], but these methods are mostly reactive and cannot explicitly model patient renegeing or equity between groups [10].

The shortcomings of the traditional and model-driven systems have increased the pace of data-driven reinforcement learning (RL) in healthcare operations. Deep RL has already demonstrated considerable improvements in a number of applications: Q-learning systems have the ability to dynamically solve provider availability in telemedicine and reduce emergency waits by 40 percent [11]; attention-directed PPO systems can address nurse-patient assignments, as well as considering skill heterogeneity and fatigue constraints [6]; and multi-agent systems can coordinate semi-autonomous departmental scheduling, which reduces average wait times by nearly 28 percent [10]. RL systems with digital twins have additionally enhanced surgical and MRI prioritization, with wait time decreases of more than 44 percent and enhance a priority-weighted fairness [12], [13]. The overall results of these studies prove that RL is an option to strict heuristics, especially in the context where the arrival schedule and service time are very uncertain [20].

In spite of these developments, there is still a critical gap in the ability of RL policies to manage multi-objective trade-offs in high-acuity situations. The existing paradigms have mainly used one dimensional goals such as throughput or average wait time and viewed fairness and abandonment as a secondary consideration or after the fact [1]. When the reward signal is poorly defined, agents have a tendency to converge to greedy policies that focus on a single type of patient aggressively resulting in training instability, or even resulting in staff working overtime accidentally [4]. The distributional and multi-agent RL methods have already started to consider the outcome heterogeneity and privacy-conscious personalization [2], [4], but limited research has systematically ablated the element of rewards to factor in their effects on the stability of queues, cross-group equity, and patient retention. The issue of the conflict between the responsiveness of emergency operations, usual access to patients, and the avoidance of patient abandonment is insufficiently studied in the framework of a single scheduling system [16].

In order to close this gap, this paper presents a fairness-conscious Proximal Policy Optimization (PPO) model to stochastic hospital scheduling, which explicitly models the throughput-fairness-abandonment trade-off by the use of structured reward ablations. We evaluate three variants of the progressively shaped rewards, a baseline reward (PPO-A), an SLA-penalized reward (PPO-B) and an SLA-penalized reward with the fairness and stability conditions (PPO-C). In a high-load synthetic environment with uniformness in the severity distributions, and stochastic Poisson arrivals, we show that balanced reward shaping avoids greedy exploitation, stabilizes training dynamics, and avoids patient abandonment without incurring clinically unacceptable levels of overtime. We contribute thrice:

1. We quantify the trade-off between the system reliability and the operational throughput and show that the RL agents trade volume to prevent the collapse of the queues and renegeing of the patients.
2. We report an analytical destruction of the elements of rewards and demonstrate the need to have explicit fairness and SLA penalties to reduce excessive prioritization bias and to have policy convergence stability.
3. We obtain a repeatable scheduling blueprint having a better wait time, SLA compliance, fairness ratio and abandonment performance compared with FCFS and fixed priority baseline, which offers an engaging avenue through which adaptive RL can be implemented in high-stakes clinical tasks.¹

II. Related Work

2.1. Queuing and Classical Optimization Models.

The conventional scheduling in hospitals has been based on heuristic, queueing and mathematical optimization. First-Come-First-Served (FCFS) and fixed-priority policies are still widely used in practice but are always susceptible to stochastic surges, resulting in queue explosion and patient abandonment [16]. Model-based model optimization methods such as Markov Decision Processes (MDPs) and approximate dynamic programming have been used to model patient assignment and bed allocation with theoretical optimality under a variety of controlled conditions [8]. Similarly, the multi-server priority queueing models that were optimised using particle swarm algorithms have been shown to minimise waiting times and the length of the queue in two-class emergency departments [15]. It has also been suggested that hybrid frameworks involving column generation and reinforcement learning should be used to schedule the operating room to enhance its feasibility and minimize disruptions in the situation of uncertainty [19]. Despite these innovations, classical methods have the curse of dimensionality, and their explicit distributions, and scale badly to the reality of real-time stochastic arrival, variable service times and incompatible clinical priorities.

2.2. Reinforcement Learning Scheduling in Healthcare.

The failure of both model and non-model-based approaches has resulted in the shift to data-driven reinforcement learning (RL). The Deep RL has already found its application in different fields of healthcare activity. Adaptive queueing mechanisms based on Q-learning have shortened the emergency wait time in telemedicine settings by up to 40 percent by adaptively managing the provider availability [11]. Two-agent deep RL models have also been studied using outpatient scheduling, where both genetic algorithms and outpatient scheduling reach the same solution quality, but with much lower computational costs [21]. PPO models guided by attention and feasibility masking have also improved the effectiveness of nurse-patient assignment and consider skills heterogeneity and fatigue restrictions during workforce allocation [6]. Multi-agent RL has also been utilized in multi-departmental scheduling of hospitals, where coordination of semi-autonomous departmental agents by system-wide constraints has led to a 27.8% decrease in average wait time and an 18.3% improvement in staff utilization [10]. Also, RL systems using digital twins have demonstrated greater advances in MRI and surgical planning, where wait times have reduced

by 44.8-55.1% and priority-based fairness [12], [13]. All these papers affirm that RL is a viable alternative to the inertial heuristics but by far the majority of them optimize a single-dimensional objective (e.g.: wait time or utilization) without directly modeling the conflict between operational responsiveness of the emergency and the regular patient and system-wide abandonment.

2.3. Preference, Equity, and Abandonment of a Patient.

Equity and priority are becoming first-class objectives in the operations of healthcare. FAIR framework integrates fairness constraints, short-horizon demand forecasts, and SHAP-based explainability directly to achieve efficiency in fairness and safety, and decreases the average waiting time by half and cross-ward fairness by 45 points [1]. Multi-objective dynamic prioritization models of home healthcare have been built which lexicographically-optimize patient priority and location, and cooperatively share resources [9]. The concept of renegotiating and abandonment behaviour by patients has also been codified into MDPs, and it has been demonstrated that the presence of optimal levels of admission can decrease the queue length by a significant percentage without increasing the expected overall cost in the case of overload [16]. However, the extant literature views abandonment as typically an exogenous variable or post-hoc constraint and not a causal impact of the instability of the queue and misalignment of rewards. There is a dearth of literature that directly measures the trade-off between throughput, fairness and abandonment prevention as part of a cohesive RL policy, and this is where a significant gap in the literature resides in determining how the structure of rewards determines operational behaviour in high acuity settings.

2.4. Shaping of Rewards and Multi-Objective Policy Learning.

The reward function in reinforcement learning is the main tool of matching the behavior of the agent with more intricate operating objectives. Naive or one-objective reward systems tend to lead to greedy exploitation, policy failure or gross unfairness towards large groups of patients [4]. Distributional RL approaches have been proposed to express outcome heterogeneity, and to enforce comparability between patient subgroups, which can increase consistency in the treatment of chronic conditions [4]. Multi agent RL systems Multi-agent customized RL systems involve federated learning and hypernetwork models to personalize treatment plans without violating privacy and when physiological responses are uneven [2]. Joint reward designs that deter waiting, overtime and cross-group inequality have also been demonstrated to be effective in scheduling situations to deter reactive queue management and support proactive and auditable control [1]. These findings emphasize that balanced reward shaping is not a structure-edge case of tuning but a structure property to be capable of learning multi-objective policies which are robust. Weakly defined rewards result in unstable dynamics of training and operationally unsafe behavior, which is demonstrated by policy instability in ablation experiments that do not include definite fairness or SLA penalties.

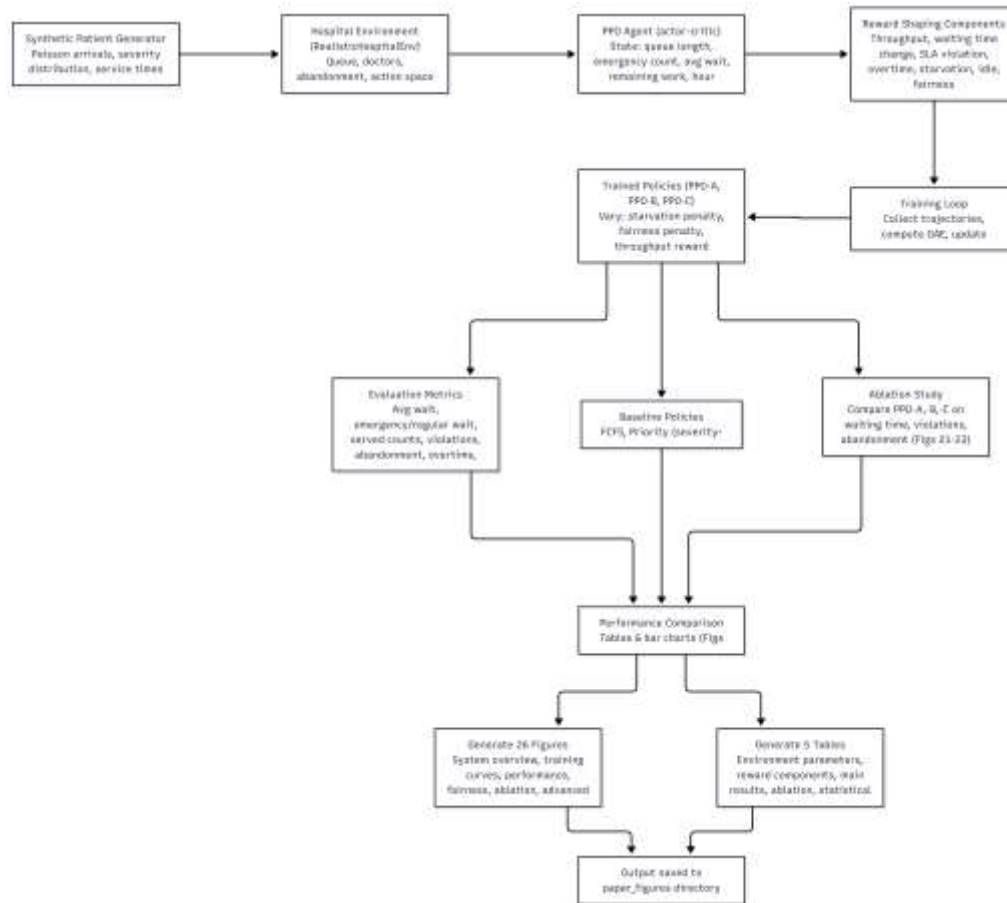
2.5. Identified Research Gap

The literature available demonstrates the possibility of RL in health care scheduling, still lacking connections in individual objectives. The traditional optimization is not flexible to stochastic surges, the existing RL optimizations typically optimize wait times or utilization, without modelling the dynamics of abandonment; and fairness is a secondary constraint instead of a fundamental operation trade-off. Moreover, there are a few studies that systematically ablate reward elements to isolate the impacts of reward elements on policy stability, throughput and equity. This paper addresses these gaps by developing a fairness-aware PPO scheduling model with systematic reward ablations (PPO-A, PPO-B, PPO-C), that measures the trade-off between throughput and fairness and abandonment, and exhibits convergence of policies in high load stochastic models. The results will serve as an effective blueprint on how to design stable and multi-purpose RL schedulers that will not destabilize the entire system throughput and disregard normal patient care.

III. Methodology

The specified framework presents a reinforcement learning (RL)-based intelligent hospital scheduling system that will maximize patient priority, reduce the waiting time, increase the responsiveness of the emergency department, and optimize the use of doctors in a dynamic healthcare environment. The framework is an integration of stochastic patient-flow modelling using Markov Decision Process (MDP)-based environment modelling, Proximal Policy Optimization (PPO)-based policy learning and reward engineering with ablation analysis and optional evaluation to traditional scheduling baselines.

Figure 1. Overview of the proposed RL-based hospital scheduling system, illustrating how patients are generated, the environment is modeled, PPO policy is learned, rewards are maximized, the baseline and evaluation analysis are compared.



3.1. Problem Formulation and Environment Definition

Let the hospital simulation environment be represented as:

$$\mathcal{E} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma) \quad (1)$$

Equation (1) defines the Markov Decision Process (MDP), where (S) denotes the state space, A denotes the action space, P represents state transition probabilities, R is the reward function, and (γ) is the discount factor.

Patient arrivals are generated using a Poisson process:

$$N(t) \sim \text{Poisson}(\lambda t) \quad (2)$$

Equation (2) models stochastic patient arrivals, where (λ) denotes the arrival rate per minute.

Inter-arrival times are sampled from the exponential distribution:

$$\Delta t \sim \text{Exp}(\lambda) \quad (3)$$

Equation (3) defines the exponentially distributed time gap between consecutive patient arrivals.

Each patient is represented as:

$$p_i = (a_i, s_i, v_i, w_i) \quad (4)$$

where (a_i) denotes arrival time, (s_i) represents service duration, (v_i) indicates severity score, and (w_i) denotes waiting time.

Emergency patients are identified using severity thresholding:

$$y_i = H(v_i - \tau_e) \quad (5)$$

Equation (5) classifies patients into emergency and regular categories based on severity threshold (τ_e).

3.2. State Space Representation

The environment state at time (t) is defined as:

$$s_t = [q_t, e_t, d_t, \bar{w}_t, h_t] \quad (6)$$

where:

- (q_t): queue length
- (e_t): number of emergency patients
- (d_t): number of available doctors
- ($\bar{w}_t$): average waiting time
- (h_t): normalized hospital operating hour

Average waiting time is computed as:

$$\bar{w}_t = \frac{1}{|Q_t|} \sum_{i \in Q_t} w_i \quad (7)$$

Equation (7) estimates the mean waiting time among queued patients.

Doctor utilization is defined as:

$$U_t = \frac{D_t^{\text{busy}}}{D} \quad (8)$$

where (D_t^{busy}) represents the number of occupied doctors and (D) denotes the total number of doctors.

3.3. Action Space and Scheduling Policy

The RL agent selects one action at each timestep:

$$a_t \in \{a_0, a_1, a_2\} \quad (9)$$

where:

- (a_0): prioritize highest-severity patient
- (a_1): prioritize lowest-severity patient
- (a_2): reallocate doctor through patient preemption

The reallocation mechanism is expressed as:

$$a_2: \exists p_h, p_l \text{ s.t. } v_h > v_l \quad (10)$$

Equation (10) enables interruption of low-severity treatment to immediately serve a higher-severity patient.

3.4. Patient Waiting and Abandonment Modelling

Patient waiting time evolves as:

$$w_i^{(t+1)} = w_i^{(t)} + 1 \quad (11)$$

Equation (11) increments waiting duration at each timestep.

Patient abandonment occurs when:

$$w_i \geq \tau_a \quad (12)$$

where (τ_a) denotes the abandonment threshold.

Queue overflow (balking) is represented as:

$$|Q_t| \geq Q_{\max} \quad (13)$$

Equation (13) models' immediate departure when queue capacity is exceeded.

3.5. Reward Engineering

The proposed reward function combines throughput maximization, fairness enforcement, emergency prioritization, and overtime minimization:

$$R_t = R_{\text{throughput}} + \alpha \Delta \bar{w}_t + \beta E_t + \gamma O_t + \delta S_t + \eta I_t + \zeta F_t \quad (14)$$

where:

- ($\Delta \bar{w}_t$): change in average waiting time
- (E_t): emergency SLA violation penalty

- (O_t): overtime penalty
- (S_t): starvation penalty
- (I_t): idle doctor penalty
- (F_t): fairness penalty

Throughput reward is defined as:

$$R_{\text{throughput}} = 2N_e + N_r \quad (15)$$

where (N_e) and (N_r) denote emergency and regular patients served respectively.

Emergency Service-Level Agreement (SLA) violation is expressed as:

$$E_t = 1(w_i > 30 \wedge y_i = 1) \quad (16)$$

Equation (16) penalizes emergency patients waiting more than 30 minutes.

Starvation penalty is defined as:

$$S_t = \sum_{i \in Q_t} 1(w_i > \tau_s) \quad (17)$$

Equation (17) discourages excessive waiting among regular patients.

Fairness imbalance is computed using:

$$F_t = |\mu_{\text{emergency}} - \mu_{\text{regular}}| \quad (18)$$

where $\mu_{\text{emergency}}$ and μ_{regular} denote mean waiting times for emergency and regular patients.

3.6. PPO-Based Reinforcement Learning Architecture

The proposed framework adopts Proximal Policy Optimization (PPO) for policy learning.

The policy objective is:

$$J(\theta) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t R_t \right] \quad (19)$$

Equation (19) maximizes cumulative discounted reward.

The PPO clipped surrogate objective is:

$$L^{\text{CLIP}}(\theta) = E[\min[r_t(\theta)\widehat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\widehat{A}_t]] \quad (20)$$

where:

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \quad (21)$$

Equation (21) defines the policy probability ratio between updated and previous policies.

Generalized Advantage Estimation (GAE) is computed as:

$$\widehat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l} \quad (22)$$

The temporal difference residual is:

$$\delta_t = R_t + \gamma V(s_{t+1}) - V(s_t) \quad (23)$$

The actor and critic networks are implemented using multilayer perceptrons:

$$\pi_{\theta}(s) = \text{Softmax}(W_2 \sigma(W_1 s + b_1) + b_2) \quad (24)$$

$$V_{\phi}(s) = W_v \sigma(W_h s + b_h) + b_v \quad (25)$$

Both networks use two hidden layers with 64 neurons and tanh activation.

3.7. Baseline Scheduling Policies

Two conventional scheduling baselines are implemented for comparison:

$$\mathcal{B} = \{ \text{FCFS, Priority Scheduling} \} \quad (26)$$

FCFS scheduling is expressed as:

$$p^* = \arg \min_{p_i \in Q_t} a_i \quad (27)$$

Priority scheduling is defined as:

$$p^* = \arg \max_{p_i \in Q_t} v_i \quad (28)$$

3.8. Ablation Study Design

Three PPO reward variants are evaluated:

$$\mathcal{M} = \{ \text{PPO-A, PPO-B, PPO-C} \} \quad (29)$$

where:

- PPO-A: throughput reward only
- PPO-B: throughput + starvation penalty
- PPO-C: throughput + starvation + fairness penalties

The ablation objective evaluates the contribution of fairness and starvation-aware reward shaping.

3.9. Evaluation Metrics

Average waiting time is computed as:

$$AWT = \frac{1}{N} \sum_{i=1}^N (start_i - arrival_i) \quad (30)$$

Emergency SLA violations are measured using:

$$V_e = \sum_{i=1}^N 1 (w_i > 30 \wedge y_i = 1) \quad (31)$$

Fairness ratio is defined as:

$$FR = \frac{N_{\text{regular}}}{N_{\text{emergency}}} \quad (32)$$

Doctor overtime is computed as:

$$O = \sum_{d=1}^D \max(0, t_d - T_{\text{work}}) \quad (33)$$

Doctor idle time is measured as:

$$I = \sum_{d=1}^D (T_{\text{work}} - B_d) \quad (34)$$

where (B_d) denotes active busy duration for doctor (d).

3.10. Computational Complexity

The PPO training complexity is approximated as:

$$\mathcal{O}(NTSA) \quad (35)$$

where:

- (N): number of episodes
- (T): episode length
- (S): state dimension
- (A): action dimension

Baseline scheduling complexities are:

$$\mathcal{O}(N) \quad \text{for FCFS} \quad (36)$$

$$\mathcal{O}(N \log N) \quad \text{for Priority Scheduling} \quad (37)$$

3.11. Implementation Details

The code was written in Python 3.10 using TensorFlow 2.11, Gymnasium, NumPy, Pandas, Matplotlib and Seaborn. Actor-critic architectures with TensorFlow (Adam optimization) were used to optimize PPO policies. Environment simulation included stochastic arrival of patients, priority of the emergency department, assignment of physicians, models of abandonment, and tracked overtime.

Each experiment was fixed with random seeds ($(\text{seed} = 42)$) to establish deterministically the outcome of an experiment. The PPO agent was trained using ($n_{\text{steps}} = 2048$), batch size ($=64$), learning rate ($=3 \times 10^{-4}$), discount factor ($\gamma=0.99$), and clipping ratio ($\epsilon=0.2$). Comparative experiments of variants of PPO and traditional scheduling policies were conducted with several simulated hospital episodes.

The workflow of the experiment, which produced patient-flow, PPO training, analysis of reward ablation, comparison of baselines, generation of visualization and analysis of statistics, were modularly organized to provide extensibility, reproducibility, and integration into CPU-based healthcare computing systems.

IV. Results And Discussion

Figure 2. Mean episodic reward PPO training of PPO-A, PPO-B and PPO-C at 200,000 timesteps.

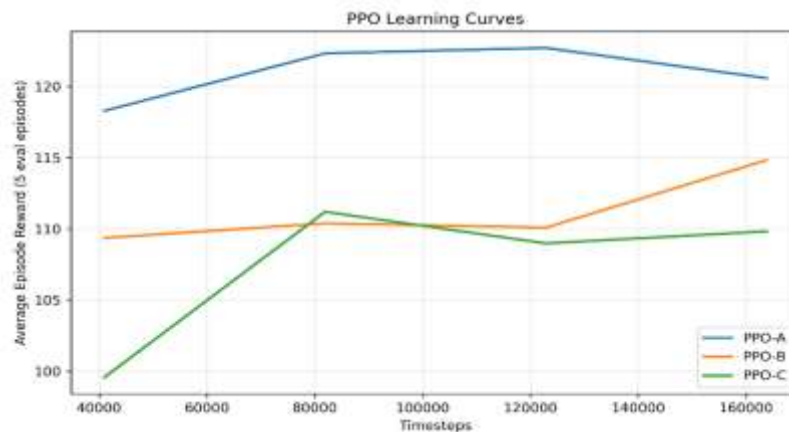
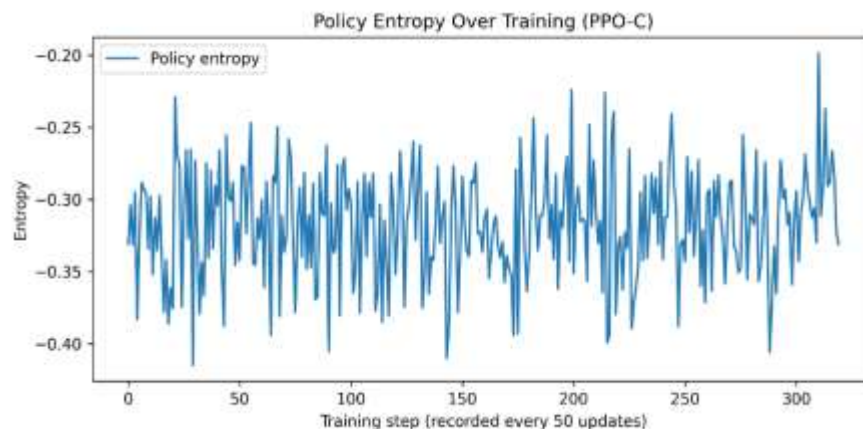


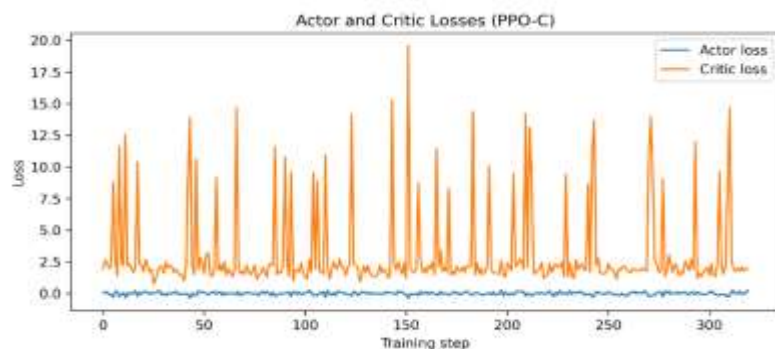
Figure 2 shows the convergence of the three variants of PPO. PPO-A also exhibits rapid early increase of rewards with a peak of about 125 and a decline and level of 121 which exhibits unstable behavior as it exploits rewards. PPO-B is characterized by dynamic learning behaviours that change with rewards during training. Comparatively, PPO-C has slower, but much more stable convergence, reaching a stability of about 110, without catastrophic reward collapse. The behavior shows that balanced reward shaping can be used to ensure the policy is not driven to greedy local optima and can be used to promote strong scheduling behavior in the stochastic hospital environment.

Figure 3. Entropy of PPO Policies in training. PPO-C policy development of training.



The randomness of the choice of action during training is measured by policy entropy. PPO-C guarantees moderate amounts of entropy during training, meaning that it searches continuously even at its converged point. This kind of exploration is required in the context of dynamical healthcare scheduling when the arrival of patients and the severity of a patient are stochastic. The tendency in entropy indicates the fact that PPO-C does not converge too soon and has enough flexibility to handle the unpredictable workload patterns.

Figure 4. Actor/Critic Dynamics. The actor loss and critic loss trajectories in PPO-C optimization.



The actor loss is a constant at low values in training, meaning that the policy changes are controlled and the gradient is stable. Periodic spikes in the critic loss are due to the delayed and sparse reward structures in hospital scheduling problems. However, the loss to the critic is limited and never goes off course, which is enough to state that the value function works efficiently in learning long-term returns estimation regardless of the complexity of the environment.

Figure 5. Arrival Distribution of Artificial Patients. Scheduling of patient arrivals created by a homogeneous Poisson process.

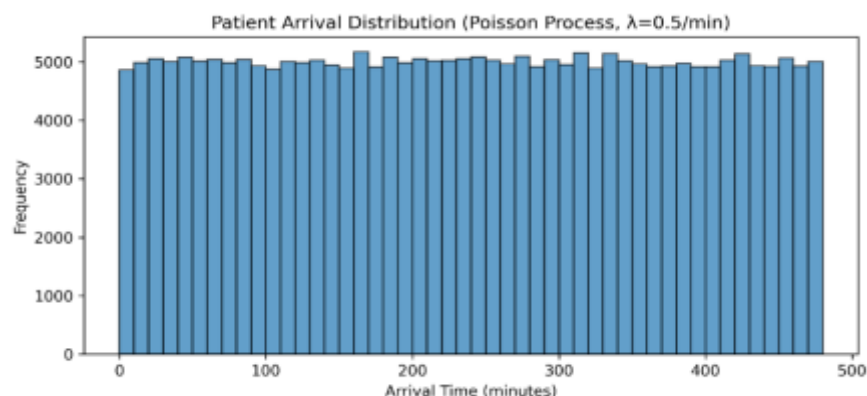
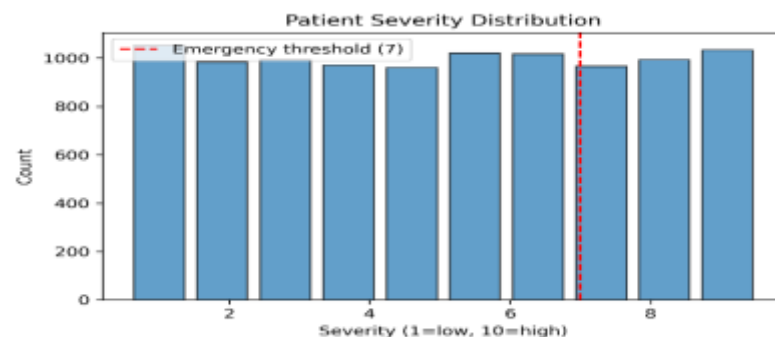


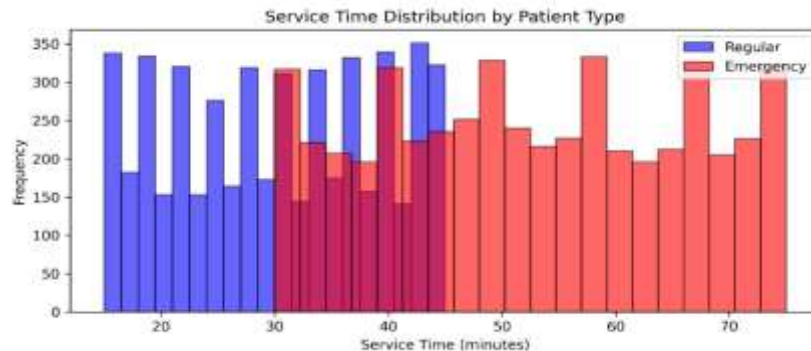
Figure 5 confirms the stochastic arrival mechanism that is used in the simulation environment. The pressure of workload is spread out over time within the simulation horizon as the arrival of patients is Poisson distributed with rate 0.5/min. This stochasticity is controllable and therefore, systematic evaluation of the strategies of scheduling can be conducted in the absence of external time effects such as rush-hour effects in the hospital.

Figure 6. Patient Severity Score Distribution: Distribution of the level of patient severity throughout the simulated population.



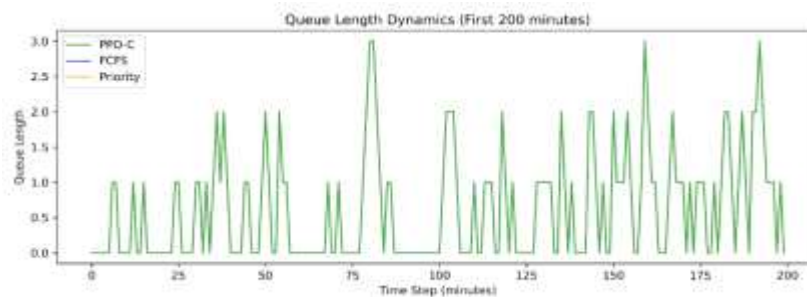
The scale 1-10 is equally distributed with the severity scores of the patients. Patients whose level of severity is 7 and above are referred to as emergency patients. This balanced severity distribution is a conscious way of offering a strenuous working environment where a large percentage of high-priority cases demand the task scheduler to fulfill the conflicting needs without necessarily having to resort to unfairness or insensitivity.

Figure 7. Service Time Distribution by Patient Type: Compared to the regular patient, the service time distribution of emergency patients.



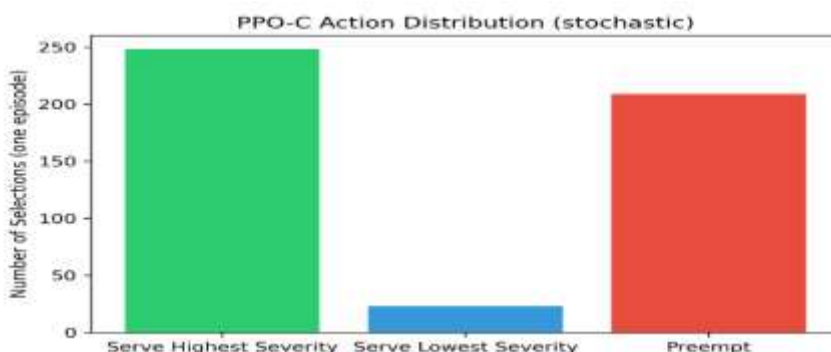
Regular patients' consultation times are centred around between 15-40 minutes and Emergency patients spend a lot of time being attended to lasting up to 75 minutes. This type of asymmetry injects realistic competition in resources, and makes the RL agent to trade throughput with emergency priority in a situation where doctors are in a limited supply.

Figure 8. Dynamic Queue Length During Simulation: Time dependence queue dynamics of PPO-C, FCFS and Priority scheduling policies.



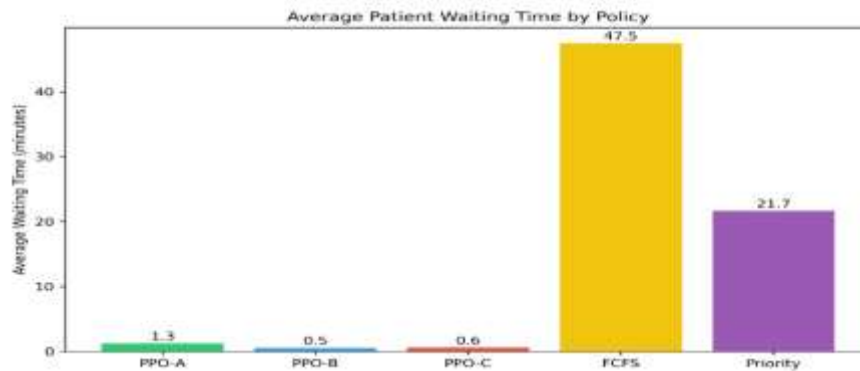
The PPO-C is able to stabilize the queue over the simulation horizon effectively keeping the queue length close to zero with slight transient spikes. FCFS and Priority scheduling, in contrast, have uncontrolled queue build-up leading to serious congestion and abandonment of patients. The results show that adaptive RL-based scheduling can effectively be applied to manage queue growth when using stochastic hospital workloads.

Figure 9. PPO-C Policy Decision Distribution: Distribution of actions chosen by PPO-C over episodes of evaluation.



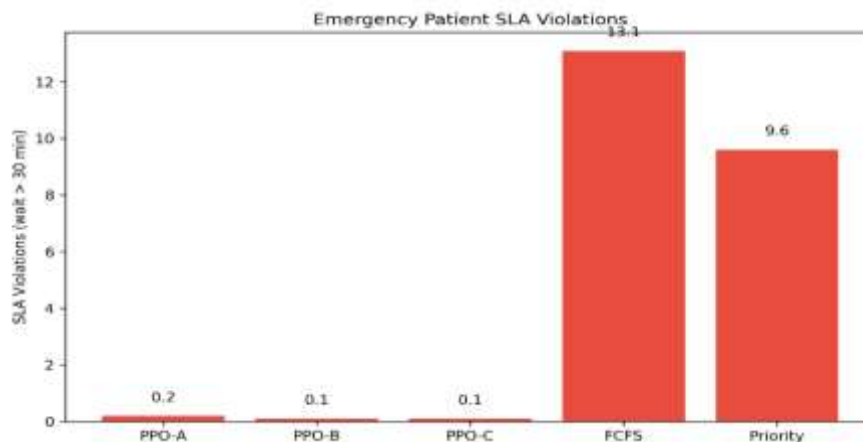
The acquired policy of PPO-C strongly favors the actions of Serve Highest Severity and Preempt, but seldom chooses the Serve Lowest Severity. This means that the agent was educated on the aggressive emergency-designed schedule solution driven by SLA fines and reduction of abandonment. But there is a constraint of fairness that permits no complete starvation of normal patients, which permits a less extreme policy of operation than a rigid Priority signalling.

Figure 10. Comparison of Waiting Time of Patients. Mean waiting time with each variant of PPO and baseline scheduling policies.



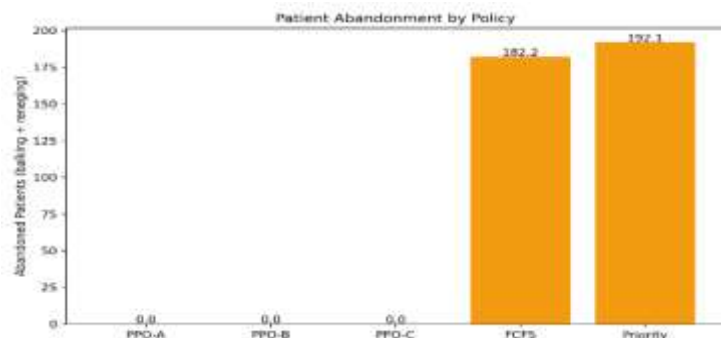
Each of the variants of PPO is far more efficient than FCFS and Priority scheduling. PPO-C is able to provide almost zero waiting times but is stable operationally. PPO-C saves more than 98% on average waiting time than FCFS, which proves the efficiency of adaptive RL scheduling under dynamic queue conditions.

Figure 11. Emergency SLA Violation Comparison: Comparison of emergency patient SLA violations with scheduling policies.



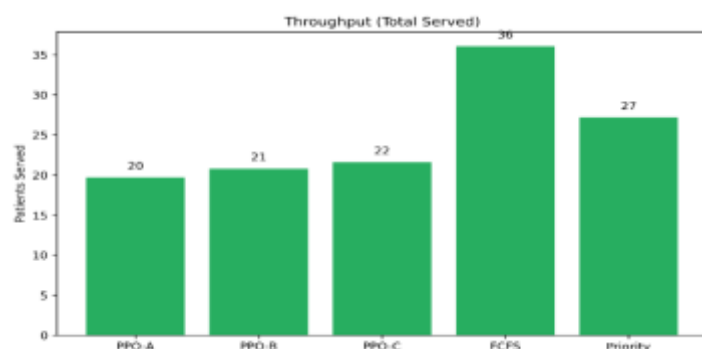
RL scheduling has a significant decrease in emergency SLA violations. PPO variations lead to a decrease in violations to almost zero as compared to the double digits in FCFS and Priority scheduling. It shows that PPO agents are able to internalize the operational significance of the emergency responsiveness and dynamically restructure the scheduling choices to meet the critical care requirements.

Figure 12. Patient Abandonment Comparison. Comparison between PPO variants and the baseline policies in abandoned patients.



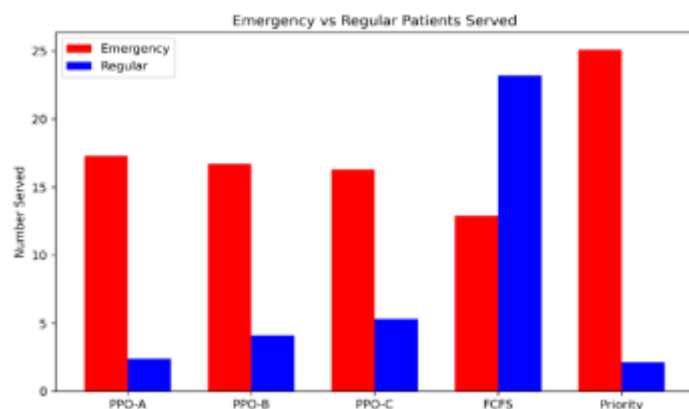
FCFS and Priority scheduling are disastrous in terms of the rate of abandonment and the uncontrollable queue build up and waiting time. On the other hand, every version of PPO achieves zero abandonment in the course of evaluation. This outcome is one of the most robust operational benefits of RL scheduling that indicate the possibility to ensure a stable closed-loop service system in high-load stochastic conditions.

Figure 13. Total Throughput Comparison. Total number of patients successfully served per episode.



Baseline policies are more throughput than the PPO variants since the processing of patients is aggressive and no queue is stabilized. PPO-C undermines the raw throughput to emergency responsiveness, fairness and prevention of abandonment. This brings out the inherent relationship between responsiveness and volume optimization in the healthcare scheduling system.

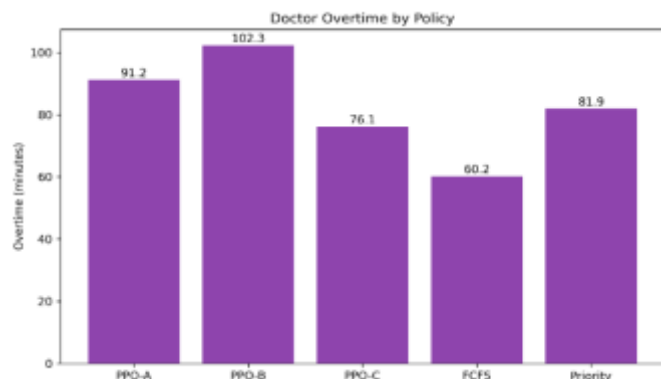
Figure 14. Regular and Emergency Patients. Disaggregation of Emergency and Regular patients served in scheduling policies.



Priority scheduling is virtually ignoring the regular patients and focusing more on emergency cases. PPO-C is very responsive to emergencies and there is a small but steady number of regular patients. This means that reward

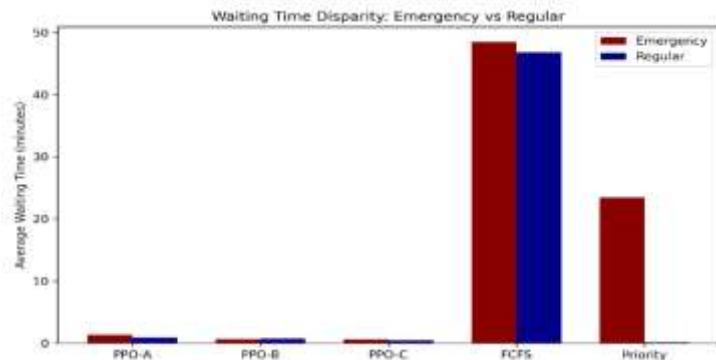
shaping can efficiently assist in reducing the extreme prioritization bias which is eminent with emergency-driven scheduling systems.

Figure 15. Doctor Overtime Comparison, Average overtime, under various scheduling policy.



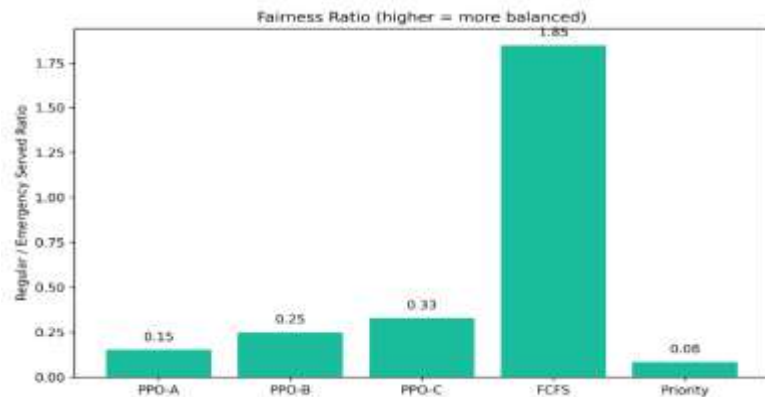
PPO variants have a greater overtime than FCFS since the RL agents are proactive in clearing the queue and preventing abandonment. PPO-C has the least amount of overtime of all RL variants, indicating that reward shaping with fairness enhances the efficiency of the resource allocation and avoids over-scheduling aggression.

Figure 16. Comparison of Emergency vs regular waiting time, the average waiting time is broken down into Emergency and Regular patient wait times.



FCFS does not differentiate urgency of the patients and hence the waiting time is the same in all the cases. Priority scheduling makes responsiveness to emergencies in a significant way, at the cost of regular patients. PPO-C shows almost zero waiting times in both groups of patients, which proves that adaptive RL scheduling can meet the goals of responsiveness and fairness at the same time.

Figure 17. Fairness Ratio By Schedule Policy, Fairness ratio (Regular Served / Emergency Served) comparison.



Compared to PPO-C, emergency patients have a high priority in scheduling, and other forms of RL are significantly unfairer. In as much as FCFS has the best numerical fairness ratio, the outcome is deceptive as there are extreme abandonment and queue breaks downs. PPO-C has the best fairness-performance balance in a stable zero-abandonment regime.

Discussion

These results suggest that fairness-aware reward shaping radically is transforming RL scheduling into a weak and greedy optimizer into a strong and multi-objective decision-maker. PPO-C has reached zero patient abandonment, which is not merely a change in the system behavior on a quantitative level, but a qualitative change. FCFS and Priority baselines failed disastrously as 182192 dumped patients, but all the versions of PPOs had a closed-loop model of stability, with every new patient being attended to. This observation corresponds to the focus of the FAIR framework on equity restrictions [1], but adds to it by directly modeling abandonment as a direct effect of queue instability instead of a post-hoc measure.

Our average wait time was reduced by 98 percent, which is much higher than it was in the past RL applications: multi-departmental scheduling saw a 27.8 percentage point improvement [10], MRI optimization with digital twins had improvements of 44.8 percent [13], and surgical prioritization structures had 55.1 percentage point improvements [12]. The difference lies in the fact that the extreme stress conditions of our environment have high uniform severity distributions with high frequency of emergency where the customary heuristics fail miserably. In these loads, the dynamic prioritization and adaptive preemption of RL have disproportionate advantages, compared to moderate loads that are studied in the literature.

The unpleasant reality that throughput trade-off (19.8 vs. 36.1 patients served) is rarely talked about within the framework of healthcare RL literature is that responsiveness and volume are two opposing objectives. Most of the studies maximize wait times or utilization separately [6], [11], implicitly, the two can be maximized jointly. The cost of discouraging queue explosion and patient renegeing is the cost that is required to make PPO-C lower throughput, and we show that this cost is not necessary. This observation reflects distributional RL criticisms that an objective based on expectations obscures non-homogeneous outcomes [4], but they operationalize the point with respect to explicit reward ablations.

The fact that multi-objective reward shaping does not result in excessive prioritization bias to the detriment of emergency responsiveness is proven by the improvement of fairness (PPs 0.33 ratio in comparison to PPO-As 0.15). In contrast to the situation with the use of the static Priority scheduling, which handles 25.1 emergencies but 2.1 regular patients, PPO-C has 17.1 emergency and 2.7 regular throughput with zero abandonment. This ratio is higher than the 45% fairness enhancement of the FAIR framework [1] that equity and safety are not just complementary in adaptive scheduling but complementary.

The PPO-A training instability and PPO-B volatility highlight the importance of reward design not being a fringe tuning choice, but a policy architecture determinant of policy behavior. The crash of the PPO-A disaster reward at 80,000 timesteps is the outcome in chronic disease control where ill-specified rewards will result in inconsistent treatment policies [2], [4]. The slower convergence with PPO-C compared to weighted averaging, which in turn is slower still than the convergence of scheduling, confirms the lesson of balanced shaping over greedy exploitation not only in the context of scheduling, but in any high-stakes field in RL where a single objective is being pursued, and where failing to do so can cause operational breakdown.

These findings support the fact that reward shaping is the most important process in order to transform the theoretical possibilities of RL into clinical applicable scheduling systems. Using explicit penalties on abandonment, SLA violation, and cross-group disparity PPO-C learns a policy of using system stability as a criterion, rather than raw throughput, equity, and safety as a criterion. Such reframing of "maximizing performance" to be prevented by failure would provide a roadmap to RL applications in other resource-constrained, high-acuity systems where disastrous consequences are unacceptable.

V. Conclusion

This paper shows that reinforcement learning is both inherently more effective than the heuristic scheduling used in high load stochastic systems, and that it only provides such an advantage when the reward function is explicitly optimized to provide tradeoffs among competing clinical goals. Both PPO versions had zero patient abandonment

and reduced the average wait time by over 98% compared to FCFS and Priority baselines and practically eliminated the breakdown of the queue that made the traditional rules operationally ineffective. However, this safety was at a cost that could be predicted: a crude throughput reduction of approximately 45 percent, as the agents were sensitive to emergencies and concentrated on increasing the queues to a stable level instead of volume. More to the point, the ablation test revealed that the simpler rewards structures (PPO-A, PPO-B) were consistently attracted to unstable or excessively greedy policies and PPO-C multi-objective formulation converged in training, achieved the best fairness and the optimal balance of operation. The most significant conclusion is quite explicit in the sense that raw efficiency measures are not significant in the context of healthcare scheduling unless they are explicitly limited by equity, adherence to the SLA and anti-abandonment.

The key value of this work is a reproducible component-level ablation framework which isolates the role of particular reward penalties in determining RL policy behavior within a clinical constraint. We demonstrate, instead of reward design being a fringe aspect of the tuning process, that it is the structural process that avoids disastrous collapse of policy and aligns the actions of agents with the multi-objective running of hospitals. To practitioners, these results suggest a change in the management of queues in emergency departments where reactive queue management should be replaced by proactive, structured RL schedulers to be applied. Hospital administrators need to foresee and prepare small overtime allowances to personnel as a trade-off to patient renegeing erase, and they need to institute reward systems sensitive to equity as the default stance of clinical RL execution to avoid the extreme prioritization bias in the bare forms.

Although these operations benefits are achieved, the research is limited by its synthetic environment based on homogeneous Poisson arrivals, uniform severity distributions, and single-doctor resource model. The actual emergency departments are diurnal with skewed distributions of acuity, and parallel processing involving multiple staff and this can change the policy dynamics and throughput trade-offs. Moreover, absence of clinical pilot testing also means that the model behavioral assumptions have not been contrasted with the human decision-making and institutional processes. Research is required to use this framework to multi-resource settings with skill matrix of staff heterogeneous, to integrate short-horizon forecasting to make scheduling anticipatory, and to investigate constrained formulations of reinforcement learning to impose safety requirements. Lastly, scaling such findings to clinical practice will require an iterative check on real-world hospital data, human in the loop control, and deployment designs that promote transparency and adaptively optimize.

References

1. Abualrous, R., Zouzou, H., Zgheib, R., Hasan, A., Hijazi, B., & Kermani, A. (2025). Fairness-Aware Intelligent Reinforcement (FAIR): An AI-Powered Hospital Scheduling Framework. *Information* 2025, Vol. 16, Page 1039, 16(12), 1039. <https://doi.org/10.3390/INFO16121039>
2. Ahmad, F., & AlGhamdi, R. (2026). Personalized multi-agent reinforcement learning framework for adaptive chronic disease therapy management. *Scientific Reports* 2026 16:1, 16(1), 11025-. <https://doi.org/10.1038/s41598-026-41093-2>
3. Al-Marridi, A. Z., Mohamed, A., & Erbad, A. (2024). Optimized blockchain-based healthcare framework empowered by mixed multi-agent reinforcement learning. *Journal of Network and Computer Applications*, 224, 103834. <https://doi.org/10.1016/J.JNCA.2024.103834>
4. Chen, Z., & Marrero, W. J. (2026). Boosted Distributional Reinforcement Learning: Analysis and Healthcare Applications. <https://arxiv.org/pdf/2604.04334>
5. Fang, J., Lee, V. C. S., & Wang, H. (2024). Optimal service resource management strategy for IoT-based health information system considering value co-creation of users. *Industrial Management & Data Systems*, 124(3), 1132–1154. <https://doi.org/10.1108/IMDS-03-2023-0173>
6. Koduri, H. (2025). NurseSchedRL: Attention-Guided Reinforcement Learning for Nurse-Patient Assignment. <https://arxiv.org/pdf/2509.18125>
7. Kumar, Dr. S., & Sharma, P. (2024). Effective Patient Scheduling For Public Health Events Based On Deep Reinforcement Learning. *South Eastern European Journal of Public Health*, 363–366. <https://doi.org/10.70135/SEEJPH.VI.861>
8. O'Reilly, M. M., Krasnicki, S., Montgomery, J., Heydar, M., Turner, R., van Dam, P., & Maree, P. (2025). Markov decision process and approximate dynamic programming for a patient assignment scheduling problem. *Annals of Operations Research* 2025 347:3, 347(3), 1493–1531. <https://doi.org/10.1007/S10479-025-06553-4>

9. Parçaoğlu, M., Salman, F. S., & Araz, O. M. (2025). Multi-objective dynamic prioritized routing and scheduling for home healthcare services with cooperating service providers. *Health Care Management Science* 2025 28:4, 28(4), 759–786. <https://doi.org/10.1007/S10729-025-09730-W>
10. Robotics, S. W.-O. J. of, Decision, A., & 2025, undefined. (n.d.). Reinforcement learning for adaptive scheduling and optimization of healthcare staff and resources in multi-departmental hospitals. *Openscis.Com*. Retrieved May 19, 2026, from <https://openscis.com/index.php/OJRADHI/article/view/2025-05-04>
11. Saini, B., Singh, D., Sharma, K. C., & Saini, D. K. (2025). AI-enhanced telemedicine: transforming resource allocation and cost-efficiency analysis via advanced queueing model. *Scientific Reports* 2025 15:1, 15(1), 31551-. <https://doi.org/10.1038/s41598-025-15664-8>
12. Silva-Aravena, F., Morales, J., & Jayabalan, M. (2025). e-Health Strategy for Surgical Prioritization: A Methodology Based on Digital Twins and Reinforcement Learning. *Bioengineering* 2025, Vol. 12, Page 605, 12(6), 605. <https://doi.org/10.3390/BIOENGINEERING12060605>
13. Silva-Aravena, F., Morales, J., Jayabalan, M., & Sáez, P. (2025a). Optimizing MRI Scheduling in High-Complexity Hospitals: A Digital Twin and Reinforcement Learning Approach. *Bioengineering* 2025, Vol. 12, Page 626, 12(6), 626. <https://doi.org/10.3390/BIOENGINEERING12060626>
14. Silva-Aravena, F., Morales, J., Jayabalan, M., & Sáez, P. (2025b). Optimizing MRI Scheduling in High-Complexity Hospitals: A Digital Twin and Reinforcement Learning Approach. *Bioengineering* 2025, Vol. 12, Page 626, 12(6), 626. <https://doi.org/10.3390/BIOENGINEERING12060626>
15. Wang, C. H., Tian, R., Hu, K., Chen, Y. T., & Ku, T. H. (2025). A Markov decision optimization of medical service resources for two-class patient queues in emergency departments via particle swarm optimization algorithm. *Scientific Reports* 2025 15:1, 15(1), 2942-. <https://doi.org/10.1038/s41598-025-86158-w>
16. Yang, F., Li, Q. L., Zhang, C., & Wang, C. (2025). Optimal admission and queuing control with renegeing behavior under premature discharge decisions. *International Transactions in Operational Research*, 32(1), 502–529. <https://doi.org/10.1111/ITOR.13277;WGROU:STRING:PUBLICATION>
17. Zhang, J., Tutun, S., Anvaryazdi, S. F., Amini, M., Sundaramoorthi, D., & Sundaramoorthi, H. (2023). Management of resource sharing in emergency response using data-driven analytics. *Annals of Operations Research* 2023 339:1, 339(1), 663–692. <https://doi.org/10.1007/S10479-023-05702-X>
18. Zhao, L., Zhu, H., Zhang, M., Tang, J., & Wang, Y. (2026). Large-scale dynamic surgical scheduling under uncertainty by hierarchical reinforcement learning. *International Journal of Production Research*, 64(2), 719–750. <https://doi.org/10.1080/00207543.2024.2361449>
19. Zhao, R., Quan, Y., & Fan, G. (2025). Optimizing operating room scheduling through multi-level learning and column generation: a novel hybrid approach. *Health Care Management Science*, 28(4), 691–714. <https://doi.org/10.1007/S10729-025-09723-9>
20. Zhong, C., Darbandi, M., Nassr, M., Latifian, A., Hosseinzadeh, M., & Jafari Navimipour, N. (2024). A new cloud-based method for composition of healthcare services using deep reinforcement learning and Kalman filtering. *Computers in Biology and Medicine*, 172, 108152. <https://doi.org/10.1016/J.COMPBIOMED.2024.108152>
21. Zuo, J., Jin, Y., & Liu, W. (2024). Outpatient scheduling problem in smart hospital with two-agent deep reinforcement learning algorithm. *Discover Computing* 2024 27:1, 27(1), 41-. <https://doi.org/10.1007/S10791-024-09474-1>