GENERALIZED NASH EQUILIBRIUM-SEEKING FAIRNESS IN MULTIAGENT HEALTHCARE AUTOMATION *

Promise Ekpo
Cornell Tech, NY, USA
poe6@cornell.edu

Saesha Argawal
Cornell University, NY, USA
sa2388@cornell.edu

Felix Grimm
Cornell University, NY, USA
fjg45@cornell.edu

Lekan Molu
Bala-Cynwyd, PA, USA
lekanmolu@molux-labs.com

Angelique Taylor
Cornell Tech, NY, USA
amt298@cornell.edu

ABSTRACT

Enforcing a fair workload allocation among multiple agents tasked to achieve an objective in *learning-enabled* demand-side healthcare worker settings is crucial for consistent and reliable performance at runtime. Existing multi-agent reinforcement learning (MARL) approaches steer fairness by shaping reward through *post-hoc* orchestrations — leaving no certifiable *self-enforceable* fairness that is immutable by individual agents at runtime. Contextualized within a setting where each agent shares resources with others, we address this shortcoming with a learning-enabled optimization scheme among self-interested decision makers whose individual actions affect those of other agents. This extends the problem to a generalized Nash equilibrium (**GNE**) game-theoretic framework where we steer group policy to a safe and locally efficient equilibrium — no agent can improve their utility function by unilaterally changing their decisions. **Fair-GNE** models MARL as a constrained generalized Nash equilibrium-seeking (**GNE**) game, prescribing an ideal equitable collective equilibrium within the problem's natural fabric. Our hypothesis is rigorously evaluated in our custom-designed high-fidelity resuscitation simulator. Across all our numerical experiments, **Fair-GNE** achieves 168% improvement in workload balance over fixed-penalty baselines (0.89 vs. 0.33 JFI, $p < 0.01$) while maintaining 86% task success, demonstrating statistically significant fairness gains through adaptive constraint enforcement. Our results communicate our formulations, evaluation metrics, and equilibrium-seeking innovations in large multi-agent learning-based healthcare systems with clarity and principled fairness enforcement.

1 Introduction

Empirical studies of hospital teamwork among multiple individuals show that when task coordination or task automation fails, workload becomes unevenly distributed — this increases mental workload and imposes time-pressure on workers significantly (up to 40%). This unevenness often delays the execution of critical actions (up to 15 seconds on average) [1, 2, 3], which may lead to deleterious effects in mission-critical operations. Thus, enforcing fairness as a quantitative determinant of balanced workloads and efficient performance is a crucial element of the hospital teamwork automation process. The automation of collaborative decision-making in high-stakes teams within emergency healthcare applications [4, 1] demands safety and equitable fairness. This is crucial for consistent conformance with functional and allocated performance baselines: enforcing an equitable workload distribution among agents that is matched to task complexity, whilst matching workload allocation to individual agent capacity.

* *Citation*: **Promise Osaine Ekpo, Brian La, Thomas Wiener, Saesha Agarwal, Arshia Agrawal, Gonzalo Gonzalez-Pumariega, Lekan P. Molu, Angelique Taylor. Skill-Aligned Fairness in Multi-Agent Learning for Collaboration in Healthcare. Pages.... DOI:000000/11111.**

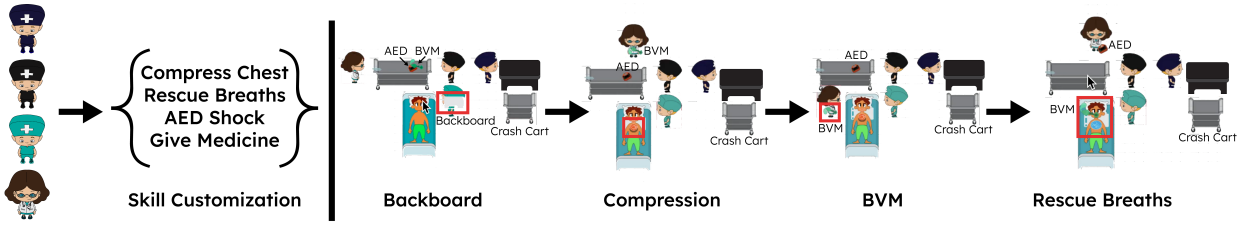

Figure 1: The **MARLHospital Environment.** The environment integrates a PDDL planner with a MARL state layer to model skill-aligned fairness and shared-task coordination among healthcare workers. The goal is to pick the backboard from the crash cart, move to the patient, place it under the patient, compress patient chest multiple times, retrieve the Bag-Valve-Mask (BVM) from the crash cart, give patient rescue breaths multiple times.

Collaborative healthcare worker (HCW) systems are often characterized by interactions among individual decision-makers — each possessing a self-interest to extremize its individual objective function. The feasible decision set by each individual may be contingent upon the (decision) strategy of other players. In these systems, each individual may be modeled as players or agents in a *game* where they compete for available shared resources. The interdependence of the various *selfish* players' decisions sets this *noncooperative* problem within a *generalized Nash equilibrium* (**GNE**) solution concept [5]. Here, objectives such as fairness and equity in resource allocation may be studied objectively.

The preeminent schemes for enforcing fairness in HCW settings in literature mostly involve multi-agent reinforcement learning (MARL). These optimize scalar rewards where the *fairness* metric is either heuristically optimized in tandem with the reward (through reward shaping) [6, 7] or evaluated *post hoc* [4, 8] to impose equity. Here, the resulting policy is not guaranteed to yield a *self-enforceable* fairness, i.e., individual agent(s) may perturb a reached equilibrium [6]. This paper introduces **Fair-GNE** as a fairness management problem among a collection of interdependent optimization problems among selfish decision-makers. In this sentiment, we embed fairness directly into the structure of the *distributed* learning process as a *shared constraint*. Each agent's feasible policy is constrained by a shared fairness condition (e.g., Jain's index), inducing coupling via a global aggregate metric rather than through explicit dependence on other agents' actions. The dynamic decision-making process by which agents reach a GNE state is the **Fair-GNE**-seeking algorithm we optimize for within the natural fabric of the equitable resource allocation problem.

We make the following key **contributions**: 1) We formulate fairness-constrained cooperative MARL as a *generalized Nash game* with workload-dependent coupling, establishing a connection between fairness and equilibrium stability. 2) We develop **Fair-GNE**, a dual-ascent algorithm that enforces a JFI-based fairness constraint during learning, ensuring convergence toward equitable equilibria under nonstationary conditions. 3) We evaluate **Fair-GNE** in a hospital-inspired multi-agent environment, demonstrating that it achieves high success rates while maintaining fairness thresholds that are higher than those of unconstrained baselines.

The remainder of this paper is organized as follows. Section 2 reviews related work in fairness and constrained MARL. Section 3 presents the Fair-GNE formulation and algorithmic details. Section 4 describes our MARL environment. Section 5 describes the experiments, and Section 6 reports empirical results and analyses, and Section 7 concludes with implications for fairness-aware multi-agent RL.

2 Related work

This section introduces an inexhaustive fairness review for fairness in MARL, constrained MARL and Nash Equilibrium:

Fairness in MARL: Prior work has sought to quantify and promote fairness in multi-agent learning, as traditional reward objectives often lead to inequitable outcomes [9]. One common strategy is to incorporate fairness directly into the agents' learning objectives, typically through intrinsic rewards or modifications to the extrinsic signal. For example, [8] introduced inequality aversion, inspired by behavioral economics, as an intrinsic reward based on pairwise differences in agents' returns, thereby improving cooperation in social dilemmas. This approach captures envy and guilt through reward shaping rather than explicit constraints. [10] formalized fairness using the Gini index as a social welfare function within a MARL framework, enabling joint optimization of efficiency, imparity, and equity. [11] proposed FAIR-PPO, which augments the PPO objective with a penalty term derived from algorithmic fairness metrics such as demographic parity, balancing reward maximization and fairness. [12] examined fairness from a welfare perspective in multi-armed bandits, introducing the notion of Nash social welfare and "Nash regret." Similar formulations have been applied to network and traffic systems, where fairness is measured through resource allocation and scheduling efficiency

[4, 13]. Additional studies have explored hierarchical or structural approaches to balance fairness and efficiency [7], penalizing workload deviation [9, 14], and enforcing team-level fairness via equivariant policies. Overall, these methods embed fairness within the optimization objective, providing soft incentives rather than guarantees of equitable outcomes.

Constrained MARL: In addition to fairness, ensuring agents follow constraints is critical for maintaining ethical or safety constraints in real-world MARL deployment. Many of the constrained MARL framework is built upon the constrained Markov decision process, where agents maximize reward subject to constraints on expected costs. [15] proposed constrained policy optimization, a trust region policy search algorithm guarantees by using surrogate functions for the objective and constraints. [16] introduced first-order constrained optimization in policy space that finds an optimal non-parametric policy and projects it back into the parametric space. This aims to reduce the surrogate approximation errors inherent in trust region methods, which rely on linearizing the constraint. These methods focus only on trajectory-wise constraints. While these methods focus on trajectory-based constraints, a critical challenge in MARL is handling state constraints across several agents. [17] addressed this need for stricter state-wise safety with a MARL framework that converges to a GNE that balances feasibility and performance. This focus on constraints is particularly vital in safety-critical domains where constraints can affect human safety, such as healthcare or autonomous vehicles. Recent work in constrained MARL includes medical resource allocation, reliable routing in medical sensor networks, surgical assistance, and medical supply chain management [18, 19, 20, 21]. Other approaches, such as in medical image segmentation, aim to adhere to constraints through reactive mechanisms such as shielding, in which a synthesized model passively monitors states. However, it can override actions if safety specifications are being violated [22].

Nash Equilibrium: A Nash equilibrium describes a stable state in a non-cooperative game in which no player can benefit from changing their strategy. The generalized Nash equilibrium (GNE) problem extends this scenario into one where each player's feasible strategy depends on the strategies of other players, which often comes from shared resources or constraints. Mamu of the research in Nash equilibrium seeks to develop distributed algorithms for finding GNEs. These algorithms are often designed for specific game structures like non-linear constraints or aggregate games with community structures [23, 24, 25]. A different thread inNash equilibrium research focuses on learning GNEs rather than computing them. [26] developed a distributed playoff-based learning algorithm for convex games with coupling constraints, where agents need to only access their own cost and constraints. This approach is increasingly being applied to distributed systems. GNEs have been used to model decentralised resource allocation and computation in the Internet of Medical Things or offloading fog computing [27, 28, 29].

3 Fairness-Constrained MARL as a Stationary Markov GNE

This section introduces our **Fair-GNE** framework for multi-agent reinforcement learning (MARL) under explicit fairness constraints. We first formalize the fairness-constrained noncooperative Markov game, then derive the associated Lagrangian dual-ascent updates. We conclude by interpreting the stationary point as a generalized Nash equilibrium (**GNE**). For clarity, all symbols are defined before first use, and the analysis assumes finite state and action spaces for theoretical results. The algorithmic framework we present afterwards is a ***GNE**-seeking* one which extends naturally to neural network function approximations with empirical validation.

3.1 Markov-Constrained Dynamic Games

Multi-agent reinforcement learning (MARL) is a special case of stochastic dynamic games, where multiple decision makers interact through a shared environment and a coupled objective. We formalize this connection using standard game-theoretic notation [25].

A discounted Markov game is defined as $\mathcal{M} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, T, r, \gamma, d)$ where $\mathcal{N} = \{1, \ldots, n\}$ denotes the set of agents, $\mathcal{S}$ the global state space, and $\mathcal{A}_i$ the local action space of agent i. The transition kernel $T : \mathcal{S} \times \prod_i \mathcal{A}_i \to \Delta(\mathcal{S})$ determines how the next state is sampled, while $r : \mathcal{S} \times \prod_i \mathcal{A}_i \to \mathbb{R}$ provides a shared instantaneous reward. The scalar $\gamma \in (0, 1)$ is the discount factor, and d denotes the initial state distribution. At each time step t, each agent samples an action $a_{i,t} \sim \pi_i(\cdot \mid s_t)$; the joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{n,t})$ is executed, and the next state follows from $s_{t+1} \sim T(s_t, \mathbf{a}_t)$.

The common expected discounted return under a stationary joint policy $\pi := \prod_i \pi_i$ is

$$J(\pi) = \mathbb{E}_{s_0 \sim d} \sum_{t=0}^{\infty} \gamma^t \, r(s_t, \mathbf{a}_t). \tag{1}$$

We consider a setting where all agents share the same objective, that is, each agent's utility equals the team return: $u_i(\pi) \equiv J(\pi)$ for all $i \in \mathcal{N}$. This assumption defines an *identical-payoff* (or *potential*) game, in which the joint

objective $J(\pi)$ serves as a common potential function. Hence, any improvement in an individual agent's policy increases the same global utility, and the team-optimal solution coincides with a (generalized) Nash equilibrium. Coupling among agents thus arises solely through the shared fairness constraint.

3.2 Problem Formulation

We consider a game $\Gamma(\mathcal{V}, \Omega, J)$ with $n \in \mathcal{V}$ players . For each player $i \in \mathcal{V}$, let pi_i be its policy subject to a local set constraint $\Omega_i \in \mathbb{R}$, with a local cost function $J_i(a_t) : \Omega \to \mathbb{R}$, (in practice, we use the same objective for all agents so that $J_i \equiv J$). It follows that the Nash equilibrium $\max_{\pi_i} \; J(\pi_i, \pi_{-i}) \; \forall i \in \mathcal{V}$ denotes the joint action of all players, whereupon no player can improve its payoff by unilaterally changing its own action. This setting naturally accommodates shared fairness coupling constraints.

To capture equitable workload distribution across agents, we augment the Markov game with a shared coupling constraint based on a differentiable fairness index. The Jain fairness index [30] is a convex, scale-invariant metric widely used in networking [31, 32] and scheduling [33], defined as

$$\mathsf{F}(w_t) = \frac{\left(\sum_{i=1}^{n} w_{i,t}\right)^2}{n \sum_{i=1}^{n} w_{i,t}^2} \in [0, 1], \tag{2}$$

and the state-based fairness constraint is

$$g(s_t) = \tau - \mathsf{F}(w_t) \leq 0, \qquad \tau \in (0, 1). \tag{3}$$

A feasible policy must therefore maintain $\mathsf{F}(w_t) \geq \tau$ throughout execution. In stochastic environments, enforcing $g(s_t) \leq 0$ at every step is intractable. Following the constrained MDP formulation [15, 34], we define a discounted cumulative constraint cost: $\bar{g}(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \, (\tau - \mathsf{F}(w_t))] \leq 0$, where the expectation is taken over trajectories induced by the stationary policy π. This relaxation replaces the hard constraint (3) with a differentiable, trajectory-level condition. The constraint function $\bar{g}(\pi)$ is smooth in the policy parameters, enabling gradient-based optimization. Suppose that each player i solves

$$\pi_i^\star \in \arg\max_{\pi_i} \; J(\pi_i, \pi_{-i}) \quad \text{s.t.} \quad \bar{g}(\pi_i, \pi_{-i}) \leq 0, \tag{4}$$

where $J(\pi_i, \pi_{-i})$ is the team objective return from defined in (1), a stationary policy profile $\pi^\star = (\pi_1^\star, \ldots, \pi_n^\star)$ is a Stationary Markov GNE for all $i \in \mathcal{N}$, $J(\pi_i^\star, \pi_{-i}^\star) \geq J(\pi_i, \pi_{-i}^\star)$ for all π_i such that $\bar{g}(\pi_i, \pi_{-i}^\star) \leq 0$.

3.3 Lagrangian and KKT Conditions

Let us introduce a nonnegative Lagrange multiplier λ for the shared fairness constraint [34]. The Lagrangian is given as

$$\mathcal{L}(\pi, \lambda) = J(\pi) - \lambda \, \bar{g}(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t, \mathbf{a}_t) - \lambda(\tau - \mathsf{F}(w_t))\big)\right]. \tag{5}$$

Under compact strategy sets and standard regularity assumptions (Slater's condition, continuity of J, boundedness of F), there exists a saddle point $(\pi^\star, \lambda^\star)$ satisfying the Karush–Kuhn–Tucker (KKT) system

$$\underbrace{\nabla_\pi \mathcal{L}(\pi^\star, \lambda^\star) = 0}_{\text{Stationarity}}, \quad \underbrace{\bar{g}(\pi^\star) \leq 0}_{\text{Primal feasibility}}, \quad \underbrace{\lambda^\star \geq 0}_{\text{Dual feasibility}}, \quad \underbrace{\lambda^\star \, \bar{g}(\pi^\star) = 0}_{\text{Complementary slackness}}. \tag{6}$$

Proposition 3.1 (KKT relationship with SM–GNE)**.** *If* $(\pi^\star, \lambda^\star)$ *satisfies* (6)*, then* $\pi^\star$ *is an SM–GNE.*

Primal–Dual update rule. We realize the KKT conditions (6) through the following coupled updates:

$$\pi^{(k+1)} \in \arg\max_{\pi} \; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t, \mathbf{a}_t) - \lambda^{(k)}(\tau - \mathsf{F}(w_t))\big)\right], \tag{7}$$

$$\lambda^{(k+1)} = \Big[\lambda^{(k)} + \eta_\lambda \, \mathbb{E}_{\pi^{(k+1)}}\left[\sum_{t=0}^{\infty} \gamma^t (\tau - \mathsf{F}(w_t))\right]\Big]_+, \tag{8}$$

where $[\cdot]_+$ projects onto $[0, \lambda_{\max}]$. These updates form the basis for the convergence analysis below.

3.4 Convergence Analysis for Finite Policy Spaces

We provide a convergence guarantee for Fair-GNE under the assumption of a finite joint policy space. Let $\mathbf{\Pi} = \Pi_1 \times \cdots \times \Pi_n$ denote the space of deterministic stationary policies for all n agents, and assume $\mathbf{\Pi}$ is finite. Each joint policy π induces a vector of long-run average workloads $\mathbf{w}^\pi$, and the fairness constraint $g(\pi) = \tau - \mathrm{JFI}(\mathbf{w}^\pi)$ defines a feasible policy set.

Following the classical Lagrangian relaxation framework, we consider the penalized objective

$$\max_{\pi \in \mathbf{\Pi}} R(\pi) - \lambda g(\pi), \tag{9}$$

where $R(\pi)$ is the long-run average reward under policy π. For fixed λ, the optimal policy $\pi^*(\lambda)$ maximizes the penalized return. We update λ via projected gradient ascent:

$$\lambda_{t+1} = \left[\lambda_t + \eta\, g(\pi_t)\right]^+. \tag{10}$$

When $\mathbf{\Pi}$ is finite and each update $\pi_{t+1} \in \arg\max_\pi R(\pi) - \lambda_t g(\pi)$ is exact, this process converges to a policy $\pi^\star$ satisfying KKT conditions:

$$g(\pi^\star) \le 0, \quad \lambda^\star \ge 0, \quad \lambda^\star g(\pi^\star) = 0. \tag{11}$$

Hence, Fair-GNE converges to a stationary policy satisfying the fairness constraint if one exists.

4 Environment Setup: MARLHospital

We evaluate Fair-GNE in the MARLHospital simulator [35], which models cooperative decision-making among healthcare workers during cardiopulmonary resuscitation (CPR). Each state s_t encodes symbolic and spatial information about the emergency room, including patient status, agent locations over six discrete stations including patient status, agent locations over six discrete stations (`cart.left`, `cart.right`, `table`, `cart.small`, `patient.legs`, `bed`), each agent's held item (`BVM`, `Backboard`, `Patient`), discrete skill levels, and binary subgoal indicators. The resulting vector has 174 binary features representing the joint symbolic–spatial configuration (station mapping details are in the Appendix). Each agent's action space $\mathcal{A}_i$ consists of eight symbolic primitives: `move`, `pick`, `place`, `stack`, `treat(patient)`, `compress_chest`, `give_rescue_breaths`, and `noop`, and the joint space scales as $|\mathcal{A}| = 8^n$ with $n = 3$ agents. The team reward is a shaped progress increment $r(s_t, \mathbf{a}_t) = H(s_t) - H(s_{t-1})$, where H measures task progress across milestones such as retrieving, placing, and using medical equipment, yielding positive reward only for medically meaningful actions. Workload counters $w_{i,t}$ accumulate validated task completions for setup or treatment, and fairness is measured by Jain's index $\mathsf{F}(w_t)$ from (2), enforcing the constraint $g(s_t) = \tau - \mathsf{F}(w_t) \le 0$ (statewise) and its discounted relaxation $\bar{g}(\pi) \le 0$ during learning. The simulator models the "Adult Basic Life Support" protocol from American Red Cross code cards [36], covering CPR and rescue-breath tasks. We focus on the *rescue-breath* goal, a long-horizon sequence involving board placement, chest compressions, and bag–valve–mask administration as seen in Figure 2 and Figure 1. All experiments use three specialised agents with different skill levels for task 3 and task 6 in Fig. 2, evaluated for 50 timesteps per episode under fairness thresholds $\tau \in [0.55, 0.85]$. Given the need for agents to cooperate to complete different tasks, this setup provides a realistic testbed for evaluating Fair-GNE's ability to enforce workload balance under shared constraints.

5 Experiments

Our experiments evaluate whether fairness can be adaptively enforced in multi-agent reinforcement learning without sacrificing task performance. Specifically, we address three questions: (1) How does fairness-constrained learning affect coordination efficiency compared to standard MARL? (2) How does adaptive constraint enforcement (Fair-GNE) compare to fixed-penalty baselines in balancing workload and success rate? (3) How reliably does Fair-GNE satisfy fairness constraints across different thresholds τ?

5.1 Baselines

We benchmark Fair-GNE against standard and fairness-aware MARL baselines implemented in EPyMARL [37] using the MARLHospital environment, which include:

- **QMIX (No Fairness)** [38]: A value-based centralized training with decentralized execution (CTDE) algorithm that factorizes the joint Q-function into individual agent Q_i values using a monotonic mixing network. The reward is purely efficiency-based with no fairness penalty.

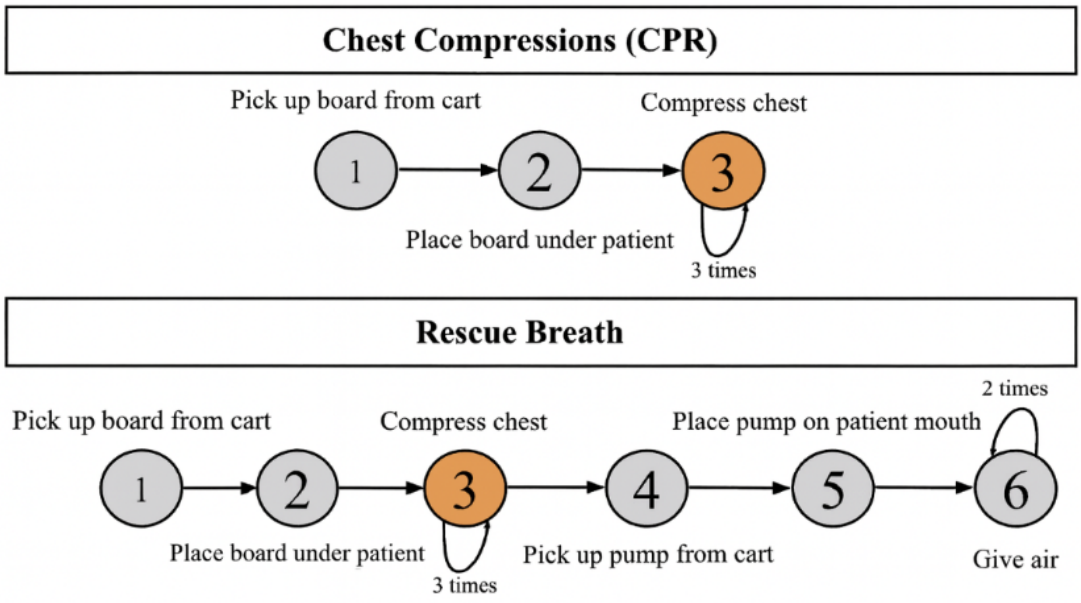

Figure 2: Task flow diagrams for the check compression (CPR) and rescue breath tasks in MARLHospital. The yellow-shaded subtask supports shared action with energy constraints.

- **FairSkillMARL (Fixed λ)** [35]: Extends QMIX with a fairness-shaped reward term encouraging balanced workload via the Gini index [11, 10]. Arbitrarily chosen fairness penalty weights $\lambda \in \{10, 50\}$ remain fixed during training, exposing the trade-off between efficiency and fairness.
- **Fair-GNE (Ours)**: Enforces fairness as a shared constraint using the GNE formulation. The fairness requirement $\mathsf{F}(w) \geq \tau$ is imposed through a Lagrange multiplier λ that adapts via dual ascent. When workload imbalance occurs ($\mathsf{F} < \tau$), λ increases to penalize unfair allocations; otherwise, it decays toward zero, allowing efficient task focus. The multiplier updates at every environment step with $\eta_\lambda = 0.01$ and a cap $\lambda_{\max} = 20$. We evaluate thresholds $\tau \in \{0.55, 0.65, 0.85\}$ to study constraint sensitivity.

5.2 Training and Evaluation

All experiments use three specialized agents in the MARLHospital CPR environment (Sec. 4). Each episode runs for up to 50 timesteps and training is repeated across four random seeds. We train QMIX for 700k timesteps. Evaluation is conducted over 50 test episodes per seed, and results are averaged across seeds. All runs use 64-core Intel Xeon CPUs with one NVIDIA GPU for parallel rollout and training.

5.3 Metrics

We report both task-level and fairness-level metrics to assess constraint satisfaction and team coordination.

- **Task Success Rate (↑)**: Fraction of episodes in which the medical task (rescue-breath) is completed successfully, reflecting coordination efficiency.
- **Workload Fairness (JFI, ↑)**: JFI ($\mathsf{F}(w)$) computed from per-agent workloads. Higher values indicate more equitable task allocation.
- **Fairness Multiplier λ**: Mean steady-state dual variable reflecting constraint tightness. Large λ indicates persistent imbalance; near-zero λ implies satisfied fairness.
- **Constraint Satisfaction Rate (↑)**: Fraction of episodes satisfying $\mathsf{F}(w) \geq \tau$. Quantifies how often fairness goals are met during training.
- **KKT Satisfaction Rate (↑)**: Fraction of episodes approximating the KKT condition $\lambda(\tau - \mathsf{F}(w)) \approx 0$. Indicates equilibrium stability between fairness enforcement and efficiency.

6 Results: Fairness–Efficiency Trade-offs with λ and τ

Table 1 presents average results over 3 random seeds, comparing Fair-GNE's adaptive constraint enforcement against fixed-penalty baselines (Gini index with $\lambda \in \{0, 10, 50\}$). Statistical significance is assessed using Welch's t-tests with Bonferroni correction ($\alpha = 0.0167$). The $\lambda = 0$ baseline is evaluated at 700k timesteps for consistency.

Convergence Validation. Fair-GNE achieves high KKT satisfaction rates ($\geq 95\%$ across all τ) and constraint satisfaction rates increasing from 88% ($\tau = 0.85$) to 100% ($\tau = 0.55$), empirically validating convergence to constrained Nash equilibria. The learned multipliers automatically adapt to workload distribution, ranging from $\lambda = 19.0 \pm 0.4$ at strict thresholds to $\lambda = 5.4 \pm 3.2$ at relaxed thresholds, demonstrating the dual ascent mechanism's responsiveness to constraint tightness.

Fair-GNE vs. Fixed Penalties. Fair-GNE substantially outperforms fixed-penalty methods in workload balance. At $\tau = 0.85$, it achieves Workload JFI of 0.89 ± 0.09 vs. 0.33 ± 0.00 for the best baseline ($\lambda = 50$), a 168% improvement ($p = 0.0082$, Cohen's $d = 8.99$). This demonstrates that adaptive constraint enforcement fundamentally outperforms static penalties. Compared to unconstrained training ($\lambda = 0$), Fair-GNE improves fairness (JFI: 0.89 vs. 0.33, $p = 0.0082$) while maintaining comparable task success (0.86 ± 0.05 vs. 0.83 ± 0.00, $p = 0.49$), showing that properly enforced fairness constraints need not sacrifice performance.

Fairness-Efficiency Trade-off. The fairness threshold τ provides principled control over the fairness-efficiency trade-off. Strict enforcement ($\tau = 0.85$) achieves maximum fairness (JFI 0.89 ± 0.09) with task success 0.86 ± 0.05. Relaxing to $\tau = 0.75$ improves efficiency (success 0.91 ± 0.06) with JFI 0.44 ± 0.19. At $\tau \in \{0.65, 0.55\}$, task success stabilizes at 0.92 with JFI of 0.52 and 0.64 respectively, all substantially exceeding fixed-penalty baselines. Even at the most relaxed threshold ($\tau = 0.55$), Fair-GNE achieves 94 % better fairness than the best fixed baseline while maintaining comparable task performance, illustrating the value of adaptive enforcement over static penalty tuning. The consistent high KKT satisfaction confirms stable equilibrium convergence across all operating points.

Adaptive Constraint Enforcement. Across all thresholds, Fair-GNE automatically adjusts λ based on constraint violations, increasing penalties when workload becomes imbalanced and reducing them when fairness targets are met. The dual ascent mechanism ensures that λ tracks constraint tightness, as validated by high KKT satisfaction rates ($\geq 95\%$) indicating proper complementary slackness. This adaptive behavior eliminates manual penalty tuning while providing controllable fairness guarantees through τ. Fair-GNE thus provides a principled framework for trading off task efficiency and workload equity in multi-agent coordination.

Statistical Testing. We evaluate statistical significance using Welch's t-tests with Bonferroni correction ($\alpha = 0.0167$ for three comparisons against $\lambda \in \{0, 10, 50\}$). All experiments use $n = 3$ independent random seeds. We report Cohen's d for effect sizes, with $|d| \geq 0.8$ indicating large effects.

Table 1: Performance comparison on the CPR task. ↑ indicates higher is better. Fair-GNE achieves a 168% improvement in workload balance (JFI) over the best baseline ($\lambda = 50$) with statistical significance (‡p<0.01, Welch's t-test with Bonferroni correction, $\alpha = 0.0167$). All results averaged over 3 independent seeds. Note: $\lambda = 0$ baseline adjusted to 700k timesteps for fair comparison.

Method	**Success (↑)**	**λ**	**Workload JFI (↑)**	**Constraint Sat. (↑)**	**KKT Sat. (↑)**
Fixed Penalty Baseline					
Gini index ($\lambda = 0$)	0.83 ± 0.00	0 (fixed)	0.33 ± 0.00	–	–
Gini index ($\lambda = 10$)	0.95 ± 0.00	10 (fixed)	0.33 ± 0.00	–	–
Gini index ($\lambda = 50$)	0.96 ± 0.01	50 (fixed)	0.33 ± 0.00	–	–
Adaptive Constraint Enforcement (Fair-GNE)					
Fair-GNE ($\tau = 0.85$)	0.86 ± 0.05	19.0 ± 0.4	**0.89 ± 0.09**‡	0.88	0.95
Fair-GNE ($\tau = 0.75$)	0.91 ± 0.06	12.4 ± 8.0	0.44 ± 0.19	0.95	0.97
Fair-GNE ($\tau = 0.65$)	0.92 ± 0.02	6.0 ± 2.1	0.52 ± 0.32	0.99	0.99
Fair-GNE ($\tau = 0.55$)	0.92 ± 0.00	5.4 ± 3.2	0.64 ± 0.30	1.00	1.00

7 Discussion and Conclusion

This work shows that fairness in cooperative multi-agent reinforcement learning can be incorporated directly into the learning process rather than imposed post hoc. By modeling workload balance as a shared constraint within a

generalized Nash equilibrium (GNE) formulation, Fair-GNE enforces fairness during training through an adaptive dual mechanism. Empirically, adaptive constraint enforcement improves workload balance by 168% over fixed-penalty baselines (JFI: 0.89 vs. 0.33, $p < 0.01$), while maintaining 86% task success. These results confirm that fairness and efficiency are not inherently conflicting objectives when fairness is embedded within the equilibrium dynamics of learning.

The approach provides a foundation for studying fairness as a controllable property in broader multi-agent systems. It offers a pathway to understanding how equilibrium constraints can regulate coordination among heterogeneous agents without introducing instability or excessive reward shaping. From a theoretical perspective, the Lagrangian formulation connects fairness-constrained equilibria with standard actor–critic optimization, highlighting a general principle for coupling efficiency objectives with fairness constraints in multi-agent learning. Future work will extend Fair-GNE to settings with multiple fairness dimensions, such as skill–task alignment or multi-objective trade-offs between workload balance and performance. Further analysis of convergence properties under function approximation and tests in human–AI collaboration scenarios remain important next steps.

Overall, Fair-GNE provides a practical solution for fair multi-agent coordination that automatically balances workload equity with task performance through adaptive constraint enforcement.

References

[1] Angelique Taylor, Tauhid Tanjim, Michael Joseph Sack, Maia Hirsch, Kexin Cheng, Kevin Ching, Jonathan St George, Thijs Roumen, Malte F. Jung, and Hee Rin Lee. Rapidly Built Medical Crash Cart! Lessons Learned and Impacts on High-Stakes Team Collaboration in the Emergency Room, February 2025. URL http://arxiv.org/abs/2502.18688. arXiv:2502.18688 [cs] version: 1.

[2] Tauhid Tanjim, Jonathan St George, Kevin Ching, and Angelique Taylor. Help or Hindrance: Understanding the Impact of Robot Communication in Action Teams, June 2025. URL http://arxiv.org/abs/2506.08892. arXiv:2506.08892 [cs].

[3] Tauhid Tanjim, Promise Ekpo, Huajie Cao, Jonathan St George, Kevin Ching, Hee Rin Lee, and Angelique Taylor. Human-Robot Teaming Field Deployments: A Comparison Between Verbal and Non-verbal Communication, June 2025. URL http://arxiv.org/abs/2506.08890. arXiv:2506.08890 [cs].

[4] Mingqi Yuan, Qi Cao, Man-On Pun, and Yi Chen. Multi-Agent Reinforcement Learning-Based Fairness-Aware Scheduling for Bursty Traffic. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, December 2021. doi: 10.1109/GLOBECOM46510.2021.9685661. URL https://ieeexplore.ieee.org/document/9685661/.

[5] Koichi Nabetani, Paul Tseng, and Masao Fukushima. Parametrized variational inequality approaches to generalized nash equilibrium problems with shared constraints. *Computational optimization and applications*, 48(3):423–452, 2011.

[6] Jiechuan Jiang and Zongqing Lu. Learning Fairness in Multi-Agent Systems, October 2019. URL http://arxiv.org/abs/1910.14472. arXiv:1910.14472 [cs].

[7] Jasmine Jerry Aloor, Siddharth Nayak, Sydney Dolan, and Hamsa Balakrishnan. Cooperation and Fairness in Multi-Agent Reinforcement Learning, October 2024. URL http://arxiv.org/abs/2410.14916.

[8] Edward Hughes, Joel Z. Leibo, Matthew G. Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas, September 2018. URL http://arxiv.org/abs/1803.08884. arXiv:1803.08884 [cs].

[9] Niko A. Grupen, Bart Selman, and Daniel D. Lee. Cooperative Multi-Agent Fairness and Equivariant Policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9350–9359, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i9.21166. URL https://ojs.aaai.org/index.php/AAAI/article/view/21166.

[10] Umer Siddique, Peilang Li, and Yongcan Cao. Fairness in Traffic Control: Decentralized Multi-agent Reinforcement Learning with Generalized Gini Welfare Functions.

[11] Gabriele La Malfa, Jie M. Zhang, Michael Luck, and Elizabeth Black. Fairness Aware Reinforcement Learning via Proximal Policy Optimization, September 2025. URL http://arxiv.org/abs/2502.03953. arXiv:2502.03953 [cs].

[12] Siddharth Barman, Arindam Khan, Arnab Maiti, and Ayush Sawarni. Fairness and Welfare Quantification for Regret in Multi-Armed Bandits, May 2022. URL http://arxiv.org/abs/2205.13930. arXiv:2205.13930 [cs].

[13] Wanqing Fang, Xintian Zhao, and Chengwei Zhang. Fairness-aware multi-agent reinforcement learning and visual perception for adaptive traffic signal control. *Optoelectronics Letters*, 20(12):764–768, December 2024. ISSN 1993-5013. doi: 10.1007/s11801-024-3267-2. URL https://doi.org/10.1007/s11801-024-3267-2.

[14] Umer Siddique. Towards Fair and Efficient Policy Learning in Cooperative Multi-Agent Reinforcement Learning. 2025.

[15] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization, May 2017. URL http://arxiv.org/abs/1705.10528. arXiv:1705.10528 [cs].

[16] Yiming Zhang, Quan Vuong, and Keith W. Ross. First Order Constrained Optimization in Policy Space, October 2020. URL http://arxiv.org/abs/2002.06506. arXiv:2002.06506 [cs].

[17] Zeyang Li and Navid Azizan. Safe Multi-Agent Reinforcement Learning with Convergence to Generalized Nash Equilibrium, November 2024. URL http://arxiv.org/abs/2411.15036. arXiv:2411.15036 [cs].

[18] Qianyue Hao, Fengli Xu, Lin Chen, Pan Hui, and Yong Li. Hierarchical Multi-agent Model for Reinforced Medical Resource Allocation with Imperfect Information. *ACM Transactions on Intelligent Systems and Technology*, 14 (1):1–27, February 2023. ISSN 2157-6904, 2157-6912. doi: 10.1145/3552436. URL https://dl.acm.org/doi/10.1145/3552436.

[19] Muhammad Shadi Hajar, Harsha Kalutarage, and M. Omar Al-Kadri. RRP: A Reliable Reinforcement Learning Based Routing Protocol for Wireless Medical Sensor Networks. In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, pages 781–789, January 2023. doi: 10.1109/CCNC51644.2023.10060225. URL https://ieeexplore.ieee.org/document/10060225. ISSN: 2331-9860.

[20] Paul Maria Scheikl, Balázs Gyenes, Tornike Davitashvili, Rayan Younis, André Schulze, Beat P. Müller-Stich, Gerhard Neumann, Martin Wagner, and Franziska Mathis-Ullrich. Cooperative Assistance in Robotic Surgery through Multi-Agent Reinforcement Learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1859–1864, September 2021. doi: 10.1109/IROS51168.2021.9636193. URL http://arxiv.org/abs/2110.04857. arXiv:2110.04857 [cs].

[21] Esha Saha and Pradeep Rathore. A smart inventory management system with medication demand dependencies in a hospital supply chain: A multi-agent reinforcement learning approach. *Computers & Industrial Engineering*, 191: 110165, May 2024. ISSN 0360-8352. doi: 10.1016/j.cie.2024.110165. URL https://www.sciencedirect.com/science/article/pii/S0360835224002869.

[22] Hanane Allioui, Mazin Abed Mohammed, Narjes Benameur, Belal Al-Khateeb, Karrar Hameed Abdulkareem, Begonya Garcia-Zapirain, Robertas Damaševičius, and Rytis Maskeliūnas. A Multi-Agent Deep Reinforcement Learning Approach for Enhancement of COVID-19 CT Image Segmentation. *Journal of Personalized Medicine*, 12(2):309, February 2022. ISSN 2075-4426. doi: 10.3390/jpm12020309. URL https://www.mdpi.com/2075-4426/12/2/309.

[23] Kaihong Lu, Guangqi Li, and Long Wang. Online distributed algorithms for seeking generalized Nash equilibria in dynamic environments, April 2020. URL http://arxiv.org/abs/2004.00525. arXiv:2004.00525 [math].

[24] Rui Huang, Yixin Gu, Yuan Fan, and Songsong Cheng. Distributed Generalized Nash Equilibria Seeking for Aggregative Games with Community Structures. In *2023 42nd Chinese Control Conference (CCC)*, pages 8137–8142, July 2023. doi: 10.23919/CCC58697.2023.10240724. URL https://ieeexplore.ieee.org/document/10240724/. ISSN: 1934-1768.

[25] Tao Li, Guanze Peng, Quanyan Zhu, and Tamer Basar. The Confluence of Networks, Games and Learning, August 2023. URL http://arxiv.org/abs/2105.08158. arXiv:2105.08158 [cs].

[26] Tatiana Tatarenko and Maryam Kamgarpour. Learning Generalized Nash Equilibria in a Class of Convex Games, October 2018. URL http://arxiv.org/abs/1703.04113. arXiv:1703.04113 [math].

[27] Zhaolong Ning, Peiran Dong, Xiaojie Wang, Xiping Hu, Lei Guo, Bin Hu, Yi Guo, Tie Qiu, and Ricky Y. K. Kwok. Mobile Edge Computing Enabled 5G Health Monitoring for Internet of Medical Things: A Decentralized Game Theoretic Approach. *IEEE Journal on Selected Areas in Communications*, 39(2):463–478, February 2021. ISSN 1558-0008. doi: 10.1109/JSAC.2020.3020645. URL https://ieeexplore.ieee.org/document/9309177.

[28] Yuxuan Yang, Xiaojie Wang, Zhaolong Ning, Joel J. P. C. Rodrigues, Xin Jiang, and Yi Guo. Edge Learning for Internet of Medical Things and Its COVID-19 Applications: A Distributed 3C Framework. *IEEE Internet of Things Magazine*, 4(3):18–23, September 2021. ISSN 2576-3199. doi: 10.1109/IOTM.0100.2000154. URL https://ieeexplore.ieee.org/document/9548976.

[29] Sungwook Kim. Learning and Game Based Spectrum Allocation Model for Internet of Medical Things (IoMT) Platform. *IEEE Access*, PP:1–1, January 2023. doi: 10.1109/ACCESS.2023.3266331.

[30] R. Jain, D. Chiu, and W. Hawe. A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems, September 1998. URL http://arxiv.org/abs/cs/9809099. arXiv:cs/9809099.

[31] Zhixiang Wei, James Yen, Jingyi Chen, Ziyang Zhang, Zhibai Huang, Chen Chen, Xingzi Yu, Yicheng Gu, Chenggang Wu, Yun Wang, Mingyuan Xia, Jie Wu, Hao Wang, and Zhengwei Qi. Equinox: Holistic Fair Scheduling in Serving Large Language Models, August 2025. URL http://arxiv.org/abs/2508.16646. arXiv:2508.16646 [cs].

[32] Mohammad Jaminur Islam and Shaolei Ren. Equity-Aware Spatial-Temporal Workload Shifting for Sustainable AI Data Centers.

[33] Ziqi Zhou, Agon Memedi, Chunghan Lee, Seyhan Ucar, Onur Altintas, and Falko Dressler. Fairness-Aware Multi-Agent Learning-based Task Offloading in Dynamic Vehicular Scenarios.

[34] Eitan Altman. *Constrained Markov Decision Processes: Stochastic Modeling*. Routledge, Boca Raton, 1 edition, December 2021. ISBN 978-1-315-14022-3. doi: 10.1201/9781315140223. URL https://www.taylorfrancis.com/books/9781315140223.

[35] Promise Osaine Ekpo, Brian La, Thomas Wiener, Saesha Agarwal, Arshia Agrawal, Gonzalo Gonzalez-Pumariega, Lekan P. Molu, and Angelique Taylor. Skill-Aligned Fairness in Multi-Agent Learning for Collaboration in Healthcare, August 2025. URL http://arxiv.org/abs/2508.18708. arXiv:2508.18708 [cs].

[36] American Red Cross. American Red Cross Code Cards, January 2025. URL https://www.redcrosslearningcenter.org/s/american-red-cross-code-cards.

[37] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks, November 2021. URL http://arxiv.org/abs/2006.07869.

[38] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning, June 2018. URL http://arxiv.org/abs/1803.11485.

[39] Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, June 2023. ISSN 00043702. doi: 10.1016/j.artint.2023.103905. URL https://linkinghub.elsevier.com/retrieve/pii/S0004370223000516.

[40] Tyler Westenbroek, Roy Dong, Lillian J. Ratliff, and S. Shankar Sastry. Competitive Statistical Estimation with Strategic Data Sources, April 2019. URL http://arxiv.org/abs/1904.12768. arXiv:1904.12768 [cs].

[41] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward Constrained Policy Optimization, December 2018. URL http://arxiv.org/abs/1805.11074. arXiv:1805.11074 [cs].

[42] Ankita Kushwaha, Kiran Ravish, Preeti Lamba, and Pawan Kumar. A Survey of Safe Reinforcement Learning and Constrained MDPs: A Technical Survey on Single-Agent and Multi-Agent Safety, May 2025. URL http://arxiv.org/abs/2505.17342. arXiv:2505.17342 [cs].

[43] V. S. Borkar and S. P. Meyn. The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, January 2000. ISSN 0363-0129. doi: 10.1137/S0363012997331639. URL https://epubs.siam.org/doi/10.1137/S0363012997331639. Publisher: Society for Industrial and Applied Mathematics.

[44] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning.

A Additional clarification on GNE and constrained optimization

Motivation for the GNE perspective. Although this is an identical-payoff game where strategic conflict is absent, the GNE framework captures a critical structural property: *coupled feasibility under aligned objectives*. The fairness constraint $\bar{g}(\pi) \leq 0$ couples all agents' feasible policy sets where agent i's constraint satisfaction depends on the policies of all other agents π_{-i}. This coupling prevents decomposition into independent single-agent problems and necessitates coordinated equilibrium-seeking. The GNE perspective provides both conceptual and algorithmic value: (1) **Conceptual:** It situates fairness-constrained MARL within the broader theory of constrained dynamic games, connecting to recent work on safe multi-agent RL [39, 25] and multi-objective coordination [40]. (2) **Algorithmic:** It justifies decentralized primal-dual algorithms where each agent independently optimizes its Lagrangian while coordinating only through a shared multiplier λ, avoiding full centralized control. This formulation emphasizes that fairness constraints induce *coupled feasibility* across agents even when objectives are aligned, distinguishing our setting from standard team optimization or unconstrained MARL.

Connection to constrained RL. Unlike conventional constrained MDPs [15, 41, 34] or single-agent safe RL [42], where safety constraints often decompose across agents, the fairness constraint here couples all agents' workloads into a shared global condition $\bar{g}(\pi) \leq 0$. This coupling transforms the problem into a generalized Nash structure, where constraint satisfaction must hold jointly across all agent policies. Recent work on multi-agent constrained RL [17] similarly adopts GNE formulations for problems with shared safety or resource constraints.

B Selected Hyperparameters for QMIX

For hyperparameter tuning in MARLHospital, we primarily conducted a grid search and performed a sweep across several hyperparameters for all the algorithms, ensuring consistency with established MARL benchmarks. The hyperparameter selection process follows a structured approach that is compatible with prior algorithms' work. The hyperparameters for each algorithm QMIX are outlined in Table 2. Key parameters include the hidden dimension size, learning rate, reward standardization, network type (e.g., GRU or fully connected networks), target update frequency, entropy coefficient (for policy gradient methods), and the number of steps used in bootstrapping. For off-policy algorithms such as QMIX, an experience replay buffer is employed to decorrelate samples and stabilise learning, following standard practices [**?**]. Onpolicy algorithms like IPPO utilise parallel synchronous workers to mitigate sample correlation and improve stability during training.

Table 2: Hyperparameters for QMIX with parameter sharing.

Hyperparameter	MARLHospital
Hidden dimension	64
Learning rate	0.001
Reward standardisation	False
Network type	GRU
Evaluation epsilon	0.1
Target update	25 (hard)

C Convergence Analysis for Finite Policy Spaces

Assumptions. We assume: (1) compact policy parameter sets and measurable Markov policies, (2) bounded r and bounded F, (3) continuity of $J(\pi)$ and $\bar{g}(\pi)$, (4) Slater condition: there exists $\tilde{\pi}$ with $\bar{g}(\tilde{\pi}) < 0$. Under these, there exists a saddle point $(\pi^\star, \lambda^\star)$ of $\mathcal{L}$ and the KKT system (6) holds.

Theoretical guarantees under idealized assumptions. We first establish convergence for the idealized case of finite (tabular) policy spaces with exact optimization in the primal step. This provides theoretical grounding for the algorithm structure.

Lemma C.1 (Descent Property of Primal–Dual Fair-GNE)**.** *Let $\pi^{(k)}$ and $\lambda^{(k)}$ denote the iterates under the updates* (7)–(8)*, with* exact *maximization in* (7)*. Under Assumptions (1)–(4) and a two-timescale rule $\eta_\pi \gg \eta_\lambda > 0$, the sequence $\{\mathcal{L}(\pi^{(k)}, \lambda^{(k)})\}$ is nonincreasing in expectation, and $(\pi^{(k)}, \lambda^{(k)})$ remains bounded almost surely.*

Proof. The proof follows from standard stochastic approximation arguments for constrained MDPs [15, 43, 17]. The primal update performs a gradient step minimizing $\mathcal{L}(\pi, \lambda^{(k)})$, while the dual update performs projected ascent on λ. Boundedness follows from the projection in (8). □

Additional regularity conditions. For completeness, we assume standard stochastic-approximation regularity: (1) $\nabla_\pi \mathcal{L}(\pi, \lambda)$ is Lipschitz continuous in (π, λ), (2) gradient noise has bounded second moments, and (3) the step sizes η_π and η_λ satisfy the two-timescale condition $\sum_k \eta_\pi = \sum_k \eta_\lambda = \infty$ and $\sum_k (\eta_\pi^2 + \eta_\lambda^2) < \infty$. Under these, the iterates track the mean ODE dynamics and converge almost surely [43].

Theorem C.2 (Convergence to Stationary Markov GNE)**.** *Under compactness, continuity, Slater's condition, and exact primal optimization, the primal–dual iterates* $(\pi^{(k)}, \lambda^{(k)})$ *generated by* (7)–(8) *converge almost surely to a saddle point* $(\pi^\star, \lambda^\star)$ *of* $\mathcal{L}$ *satisfying* (6)*. Consequently,* $\pi^\star$ *constitutes a Stationary Markov GNE (Def.* **??***).*

Corollary C.3 (Fairness satisfaction)**.** *At convergence,* $\pi^\star$ *satisfies* $\bar{g}(\pi^\star) \leq 0$*, and the complementarity condition* $\lambda^\star \, \bar{g}(\pi^\star) = 0$ *holds. Hence, the learned equilibrium policy achieves the prescribed fairness threshold* $\mathsf{F} \geq \tau$ *while optimizing collective return* $J(\pi^\star)$*.*

Remark C.1. Theorem C.2 establishes that the primal-dual structure is theoretically sound under idealized conditions. The theoretical analysis provides insight into why the algorithm should enforce fairness: the dual variable $\lambda^\star$ adapts to ensure constraint satisfaction at equilibrium.

Extension to neural network function approximation. While Theorem C.2 requires finite policy spaces and exact optimization, practical deep RL implementations use neural network parameterizations with approximate policy updates. The primal-dual algorithm structure—shaped reward training followed by dual ascent generalizes naturally to function approximation settings.

Empirically, we observe stable constraint satisfaction and consistent fairness enforcement across different random seeds and hyperparameter settings (see §5). This empirical stability aligns with observations in the constrained MDP literature, where algorithms like CPO [15] and PPO-Lagrangian [44] demonstrate reliable constraint satisfaction despite the lack of formal guarantees with neural networks. Recent work on safe deep RL [17] provides finite-sample analysis for specific function approximation schemes, which may extend to our multi-agent setting in future work.

C.1 Practical Algorithm: Primal–Dual MARL with Neural Networks

We now present the practical algorithm that instantiates the equilibrium conditions (6) for deep MARL with function approximation.

Primal step (policy update). For a fixed λ, agents maximize the Lagrangian via the shaped reward

$$\tilde{r}_t = r(s_t, \mathbf{a}_t) - \lambda \, (\tau - \mathsf{F}(w_t)). \tag{12}$$

We train the underlying MARL backbone (e.g., QMIX) on this shaped reward using standard policy gradient or value-based methods. This produces an approximate optimizer $\pi^{(k+1)} \approx \arg\max_\pi \tilde{J}(\pi; \lambda^{(k)})$ where $\tilde{J}(\pi; \lambda) = \mathbb{E}[\sum_t \gamma^t \tilde{r}_t]$ denotes the expected discounted return under the shaped reward.

The constraint $\bar{g}(\pi)$ is an expectation over trajectories, which is smooth in policy parameters under standard regularity conditions on the MDP. Specifically, the Jain fairness index $\mathsf{F}(w) = (\sum_i w_i)^2 / (n \sum_i w_i^2)$ is a rational function with well-defined gradients $\partial \mathsf{F} / \partial w_i$ for any $w \in \mathbb{R}_+^n$. For value-based methods such as QMIX [38], we do not require explicit differentiation through the constraint. Instead, the fairness signal $\mathsf{F}(w_t)$ is computed online at each timestep and incorporated directly into the shaped reward $\tilde{r}_t = r(s_t, \mathbf{a}_t) - \lambda(\tau - \mathsf{F}(w_t))$. The Q-network parameters are then updated via standard temporal difference learning using this shaped reward, with the TD loss $\mathcal{L}_{\text{TD}} = \mathbb{E}[(Q(s_t, \mathbf{a}_t) - y_t)^2]$ where $y_t = \tilde{r}_t + \gamma \max_{\mathbf{a}'} Q(s_{t+1}, \mathbf{a}')$. Although workload counters $w_{i,t}$ increment discretely, they are treated as state features observable to the agents, and no backpropagation through these counters is required.

Dual step (multiplier update). After each policy improvement phase, we estimate $\bar{g}(\pi^{(k+1)})$ via Monte Carlo rollouts and update the dual variable:

$$\lambda \leftarrow \big[\lambda + \eta_\lambda \, \bar{g}(\pi)\big]_+, \tag{13}$$

where $\eta_\lambda > 0$ is the dual learning rate and $[\cdot]_+$ denotes projection to $[0, \lambda_{\max}]$. Empirical convergence to a stationary point corresponds to satisfaction of the fairness constraint with minimal violations.

Implementation details. For value-based methods such as QMIX, we compute workload counters $w_{i,t}$ online during training so that $\mathsf{F}(w_t)$ and $g(s_t)$ are available at each timestep. We pass the shaped reward $\tilde{r}_t$ from (12) directly to the learner. This keeps the MARL backbone unchanged while enforcing fairness through the adaptive dual variable. For policy-gradient methods (e.g., IPPO, MAPPO), we support both stepwise and episodic reward shaping. The dual update (13) operates independently of the base learner, making Fair-GNE modular and compatible with existing MARL algorithms.

Hyperparameters and configuration. We provide key hyperparameter settings used throughout our experiments:

- **Dual learning rate:** $\eta_\lambda = 5 \times 10^{-4}$, approximately $100\times$ smaller than the policy learning rate to satisfy the two-timescale requirement.
- **Dual variable cap:** $\lambda_{\max} = 20.0$ to prevent runaway penalties.
- **Update frequency:** Dual variable updates occur every 50 policy gradient steps (or every 5000 environment steps for value-based methods) rather than at every timestep, improving stability.
- **Fairness threshold:** $\tau \in [0.5, 0.95]$ depending on task difficulty. Higher τ requires stricter workload balance.

This formulation connects fairness-aware MARL to the broader theory of constrained dynamic games, grounding empirical algorithms in the KKT and GNE frameworks that underpin equilibrium analysis in control and learning.

Implementation. For value based methods such as QMIX, we compute $w_{i,t}$ online so that $\mathsf{F}(w_t)$ and $g(s_t)$ are available at each step. We pass the shaped reward $\tilde{r}_t$ from (12) directly to the learner. This keeps the backbone unchanged while enforcing fairness through the dual variable.

We now give a practical algorithm that alternates between shaped–reward learning and a projected dual update. Expectations below are with respect to the trajectory distribution induced by the current joint policy.

Algorithm 1: Fair-GNE: Primal-Dual Policy Iteration for Constrained MARL

Input: Initial policy $\pi^{(0)}$; dual variable $\lambda^{(0)} \leftarrow 0$; fairness threshold $\tau \in (0,1)$; dual stepsize $\eta_\lambda > 0$; cap $\lambda_{\max} > 0$; discount $\gamma \in (0,1)$; number of rollouts M for constraint estimation.

for $k = 0, 1, 2, \ldots$ **do**

`// Primal: Policy update at fixed` $\lambda^{(k)}$

Define shaped reward: $\tilde{r}_t \leftarrow r(s_t, \mathbf{a}_t) - \lambda^{(k)}(\tau - \mathsf{F}(w_t))$;

Train MARL backbone on $\tilde{r}_t$ for N_{policy} steps to obtain $\pi^{(k+1)} \approx \arg\max_\pi \tilde{J}(\pi; \lambda^{(k)})$;

`// Constraint violation estimate via Monte Carlo`

Roll out $\pi^{(k+1)}$ for M episodes and compute:

$\bar{g}^{(k+1)} \leftarrow \frac{1}{M} \sum_{m=1}^{M} \sum_{t=0}^{T_m} \gamma^t \big(\tau - \mathsf{F}(w_t^{(m)})\big)$;

`// Dual: Projected gradient ascent on multiplier`

$\lambda^{(k+1)} \leftarrow \text{clip}\big(\lambda^{(k)} + \eta_\lambda \, \bar{g}^{(k+1)}, \, 0, \, \lambda_{\max}\big)$;

end

Output: Policy $\pi^{(K)}$ satisfying fairness constraint $\mathsf{F}(w_t) \geq \tau$ with high probability.

Summary. This methodology connects fairness-aware MARL to the broader theory of constrained dynamic games and generalized Nash equilibria. The primal-dual Fair-GNE algorithm provides: (1) theoretical grounding via KKT conditions and convergence analysis for finite policy spaces, (2) practical instantiation for deep MARL with neural function approximation, (3) automatic adaptation of penalty parameters through dual ascent, eliminating manual tuning, (4) modular integration with existing MARL algorithms via reward shaping. Our experiments in Section (5) validate that this approach achieves stable constraint satisfaction and superior fairness-efficiency tradeoffs compared to fixed-penalty baselines.