Fairness in Traffic Control: Decentralized Multi-agent Reinforcement Learning with Generalized Gini Welfare Functions

Umer Siddique, Peilang Li, Yongcan Cao

Department of Electrical and Computer Engineering
The University of Texas at San Antonio
{muhammadumer.siddique, peilang.li}@my.utsa.edu, yongcan.cao@utsa.edu

Abstract

In this paper, we address the issue of learning fair policies in decentralized cooperative multi-agent reinforcement learning (MARL), with a focus on traffic light control systems. We show that standard MARL algorithms that optimize the expected rewards often lead to unfair treatment across different intersections. To overcome this limitation, we design control policies that optimize a generalized Gini welfare function that explicitly encodes two aspects of fairness: efficiency and equity. Specifically, we propose three novel adaptations of MARL baselines that enable agents to learn decentralized fair policies, where each agent estimates its local value function while contributing to welfare optimization. We validate our approaches through extensive experiments across six traffic control environments with varying complexities and traffic layouts. The results demonstrate that our proposed methods consistently outperform existing MARL approaches both in terms of efficiency and equity.

Introduction

Efficient traffic light control is essential for improving traffic flow, reducing commute times, and preventing accidents (Ghazal et al. 2016). Studies indicate that inefficient traffic light control can increase commute times by 12-55% and exacerbate traffic congestion (Ault, Hanna, and Sharon 2019). Beyond time delays, traffic congestion also imposes significant economic costs and increases the likelihood of car accidents. For instance, in the United States, the trucking industry incurred a record cost of $94.6 billion in 2021 due to traffic congestion, as reported by the American Transportation Research Institute (ATRI)[1]. A standard method for managing traffic lights and reducing congestion is the fixed-time strategy, which operates on pre-determined schedules. While straightforward to implement, this approach suffers significant limitations, particularly in adapting to non-stationary traffic patterns or emergencies. For example, during accidents or when emergency vehicles are present, fixed-time strategies often fail, leading to increased delays and inefficiencies. This highlights the need for adaptive, data-driven methods capable of responding to real-time traffic conditions.

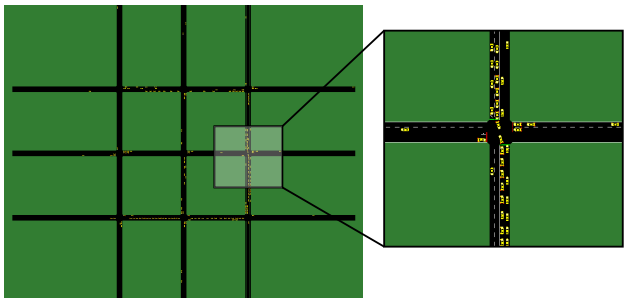


Figure 1: Example scenarios illustrating traffic imbalance: (left) a simulated 3x3 grid environment with non-uniform traffic injection across intersections and (right) a single intersection with unbalanced directional traffic flow.

Reinforcement learning (RL) has emerged as a promising approach for traffic signal control which enables agents to observe the state of the system—such as vehicle counts, queue lengths, elapsed time, and approaching vehicle speeds—and take actions that determine phase transitions (e.g., green, yellow, or red signals) at intersections (Wiering et al. 2000). The main objective of RL-based traffic control is to minimize total waiting times or cumulative delays at intersections. Despite the significant progress that has been made in both single-agent and multi-agent RL (MARL) for traffic signal control (Wiering et al. 2000; Chu et al. 2020; Wu et al. 2021; Kazemkhani et al. 2024), a critical aspect often overlooked is ensuring fairness among all road users.

Consider a 3x3 grid of intersections with varying traffic densities. In this grid, certain intersections may experience heavy congestion, while others have minimal congestion. Many simulated environments assume uniform traffic injection across intersections and use deep RL agents for optimization. However, such assumptions are far from reality, where traffic patterns are inherently non-uniform. For instance, during peak hours, intersections near downtown areas may experience significantly higher traffic volumes. Even within a single intersection, directional traffic flows can be imbalanced (see Figure 1 for an example). Standard

Multi-Agent reinforcement Learning for Transportation Autonomy (MALTA) Workshop at the AAAI Conference on Artificial Intelligence, Philadelphia, Pennsylvania, USA, 2025.

[1]https://www.freightwaves.com/news/trucking-congestion-costs-hit-record-94-6b.

RL approaches, which prioritize the minimization of overall waiting time, may allocate disproportionate green time to high-traffic lanes while neglecting others. Although this approach improves overall efficiency, it raises concerns that some vehicles may be overlooked, which could lead to unfair treatment. Therefore, to gain public trust and broaden the adoption of intelligent traffic control systems, fairness considerations are crucial.

Fairness has been explored in various forms in multi-agent systems. Early works on fair division (Beynier et al. 2019), fair mixing (Aziz, Bogomolnaia, and Moulin 2019), and fairness in non-cooperative games (de Jong, Tuyls, and Verbeeck 2008; Hao and Leung 2016) emphasized concepts such as demographic parity and proportionality but were limited to static settings without the need for learning. Recent works investigated fairness in MARL. For instance, Zhang and Shah (2014) used an egalitarian welfare function to optimize for the least advantaged agents but did not guarantee optimal outcomes for all agents. Similarly, Jiang and Lu (2019) proposed FEN, a decentralized approach combining Pareto dominance and symmetry, but it required agents to share utility information, which is not always feasible. Zimmer et al. (2021) introduced SOTO, a method that learns self-oriented and team-oriented policies for individual and collective utility optimization, respectively, but it relied on hierarchical structures and utility sharing. Within traffic signal control, fairness has been investigated to some extent. For example, Maslekar et al. (2011) developed adaptive systems to reduce outlier waiting times, while Wunderlich et al. (2008) incorporated queue lengths to enhance service quality. More recently, Wan et al. (2024) proposed a DQN-based algorithm to minimize waiting time disparities among drivers while maintaining traffic throughput. However, these methods either rely on simplified assumptions or are restricted to single-agent settings, which limits their applicability to real-world scenarios.

To address these limitations, this paper presents novel approaches designed to scale and operate in complex, real-world traffic conditions. Unlike existing fairness MARL baselines, our approaches eliminate the need for specialized network architectures or hierarchical structures. Instead of optimizing the discounted sum of rewards—a common objective in standard MARL algorithms—our methods optimize the generalized Gini welfare function to ensure equitable reward distribution among agents. By focusing on fairness within traffic light control, our work fills a significant gap where fairness considerations have been largely ignored.

Contributions. In this paper, we investigate an unresolved problem of learning fair solutions for large-scale, decentralized traffic signal control. We propose a novel MARL framework where agents independently learn policies optimized for a generalized Gini welfare function (GGF), ensuring fairness and equitable treatment across all agents. Our methods are validated across multiple traffic control environments with varying intersection densities and traffic patterns, demonstrating their scalability and adaptability. Through extensive experiments, we show that our approaches achieve fair outcomes while maintaining competitive performance against state-of-the-art MARL baselines.

Related Work

As artificial intelligence solutions are increasingly used to make decisions that affect human lives, fairness has become a crucial topic (Rawls 1971). With the widespread applications of machine learning (ML) algorithms, ensuring fairness becomes necessary to avoid harmful biases, particularly when these algorithms influence multiple end-users (Thomas et al. 2019). Numerous works have studied fairness in ML, both in supervised and unsupervised learning, using various definitions of fairness (Dwork et al. 2012; Zafar et al. 2017; Sharifi-Malvajerdi, Kearns, and Roth 2019; Agarwal et al. 2018). Among these, social welfare functions have seen increased interest, particularly in supervised learning (Cousins 2021, 2023).

In RL, defining fairness raises unique challenges due to the sequential nature of decision-making (Nashed, Svegliato, and Blodgett 2023; Wu et al. 2024). Unlike ML settings, RL requires fairness to be distributed both temporally and across agents. Recently, fairness in RL has gained significant attention, notably through early work by Jabbari et al. (2017), which introduced a fairness constraint specifically suited for Markov Decision Processes (MDPs). This work also provided a provably fair algorithm under an approximate notion of this constraint. Other approaches have focused on group fairness in online RL (Huang et al. 2022; Schumann et al. 2019), often employing fairness metrics like demographic parity. In the context of multiple objectives in RL, recent works have begun addressing fairness through social welfare function optimization (Siddique, Weng, and Zimmer 2020; Yu, Siddique, and Weng 2023a,b; Fan et al. 2022; Cousins 2022). Notably, Siddique, Weng, and Zimmer (2020) and Yu, Siddique, and Weng (2023b) proposed to learn welfare-optimal policies for multi-objective deep RL, while Cousins (2022) examined welfare objectives in a tabular setting. Fan et al. (2022) presented methods for optimizing the Nash social welfare function by utilizing its differentiability and linearizability. Siddique, Sinha, and Cao (2023) introduced FPbRL, a fairness-enhanced method in preference-based RL, to learn fair policies without relying on explicit rewards.

Building on the success of fairness in RL, fairness in multi-agent RL (MARL) has emerged as an active research area (Zhang and Shah 2014; Jiang and Lu 2019; Mandal and Gan 2022; Ju, Ghosh, and Shroff 2023; Siddique, Li, and Cao 2024). In a multi-agent MDP model, Zhang and Shah (2014) proposed a regularized maximin policy, where the regularizer balances utilitarian and max-min fairness. Jiang and Lu (2019) introduced FEN, a decentralized method using a gossip algorithm to estimate average utility, coupled with a hierarchical policy structure to choose a policy from the Pareto frontier. The closest work to ours is Zimmer et al. (2021), which proposed SOTO, a method that learns self-oriented and team-oriented policies, optimizing individual utility and social welfare functions, respectively. However, unlike FEN and SOTO, our methods do not assume the need to share individual utilities to learn a fair solution, nor does it rely on hierarchical or specialized network architectures.

Instead, our methods independently learn decentralized policies that ensure fairness through the optimization of social welfare functions.

RL and particularly MARL for traffic light control has received growing attention for its potential to optimize urban traffic systems. Early work by Mousavi, Schukat, and Howley (2017) applied deep RL agents to optimize traffic lights at a single intersection. Wei et al. (2018) proposed IntelliLight, an interpretable deep RL framework for real-time traffic signal control, while Wu et al. (2021) introduced Flow, a modular deep RL framework for managing complex traffic dynamics. Similarly, Chen et al. (2020) developed a MARL approach for controlling multi-intersection traffic systems. Other notable works include MARL algorithms for urban traffic (Wu et al. 2020) and GPU-accelerated simulators for scalable RL research (Kazemkhani et al. 2024). Despite these advances, fairness in traffic light control remains underexplored, with few exceptions. Maslekar et al. (2011) addressed fairness by designing adaptive systems to reduce extreme waiting times, while Wunderlich et al. (2008) incorporated queue lengths to improve service quality. Recently, Wan et al. (2024) proposed a DQN-based method to minimize waiting time disparities while optimizing overall throughput. However, these approaches either rely on single-agent settings or make simplified assumptions that limit their applicability in real-world scenarios. In this paper, we tackle the issue of fairness in large-scale, decentralized traffic light control. Unlike existing methods, our methods learn fully decentralized fair policies using a generalized Gini welfare function to ensure equitable treatment across diverse intersections and traffic patterns.

Preliminaries

Dec-POMDPs

Cooperative multi-agent tasks are formalized as decentralized partially observable Markov decision processes (Dec-POMDPs) (Oliehoek, Amato et al. 2016). A Dec-POMDP model is defined by the following tuple $\langle \mathcal{S}, \mathcal{U}, \mathcal{P}, \mathcal{R}, \mathcal{Z}, \mathcal{O}, N, \rho, \gamma \rangle$, where $\mathcal{S}$ describes the set of the true state of the environment, $\mathcal{U}$ represents the action space (which can be discrete or continuous), $\mathcal{P} : \mathcal{S} \times \boldsymbol{\mathcal{U}} \times \mathcal{S} \to [0, 1]$ denotes the transition function, $N = \{1, \ldots, n\}$ denotes the set of n agents, ρ represents the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor that determines the importance of future rewards. In this model, at each time step t, each agent $a \in A \equiv \{1, \ldots, n\}$ selects an action $u_t^a \in \mathcal{U}$, resulting in a joint action $\boldsymbol{u}_t = \{u_t^a\}_{a=1}^n$ in the environment. This causes a transition on the environment according to the state transition function $\mathcal{P}(s'_t \mid s_t, \boldsymbol{u}_t)$. Since we are in a fully decentralized setting, each agent receives an individual reward r_t^a, which is part of the joint reward vector $\boldsymbol{r}_t = \mathcal{R}(s_t, \boldsymbol{u}_t)$. While rewards are distributed, they are not independent among agents. Each agent's reward depends on the state and the joint actions of all agents.

We consider a partially observable scenario, which means that the agents have access only to partial observations of the environment $z_t \in \mathcal{Z}$ instead of the full state s_t, according to the observation function $\mathcal{O}(s_t, a) : \mathcal{S} \times A \to \mathcal{Z}$. The joint observation $\boldsymbol{z}_t = \{z_t^a\}_{a=1}^n$ represents the collective observations of all agents and can be referred to as the full state of the environment. Each agent has an action-observation history, which is denoted by $\tau_t^a \in T_t \equiv (\mathcal{Z} \times \mathcal{U})^t \times \mathcal{Z}$, where T_t is the set of all possible histories up to time t for each agent, and $\boldsymbol{\tau}_t = \{\tau_t^a\}_{a=1}^n$ is the set of all agents' histories. Each agent a selects its actions with a decentralized policy $u_t^a \sim \pi^a(\cdot \mid \tau_t^a)$ based only on its individual action-observation history. All agents in a team aim to learn a joint policy $\pi(\boldsymbol{u}_t|\boldsymbol{\tau}_t) \equiv \prod_{a=1}^n \pi^a(u_t^a|\tau_t^a)$ that maximizes some performance metric, such as the expected discounted return: $J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_t\right]$.

Centralized Training and Decentralized Execution

We adopt the centralized training with decentralized execution (CTDE) learning paradigm (Oliehoek, Amato et al. 2016; Sunehag et al. 2017; Foerster et al. 2018), a widely adopted approach in MARL systems. This paradigm effectively balances the advantages of having global knowledge during training with the scalability and independence needed during execution. In CTDE, centralization is exploited during the training phase while maintaining decentralization during the execution phase. In other words, during training, agents have access to the full environment state in addition to their local observation histories, and they can also share policies and experiences. This access to global state information is essential in mitigating the non-stationarity issues that often arise in dynamic multi-agent environments, where the constantly evolving system makes it difficult for agents to learn stable policies when operating independently. By considering the global state and the actions of other agents during training, the learning process becomes more stable and allows agents to converge on more effective strategies (Papoudakis et al. 2019). However, during execution, agents must operate in a decentralized manner as agents may not have access to other agents' observations or full state information. This decentralization is important in real-world applications, where access to global information is often impractical due to bandwidth constraints, latency issues, or security concerns. By adhering to this paradigm, agents trained under CTDE can adapt to environments where real-time communication is limited or costly.

The Generalized Gini Function

In this paper, we require a fair solution to satisfy three key properties: *efficiency*, *equity*, and *impartiality*. The efficiency property requires that, between two feasible solutions, if one is (weakly or strictly) preferred by all users, it should be chosen. This ensures that the solution is Pareto-optimal, meaning no agent's utility can be improved without reducing another's utility. Formally, a solution $\boldsymbol{x}$ Pareto-dominates another solution $\boldsymbol{x'}$ if $\forall i, x_i \geq x'_i$ and $\exists j, x_j > x'_j$, denoted as $\boldsymbol{x} \succ \boldsymbol{x'}$. Equity, a stronger property than efficiency, ensures that a fair solution follows the Pigou-Dalton principle (Pigou 1912; Dalton 1920), which states transferring utility from more advantaged agents to less advantaged ones, provided this transfer does not reverse their ranking. Finally, impartiality property ensures fairness by treating

identical agents equally. This property implies that permutations of the utility vector represent equivalent solutions.

To make this notion of fairness operational, social welfare functions are commonly used (Siddique, Weng, and Zimmer 2020; Fan et al. 2022; Mandal and Gan 2022; Zimmer et al. 2021; Yu, Siddique, and Weng 2023a; Siddique, Sinha, and Cao 2023) A *social welfare function*, denoted as $\phi : \mathbb{R}^n \to \mathbb{R}$, aggregates the utilities of all agents and measures how good it is in terms of social welfare. It establishes a preference or ranking over policies and the goal becomes to select a policy that maximizes this social welfare utility instead of rewards. For example, the utilitarian welfare function, defined as $\phi(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$, prioritizes aggregate efficiency but ignores equity. In contrast, the egalitarian welfare function, defined as $\phi_{-\infty}(\boldsymbol{x}) = \min_{i=1}^{n} x_i$, maximizes the utility of the least advantaged agent, potentially at the expense of overall efficiency.

The Generalized Gini Function (GGF) lies between these extremes and provides a balanced approach that incentivizes equity without disproportionately focusing on the least advantaged agents. The GGF is defined as:

$$\text{GGF}_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{x}_i^{\uparrow}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{w} \in \Delta_n$ is a fixed positive weight vector whose components are strictly decreasing (i.e., $\boldsymbol{w}_1 > \ldots > \boldsymbol{w}_n > 0$). Intuitively, by assigning higher weights to smaller agents' utilities, the GGF generated a balanced distribution of rewards, naturally encouraging fairness.

The GGF satisfies all three fairness properties. Because of positive weights, it ensures monotonicity with respect to Pareto dominance, satisfying efficiency property. The symmetry in the reordering of utilities guarantees impartiality, as identical agents are treated equally regardless of their positions in the utility vector. Lastly, its adherence to the Pigou-Dalton principle implies equity, as it is Schur-concave (i.e., favors utility redistributions that reduce disparities).

As discussed in Siddique, Weng, and Zimmer (2020), the GGF generalizes several social welfare functions by varying the weight configuration. For instance, it reduces to the egalitarian welfare function by setting $w_1 \to 1$ and $w_2, \ldots, w_n \to 0$ (Rawls 1971). A regularized egalitarian function is reduced by introducing small weights ϵ for the lower-ranked utilities. The leximin fairness approach is obtained by setting $w_i/w_{i+1} \to \infty$. Finally, equal weights reduce the GGF to the utilitarian welfare function.

Proposed Method

We consider fully cooperative MARL tasks, where a set of agents cooperate to solve a shared task. In such tasks, the final decision can impact multiple end-users. Therefore, it is crucial to consider fairness in the design of these systems to ensure their successful deployment. We propose to achieve this by optimizing a generalized Gini welfare function (GGF), which provides a balanced trade-off between utility maximization and fairness. Thus, our objective becomes to learn fair policies by maximizing the GGF, which can be formulated as:

$$\max_{\boldsymbol{\pi_\theta}} \text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\boldsymbol{\pi_\theta})), \tag{2}$$

where $\boldsymbol{\pi_\theta}$ represents the joint policy parameterized by $\boldsymbol{\theta}$, $\boldsymbol{J}(\boldsymbol{\pi_\theta}) = \mathbb{E}_{\boldsymbol{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_t\right]$ denotes the joint expected discounted returns, and $\text{GGF}_{\boldsymbol{w}}$ is the welfare function. Since we are in an independent learning setting, this optimization objective can be reformulated as:

$$\max_{\boldsymbol{\pi_\theta}} \text{GGF}_{\boldsymbol{w}}(J^1(\pi_{\theta_1}), \ldots, J^n(\pi_{\theta_n})), \tag{3}$$

where $J^a(\pi_{\theta_a})$ represents the expected return for agent a.

GGF-Based Policy Optimization

To optimize the GGF, we adopt multi-agent policy gradient methods. Our proposed framework is general and can be adapted to any policy gradient method, as it primarily modifies the optimization objective while preserving the underlying learning mechanism. The key to our proposed methods is that instead of directly optimizing expected returns, we optimize the GGF of expected returns using a variant of policy gradient theorem (Sutton et al. 2000):

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\boldsymbol{\pi_\theta})) =& \nabla_{\boldsymbol{J}(\boldsymbol{\pi_\theta})} \text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\boldsymbol{\pi_\theta}))^\top \cdot \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\pi_\theta}) \\ =& \boldsymbol{w}_\sigma^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\pi_{\boldsymbol{\theta}}),\end{aligned} \tag{4}$$

where $\nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\boldsymbol{\pi_\theta})$ is a $n \times D$ matrix representing the joint policy gradient over the n agents, $\boldsymbol{w}_\sigma$ is a weight vector sorted based on the approximated values of initial states computed by the critic, and D denotes the number of policy parameters. This weighting scheme, following GGF formulation (1), ensures equitable treatment across all agents by giving higher weights to agents with lower expected returns. The generality of our framework allows it to be seamlessly integrated with various policy gradient methods. To demonstrate this, next we explain how three popular independent MARL algorithms can be adapted to optimize GGF.

GGF-IPPO. Building on IPPO (de Witt et al. 2020), GGF-IPPO learns individual policies for agents based on local observations. Each agent maintains a local critic and computes advantages using TD(λ) estimation:

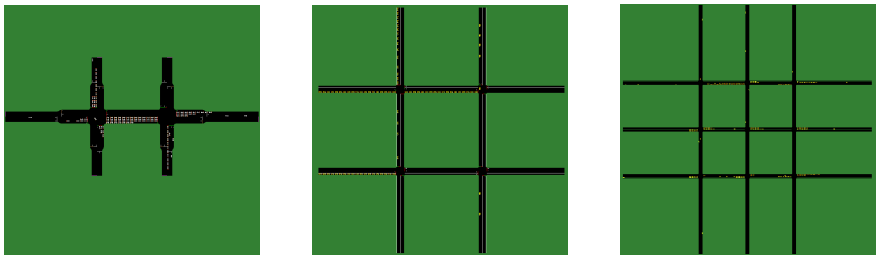
$$A^a_{\text{IPPO}} = \sum_t (\gamma\lambda)^{t-1}(r_t(z_t^a, u_t^a) + \gamma V_\theta(z_{t+1}^a) - V_\theta(z_t^a))$$

The policy update $J^a(\pi_{\theta_a})$ for each agent a derives from the policy gradient obtained from:

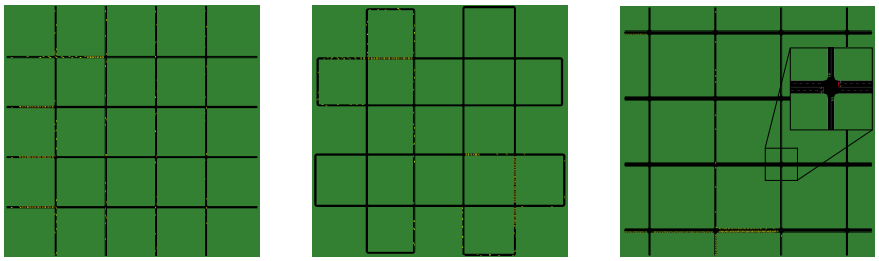
$$\mathbb{E}_{z_t^a \sim \rho_\pi, u_t^a \sim \pi_\theta(\cdot|z_t^a)}\left[\min(\rho_\theta A^a_{\text{IPPO}}(u_t^a|z_t^a), \bar{\rho}_\theta A^a_{\text{IPPO}}(u_t^a|z_t^a))\right],$$

where $\rho_\theta = \dfrac{\pi_\theta(u_t^a|z_t^a)}{\pi_{\theta_{\text{old}}}(u_t^a|z_t^a)}$, $\bar{\rho}_\theta = \text{clip}(\rho_\theta, 1-\epsilon, 1+\epsilon)$, $\pi_{\theta_{\text{old}}}$ represents the policy generating the transitions, and ϵ is a hyperparameter controlling the constraint. The GGF-IPPO policy gradient becomes:

$$\nabla_{\boldsymbol{J}_{\text{IPPO}}(\boldsymbol{\pi_\theta})} \text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}_{\text{IPPO}}(\boldsymbol{\pi_\theta}))^\top \cdot \nabla_{\boldsymbol{\theta}} \boldsymbol{J}_{\text{IPPO}}(\boldsymbol{\pi_\theta})$$



(a) Double Intersections. (b) 2x2 Grid. (c) 3x3 Grid.



(d) 4x4 Grid. (e) 4x4 Loop Grid. (f) Arterial Grid.

Figure 2: Illustration of traffic environments used in our experiments.

GGF-MAPPO. GGF-MAPPO extends MAPPO (Yu et al. 2022) by employing a centralized critic for advantage estimation while maintaining decentralized execution:

$$A^a_{\text{MAPPO}} = \sum_t (\gamma\lambda)^{t-1}(r_t(z^a_t, u^a_t) + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)).$$

The policy update $J^a(\pi_{\theta_a})$ for agent a follows:

$$\mathbb{E}_{s_t\sim\rho_\pi, u^a_t\sim\pi_\theta(\cdot|z^a_t)}\left[\min(\rho_\theta A^a_{\text{MAPPO}}(u^a_t|s_t), \bar{\rho}_\theta A^a_{\text{MAPPO}}(u^a_t|s_t))\right].$$

The final GGF-MAPPO policy gradient reformulates as:

$$\nabla_{\boldsymbol{J}_{\text{MAPPO}}(\boldsymbol{\pi_\theta})}\text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}_{\text{MAPPO}}(\boldsymbol{\pi_\theta}))^\top \cdot \nabla_{\boldsymbol{\theta}}\boldsymbol{J}_{\text{MAPPO}}(\boldsymbol{\pi_\theta}).$$

GGF-IA2C. This method uses one-step advantages with local critics as, $A^a_{\text{IA2C}} = r_t(z^a_t, u^a_t) - V_\theta(z^a_t)$. In GGF-IA2C, each actor update derives from the policy gradient obtained from:

$$J^a_{\text{IA2C}}(\pi_{\theta_a}) = \mathbb{E}_{z^a_t\sim\rho_\pi, u^a_t\sim\pi_\theta(\cdot|z^a_t)}\left[A^a_{\text{IA2C}}(u^a_t|z^a_t)\right],$$

with the GGF-modified policy gradient:

$$\nabla_{\boldsymbol{J}_{\text{IA2C}}(\boldsymbol{\pi_\theta})}\text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}_{\text{IA2C}}(\boldsymbol{\pi_\theta}))^\top \cdot \nabla_{\boldsymbol{\theta}}\boldsymbol{J}_{\text{IA2C}}(\boldsymbol{\pi_\theta}).$$

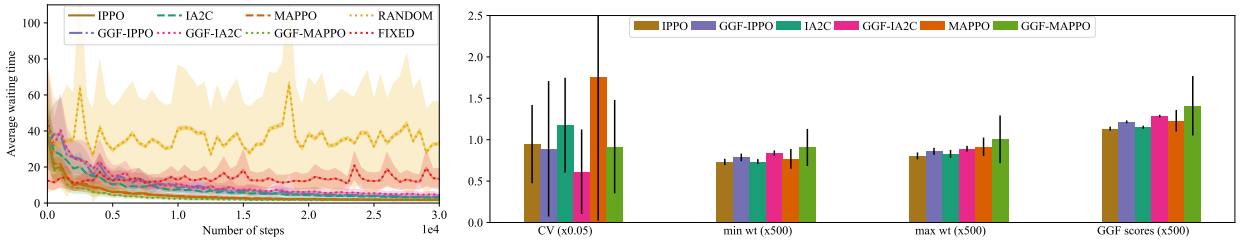
Experimental Setup and Results

To evaluate the effectiveness of the proposed methods, we run extensive experiments in six different traffic light control environments shown in Figure 2. These environments are ranked by increasing complexity and number of intersections. For GGF, we use decreasing weights $w_i = 1/2^i$ where $i \in N$. All experiments are repeated five times with different random seeds, and we report the averaged results. To simulate the traffic signals, we employed the Simulation of Urban MObility (SUMO) (Lopez et al. 2018) framework. The synthetic environments are adapted from (Alegre 2019), while realistic traffic scenario layouts are sourced from RESCO (Ault and Sharon 2021) where we inject high traffic in some intersections to mimic the real-world traffic patterns.

Environment Design. In traditional traffic control optimization problems, the main objective typically focuses on minimizing the total waiting time of all vehicles traversing intersections. However, such an objective can lead to unfair treatment of certain vehicles, particularly in real-world scenarios where traffic density varies across intersections. This disparity becomes particularly problematic as adaptive agents, trained to minimize overall waiting times, naturally prioritize high-traffic routes while potentially neglecting lower-traffic areas. Our work specifically addresses this fairness consideration in traffic light control management.

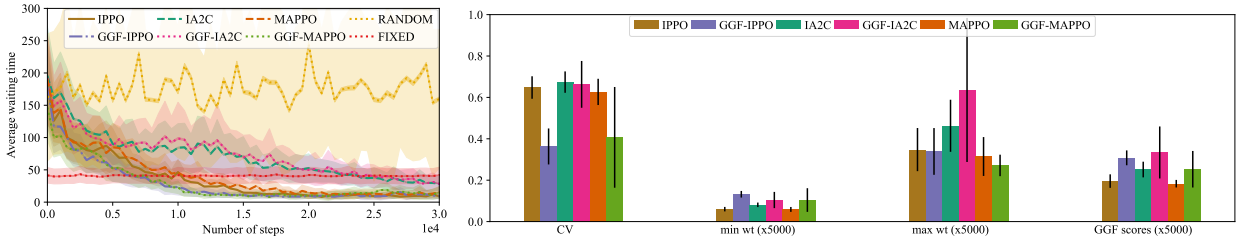
In our experiments, each intersection is modeled as an independent agent capable of controlling traffic coming from all sides of the road at an intersection. Traffic lights at an intersection are managed via four phases, with each phase specifying which lanes receive green lights. In these environments, the global state comprises the current traffic light phases, traffic density, queue lengths, and waiting times of vehicles at each intersection. Each agent's action space consists of four phase transitions that influence traffic flow. The reward for each agent is defined as the negative total waiting time at its respective intersections. Given that each intersection affects multiple road users or sides, we define fairness as achieving consistently low waiting times while ensuring equitable treatment across all sides/users. This fairness objective becomes particularly challenging in scenarios with naturally uneven traffic flow, requiring careful balancing between prioritizing less congested routes without disproportionately impacting overall system efficiency.

Experimental Environments. Our experimental setup encompasses six different traffic grid environments of increasing complexities. Our first environment is a simple environment consisting of two consecutive intersections. The 2x2 grid consists of four intersections formed by two hor-



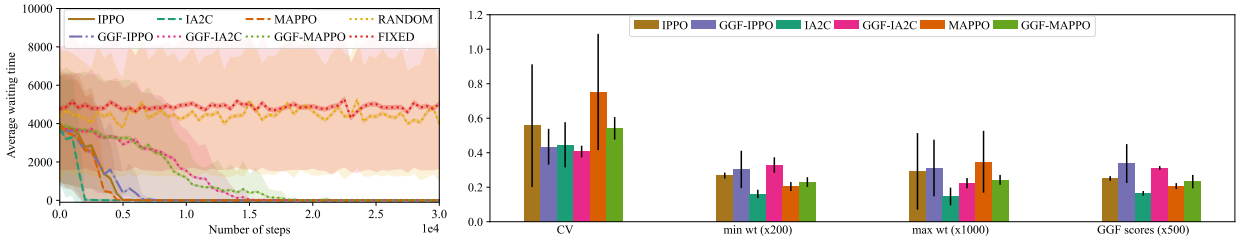
(a) Learning curves. (b) CV, min wait time, max wait time, and GGF scores.

Figure 3: Performances of IPPO, MAPPO, IA2C, and their GGF counterparts in double intersection environment.



(a) Learning curves. (b) CV, min wait time, max wait time, and GGF scores.

Figure 4: Performances of IPPO, MAPPO, IA2C, and their GGF counterparts in 2x2 grid environment.



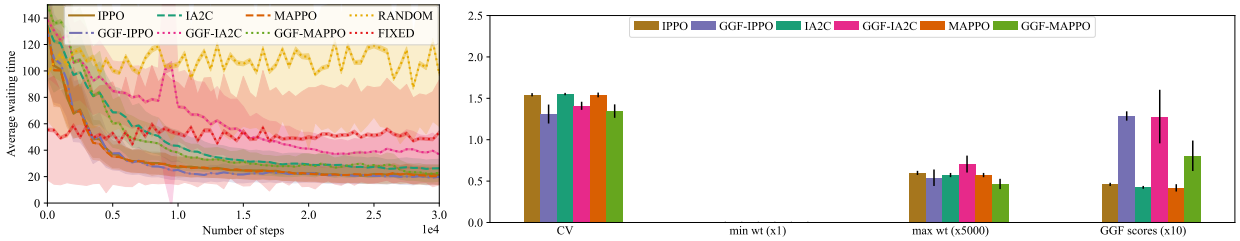
(a) Learning curves. (b) CV, min wait time, max wait time, and GGF scores.

Figure 5: Performances of IPPO, MAPPO, IA2C, and their GGF counterparts in 3x3 grid environment.

izontal and two vertical lanes. The 3x3 grid environment contains nine intersections arranged in three horizontal and three vertical intersections. The 4x4 grid and 4x4 loop contain sixteen intersections in a grid, where the latter allows vehicles to cycle recursively at the endpoint of each lane, hence generating increased traffic stress. The arterial grid features two-lane arterial horizontal streets intersecting with single-lane vertical avenues, causing unbalanced traffic flow due to turning and potential traffic congestion. The diverse set of environments provides a rich and comprehensive evaluation of our methods across varying levels of complexity and traffic patterns.

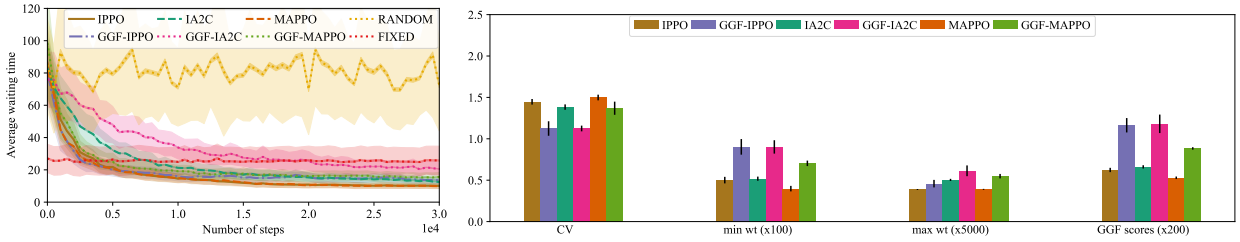
Results. Figure 3 presents the results for the double intersection environment. The learning curves (Figure 3a) show the performance of IPPO, MAPPO, IA2C, and their GGF counterparts. It also includes results for two baseline approaches: a random strategy that arbitrarily selects actions, and a fixed strategy that cycles between all phases at a predetermined frequency. As expected, the random agent performs the worst. Although the fixed strategy works better than the random agent, it is outperformed by MARL baselines and their GGF counterparts. The GGF-based methods, sometimes exhibit a higher overall waiting times as compared to their MARL algorithms. However, they achieve a more equitable distribution of waiting times as demonstrated in Figure 3b, where GGF-based methods achieve a lower Coefficient of Variation (CV), indicating more equitable waiting time distribution across intersections. Recall that, CV measures the level of dispersion in waiting times across intersections, with a lower CV indicating a fair and equitable solution. Additionally, the MARL algorithms that optimize GGF obtain higher GGF scores which further validate the successful optimization of the GGF objective.

Figure 4 depicts the results for the 2x2 Grid environment. The learning curves (Figure 4a) and bar plots (Figure 4b) show training and testing results, respectively. Once again, GGF-based methods consistently achieve lower CV scores, which indicates they achieve fair outcomes than standard MARL algorithms. Interestingly, while IA2C, IPPO, and MAPPO lower minimum and maximum waiting times, their reward distributions are much higher than GGF-based



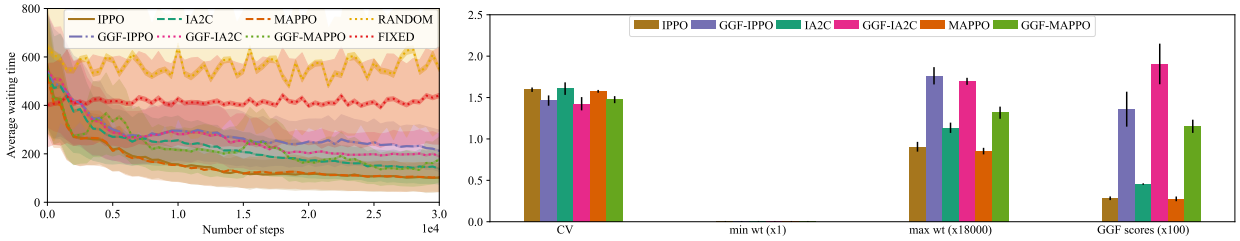
(a) Learning curves. (b) CV, min wait time, max wait time, and GGF scores.

Figure 6: Performances of IPPO, MAPPO, IA2C, and their GGF counterparts in 4x4 grid environment.



(a) Learning curves. (b) CV, min wait time, max wait time, and GGF scores.

Figure 7: Performances of IPPO, MAPPO, IA2C, and their GGF counterparts in 4x4 loop grid environment.



(a) Learning curves. (b) CV, min wait time, max wait time, and GGF scores.

Figure 8: Performances of IPPO, MAPPO, IA2C, and their GGF counterparts in arterial grid environment.

methods. That is why, they have higher CVs and lower GGF scores. On other hand, our proposed methods have higher GGF scores which indicates the effectiveness of our methods in balancing overall performance and fairness.

Similar trends are observed in more complex environments, as illustrated in Figures 5 to 8, which include 3x3 grid, 4x4 grid, 4x4 Loop, and arterial layouts. These scenarios contain a larger number of intersections, which significantly increases the complexity of fairness optimization. Notably, the 4x4 Loop and Arterial environments are challenging as they simulate realistic traffic patterns. Furthermore, these environments exhibit notable variations in traffic distribution, as demonstrated in Figures 6b and 8b. In certain cases, we observe substantial disparities in traffic density across intersections, where a subset of intersections experiences high traffic volumes while others remain relatively uncongested. Under these challenging conditions, our proposed methods demonstrate better performance, achieving higher GGF scores while effectively minimizing waiting times, as shown in Figures 6a and 8a. These results underscore the robustness of our approaches in managing heterogeneous traffic distributions while maintaining equitable outcomes across all intersections.

Conclusions and Future Work

In this work, we formalized and addressed the challenge of incorporating fairness in decentralized MARL for traffic signal control through the optimization of generalized Gini welfare functions. We proposed three novel adaptations of MARL algorithms that enable agents to independently learn policies that are both efficient and equitable. Through extensive experimental validation across six diverse traffic control environments with varying numbers of intersections and complexities, we demonstrated that our approaches can consistently achieve better performance in terms of efficiency and equity than the standard MARL methods.

Our future work includes (1) exploration of different welfare functions beyond the generalized Gini welfare functions, (2) scalability study of the proposed approaches in more complex environments with city-level infrastructure, and (3) quantitative measure and analytic study of the proposed approaches in achieving fairness.

Acknowledgements

This work was supported by the Office of Naval Research under Grant N000142412405 and the Army Research Office under Grant W911NF2310363.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*.

Alegre, L. N. 2019. SUMO-RL. https://github.com/LucasAlegre/sumo-rl.

Ault, J.; Hanna, J. P.; and Sharon, G. 2019. Learning an interpretable traffic signal control policy. *arXiv preprint arXiv:1912.11023*.

Ault, J.; and Sharon, G. 2021. Reinforcement Learning Benchmarks for Traffic Signal Control. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*.

Aziz, H.; Bogomolnaia, A.; and Moulin, H. 2019. Fair mixing: the case of dichotomous preferences. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 753–781.

Beynier, A.; Chevaleyre, Y.; Gourvès, L.; Harutyunyan, A.; Lesca, J.; Maudet, N.; and Wilczynski, A. 2019. Local envy-freeness in house allocation problems. *AAMAS*.

Chen, C.; Wei, H.; Xu, N.; Zheng, G.; Yang, M.; Xiong, Y.; Xu, K.; and Li, Z. 2020. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3414–3421.

Chu, T.; Wang, J.; Codeca, L.; and Li, Z. 2020. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*.

Cousins, C. 2021. An axiomatic theory of provably-fair welfare-centric machine learning. *Advances in Neural Information Processing Systems*, 34: 16610–16621.

Cousins, C. 2022. Uncertainty and the social planner's problem: Why sample complexity matters. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2004–2015.

Cousins, C. 2023. Revisiting fair-PAC learning and the axioms of cardinal welfare. In *International Conference on Artificial Intelligence and Statistics*, 6422–6442. PMLR.

Dalton, H. 1920. The measurement of inequality of incomes. *Economic J.*, 30(348–361).

de Jong, S.; Tuyls, K.; and Verbeeck, K. 2008. Fairness in multi-agent systems. *The Knowledge Engineering Review*, 23(2): 153–180.

de Witt, C. S.; Gupta, T.; Makoviichuk, D.; Makoviychuk, V.; Torr, P. H.; Sun, M.; and Whiteson, S. 2020. Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge? *arXiv preprint arXiv:2011.09533*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Fan, Z.; Peng, N.; Tian, M.; and Fain, B. 2022. Welfare and Fairness in Multi-objective Reinforcement Learning. *arXiv preprint arXiv:2212.01382*.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ghazal, B.; ElKhatib, K.; Chahine, K.; and Kherfan, M. 2016. Smart traffic light control system. In *2016 third international conference on electrical, electronics, computer engineering and their applications (EECEA)*, 140–145. IEEE.

Hao, J.; and Leung, H.-F. 2016. *Fairness in Cooperative Multiagent Systems*, 27–70. Springer.

Huang, W.; Labille, K.; Wu, X.; Lee, D.; and Heffernan, N. 2022. Achieving user-side fairness in contextual bandits. *Human-Centric Intelligent Systems*, 2(3): 81–94.

Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2017. Fairness in reinforcement learning. In *International conference on machine learning*, 1617–1626. PMLR.

Jiang, J.; and Lu, Z. 2019. Learning Fairness in Multi-Agent Systems. In *Advances in neural information processing systems*.

Ju, P.; Ghosh, A.; and Shroff, N. 2023. Achieving Fairness in Multi-Agent MDP Using Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

Kazemkhani, S.; Pandya, A.; Cornelisse, D.; Shacklett, B.; and Vinitsky, E. 2024. GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. *arXiv preprint arXiv:2408.01584*.

Lopez, P. A.; Behrisch, M.; Bieker-Walz, L.; Erdmann, J.; Flötteröd, Y.-P.; Hilbrich, R.; Lücken, L.; Rummel, J.; Wagner, P.; and Wießner, E. 2018. Microscopic Traffic Simulation using SUMO. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE.

Mandal, D.; and Gan, J. 2022. Socially Fair Reinforcement Learning. *arXiv preprint arXiv:2208.12584*.

Maslekar, N.; Boussedjra, M.; Mouzna, J.; and Labiod, H. 2011. VANET based adaptive traffic signal control. In *2011 IEEE 73rd vehicular technology conference (VTC Spring)*, 1–5. IEEE.

Mousavi, S. S.; Schukat, M.; and Howley, E. 2017. Traffic light control using deep policy-gradient and value-function-based reinforcement learning. *IET Intelligent Transport Systems*, 11(7): 417–423.

Nashed, S. B.; Svegliato, J.; and Blodgett, S. L. 2023. Fairness and sequential decision making: Limits, lessons, and opportunities. *arXiv preprint arXiv:2301.05753*.

Oliehoek, F. A.; Amato, C.; et al. 2016. *A concise introduction to decentralized POMDPs*, volume 1. Springer.

Papoudakis, G.; Christianos, F.; Rahman, A.; and Albrecht, S. V. 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*.

Pigou, A. 1912. *Wealth and Welfare*. Macmillan.

Rawls, J. 1971. *The Theory of Justice*. Havard university press.

Schumann, C.; Lang, Z.; Mattei, N.; and Dickerson, J. P. 2019. Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802*.

Sharifi-Malvajerdi, S.; Kearns, M.; and Roth, A. 2019. Average Individual Fairness: Algorithms, Generalization and Experiments. In *NeurIPS*. NeurIPS.

Siddique, U.; Li, P.; and Cao, Y. 2024. Towards Fair and Equitable Policy Learning in Cooperative Multi-Agent Reinforcement Learning. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*.

Siddique, U.; Sinha, A.; and Cao, Y. 2023. Fairness in Preference-based Reinforcement Learning. *arXiv preprint arXiv:2306.09995*.

Siddique, U.; Weng, P.; and Zimmer, M. 2020. Learning Fair Policies in Multi-Objective (Deep) Reinforcement Learning with Average and Discounted Rewards. In *International Conference on Machine Learning*.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. In *International Conference on Autonomous Agents and Multiagent Systems*, 2085–2087. Springer.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in neural information processing systems*.

Thomas, P. S.; Castro da Silva, B.; Barto, A. G.; Giguere, S.; Brun, Y.; and Brunskill, E. 2019. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004.

Wan, Y.; Wu, K.; Shi, T.; and Wang, J. 2024. Fair and Efficient Traffic Light Control with Reinforcement Learning. In *International Symposium on Intelligent Computing and Networking*, 17–33. Springer.

Wei, H.; Zheng, G.; Yao, H.; and Li, Z. 2018. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2496–2505.

Wiering, M. A.; et al. 2000. Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*, 1151–1158.

Wu, C.; Kreidieh, A. R.; Parvate, K.; Vinitsky, E.; and Bayen, A. M. 2021. Flow: A modular learning framework for mixed autonomy traffic. *IEEE Transactions on Robotics*, 38(2): 1270–1286.

Wu, M.; Siddique, U.; Sinha, A.; and Cao, Y. 2024. Offline Reinforcement Learning with Failure Under Sparse Reward Environments. In *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, 1–5. IEEE.

Wu, T.; Zhou, P.; Liu, K.; Yuan, Y.; Wang, X.; Huang, H.; and Wu, D. O. 2020. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Transactions on Vehicular Technology*, 69(8): 8243–8256.

Wunderlich, R.; Liu, C.; Elhanany, I.; and Urbanik, T. 2008. A novel signal-scheduling algorithm with quality-of-service provisioning for an isolated intersection. *IEEE Transactions on intelligent transportation systems*, 9(3): 536–547.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35: 24611–24624.

Yu, G.; Siddique, U.; and Weng, P. 2023a. Fair Deep Reinforcement Learning with Generalized Gini Welfare Functions. In *Adaptive and Learning Agents (ALA) Workshop*.

Yu, G.; Siddique, U.; and Weng, P. 2023b. Fair Deep Reinforcement Learning with Preferential Treatment. In *ECAI*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; Gummadi, K. P.; and Weller, A. 2017. From Parity to Preference-Based Notions of Fairness in Classification. In *NIPS*.

Zhang, C.; and Shah, J. A. 2014. Fairness in multi-agent sequential decision-making. In *Advances in neural information processing systems*.

Zimmer, M.; Glanois, C.; Siddique, U.; and Weng, P. 2021. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*.