



Adaptive Welfare Weighting for Enhanced Fairness Integration in Cooperative Multi-Agent Proximal Policy Optimization

By

Bizuhan Abate Tayu

Submitted to the School of Information Technology and Engineering in
partial fulfillment of the requirements for the degree of Master of Science in
Artificial Intelligence

Supervised by: Dr. Beakal Gizachew

College of Technology and Built Environment
Addis Ababa University
Addis Ababa, Ethiopia.

February, 2026

Approval

This is to certify that this thesis titled "**Adaptive Welfare Weighting for Enhanced Fairness Integration in Cooperative Multi-Agent Proximal Policy Optimization**" is prepared by Bizuhan Abate Tayu and submitted in partial fulfillment of the thesis-option requirements for the Degree of Master of Science in Artificial Intelligence at the School of Information Technology & Engineering, College of Technology and Built Environment.

Name	Signature	Date
<u>Dr. Beakal Gizachew Assefa</u> <i>(Advisor)</i>	_____	_____
_____ <i>(Internal Examiner)</i>	_____	_____
_____ <i>(External Examiner)</i>	_____	_____
_____ <i>(School Dean / Department Head)</i>	_____	_____

Publication Notice

The version of this thesis in manuscript form is submitted to an IEEE peer-reviewed journal for publication. The submission was completed in accordance with IEEE publication standards and guidelines, and at the time of submission of this thesis, the manuscript is currently under review and undergoing the standard IEEE editorial and peer-review process.

The submitted manuscript is derived directly from the results, analyses, and conclusions presented in this thesis, and while minor revisions may be requested during the review process, the core contributions and findings remain consistent with the content documented herein. Any updates regarding the publication status will be reflected in future revisions of this thesis, where applicable.

Abstract

Cooperative Multi-Agent Reinforcement Learning (MARL) primarily focuses on maximizing the cumulative reward to achieve high system efficiency. However, this singular focus often leads to a systemic situation where early success allows a subset of dominant agents to monopolize rewards which reflect Matthew Effect. These results long-term disparities in resource allocation among agents and the marginalization of agents. Matthew effect or rich-get-richer dynamic is particularly destructive in environments requiring sustained coordination, as it results in unfair reward distribution among agents, leading agent starvation. To address this, we propose the Weight Adaptive Fairness Integrated Proximal Policy Optimization (AW-FIPPO) algorithm. AW-FIPPO advance static fairness regularizers by introducing a dedicated Fairness Critic Network ($\mathcal{F}$). This network processes the state of the environment and agents reward distribution to generate context-aware weight vectors ($\mathbf{w}_t$). These weights are then used to dynamically modulate a hybrid objective function that blends the standard PPO clipped surrogate loss with a concave, logarithmic utility transformation, enabling the system to automatically transitioning from utilitarian maximization to egalitarian equity based on the learned, state dependent weight. Empirical results across two distinct cooperative multi-agent environments demonstrate that the adaptive weighting mechanism is the superior engine for equitable outcomes and enhanced the welfare score and minimal agent reward while maintaining competitive total return against the baselines. The improvements are significantly larger in the Matthew Effect environment, where reward accumulation bias is inherent. In this setting, our method AW-FIPPO outperforms FAPPO by 122.1% gain in total score and a 84.3% gain in max score while reducing the Gini index and the Coefficient of Variation (CV) by 24.5% and 30.6%, respectively. Additionally, our attention-based variation, AT(AW-FIPPO) compared to baseline AT-FAPPO yields a 79.5% improvement in total score and a 40.1% gain in max score, alongside reductions of 27.5% in the Gini index and 29.8% in CV.

keywords: Adaptive Fairness, Adaptive Weight, Fairness Critic Network, Generalized Gini Welfare Function, Multi-Agent Reinforcement Learning (MARL), Proximal Policy Optimization (PPO), Weight Adaptive Fairness Integrated Proximal Policy Optimization (AW-FIPPO).

Acknowledgements

Above all, I offer my deepest gratitude to **God Almighty**, whose mercy, guidance, and unfailing strength sustained me throughout my life and the completion of this work. Every step of this journey was made possible through His grace. I also give thanks to the glorious **St. Gabriel**, messenger of God and bearer of good news, for his faithful service to the Lord; through his prayers, I was granted strength of faith and courage of spirit.

I would like to express my profound gratitude to my advisor, **Dr. Beakal Gizachew**, whose insightful guidance and consistent support were invaluable throughout my master's study and research. His invaluable insights, constructive feedback, and mentorship significantly influenced the direction and quality of this work.

Here is also a special thank you for someone whose mentorship was the foundation of this work. **Mr. Befekadu Bekuretsion**, thank you for walking me through what it truly means to do research and for providing me the clarity I needed to refine my thoughts so that I can stay on track at every phase of the process. You were there from the start, and I won't forget that. You are a true example of a **teacher**.

Finally, I would love to express my deepest gratitude to **My family and Friends** for their unwavering love and patience. Thank you for listening to my endless frustrations, offering your prayers, and believing in me when I felt overwhelmed. Your constant support provided the strength and motivation I needed to see this research through to completion. I couldn't have done it without you.

Dedication

This thesis is dedicated to my father, a grounded yet devoted man who is my silent strength, and to my mama, my true reflection of unconditional love.

My dear father Abate, you taught me that love doesn't need a language; it's found in the things you do. Your love was never a spoken promise, but a lived reality. You gave us your entire self, setting aside your own comfort to ensure our well-being. You raised us with a sense of dignity and grace that transcended our circumstances, bearing the burden so we would never feel the weight. I am forever humbled by your quiet selflessness—the man who would wear the same shoes for a year just to see his child have two.

From you, I learned not only what it means to be an extraordinary father, but also what it means to be a true man and a devoted husband. You are my role model. Every day I wake up with hope and the determination to give my best to the world because of you. I promise to keep growing, to keep striving, and to make you proud because you taught me that growth never ends.

This work is also dedicated to **my dear mother, Mulualet**—a woman who truly lived up to her name by being the whole world to our family. My hero, I don't know how someone can simultaneously be a mother, a friend, and a sister, but you are everything someone needs. Thank you for your prayers, your sacrifices, and your undying commitment.

This thesis is for you both, with endless love and gratitude.

Contents

Abstract	iii
Acknowledgements	iv
Dedication	v
List of Acronyms	viii
List of Figures	x
List of Tables	xi
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Statement of the Problem	3
1.4 Research Question	5
1.5 Objective	5
1.5.1 General objective	5
1.5.2 Specific objective	5
1.6 Significance	6
1.7 Scope and Limitations	6
1.8 Contribution	7
1.9 Thesis Structure	8
2 Literature Review	9
2.1 MARL Paradigms	9
2.1.1 Centralized Training with Centralized Execution (CTCE)	12
2.1.2 Decentralized Training with Decentralized Execution (DTDE)	13
2.1.3 Centralized Training with Decentralized Execution (CTDE)	13
2.2 Fairness in MARL	14
2.3 Existing Approaches in Cooperative MARL	17
2.3.1 Utilitarian Approaches	17
2.3.2 Fairness Integrated Approaches	19

2.4	Challenges in Addressing Fairness in MARL	27
2.5	Related work	29
3	Methodology	32
3.1	Research Methodology Framework	32
3.2	Proposed Approach Description	36
3.2.1	Fairness Formulation: Generalized Gini Welfare Function (GGW) in MARL	37
3.2.2	The Novel Approach: Weight Adaptive Fairness Integrated PPO (AW-FIPPO)	39
3.2.3	The Adaptive Blended Policy Objective	39
3.2.4	The Unified Optimization Objective	40
3.2.5	Attention-Based Fairness PPO (AT-FAPPO)	40
3.3	The Proposed AW-FIPPO Architecture Detail	42
3.4	Fairness Objectives	43
3.4.1	Fairness Formulations of Baseline Paper	43
3.4.2	Fairness Formulations of This Study	44
3.5	Benchmark Environment	45
3.5.1	Random MDP	46
3.5.2	Matthew Effect	46
3.6	Evaluation Metrics	47
3.6.1	Efficiency and Utility Metrics	48
3.6.2	Disparity and Distribution Metrics	49
4	Experimentation	51
4.1	Implementation Details	51
4.2	Experimental Results	54
4.2.1	Performance Evaluation in the Random MDP Environment	54
4.2.2	Performance Evaluation in the Matthew Environment	58
4.3	Discussion	63
4.4	Ablation Study	64
4.5	Research Findings and Detailed Answers	67
5	Conclusion and future work	70
5.1	Conclusion	70
5.2	Future Work	71
	References	77

List of Acronyms

Acronym	Description
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
MARL	Multi-Agent Reinforcement Learning
DRL	Deep Reinforcement Learning
MDPs	Markov Decision Processes
PG	Policy Gradient
FCN	Fairness Critic Network
PPO	Proximal Policy Optimization
IPPO	Independent Proximal Policy Optimization
CTCE	Centralized Training, Centralized Execution
DTDE	Decentralized Training, Decentralized Execution
CTDE	Centralized Training, Decentralized Execution
FA2C	Fairness Aware Actor-Critic
FEN	Fair-Efficient Network
FSPGRL	Fairness Sensitive Policy Gradient Reinforcement Learning
FAPPO	Fairness-Aware Proximal Policy Optimization
AT-FAPPO	Attention-Based Fairness-Aware PPO
AW-FIPPO	Weight Adaptive Fairness-Integrated PPO
AT-AWFIPPO	Attention-Based Weight Adaptive FIPPO
GAE	Generalized Advantage Estimation
MLP	Multi-Layer Perceptron

List of Acronyms

Acronym	Description
Adam	Adaptive Moment Estimation
SWF	Social Welfare Function
GGWF	Generalized Gini Welfare Function
VDN	Value Decomposition Networks
COMA	Counterfactual Multi-Agent Policy Gradients
QMIX	Monotonic Value Mixing Network
CV	Coefficient of Variation
DSRP	Design Science Research Process
Dec-POMDP	Decentralized Partially Observable Markov Decision Process
IQL	Independent Q-Learning
MADDPG	Multi-Agent Deep Deterministic Policy Gradient

List of Figures

Figure 2.1:	Multi-Agent Reinforcement Learning (MARL) Overview	11
Figure 2.2:	Reward accumulation vs agent starvation and de-motivation. . .	19
Figure 3.1:	Design science research approach	33
Figure 3.2:	Architecture of our AW-FIPPO framework with decentralized poli- cies and adaptive welfare-based optimization objective	43
Figure 4.1:	Random MDP aggregated performance result illustrating efficiency (total reward, minimum reward, maximum reward, and welfare score) and disparity (coefficient of variation and Gini index). . .	54
Figure 4.2:	Random MDP welfare scores.	54
Figure 4.3:	Individual rewards across the baselines.	55
Figure 4.4:	Efficiency-equity trade-off(Random MDP): Total Reward vs. Gini index	57
Figure 4.5:	Efficiency-equity trade-off(Random MDP): Total Reward vs. CV	58
Figure 4.6:	Matthew aggregated performance results illustrating efficiency (to- tal reward, minimum reward, maximum reward, and welfare score) and disparity (coefficient of variation and Gini index).	59
Figure 4.7:	Matthew welfare scores.	60
Figure 4.8:	Efficiency-equity trade-off(Matthew): Total Reward vs. Gini in- dex	61
Figure 4.9:	Efficiency-equity trade-off(Matthew): Total Reward vs. CV . .	62

List of Tables

Table 2.1:	Comprehensive summary of of multi-agent algorithms	31
Table 3.1:	Summary: Comparative Analysis of FAPPO and Weight Adaptive PPO (AW-FIPPO)	42
Table 4.1:	Hyper-parameters and Configuration Settings for Training	53
Table 4.2:	The mean Reward distribution among individual agents	56
Table 4.3:	The absence and presence of architectural components in each variants	66
Table 4.4:	The Performance comparison of each variant	66

Chapter 1

Introduction

1.1 Background

Reinforcement Learning (RL) provides a principled framework for sequential decision-making under uncertainty, enabling agents to learn optimal behaviors through interaction with dynamic environments. With the integration of deep neural networks, RL has demonstrated remarkable success in solving high-dimensional and complex control problems. As many real-world systems involve multiple interacting decision-makers, Multi-Agent Reinforcement Learning (MARL) has emerged as a critical extension of traditional RL, allowing multiple agents to learn concurrently within shared environments.

The abilities of artificial agents to cooperate and solve complex large-scale tasks has been greatly improved by MARL. Applications span diverse domains, including traffic light systems [1] autonomous vehicles [2, 3], data center managements [4], and wireless communication networks [5]. MARL typically involve decentralized entities that must coordinate their actions while operating under partial observability and dynamic environmental conditions. Effective cooperation in such settings requires formal models capable of capturing decentralized decision-making and stochastic environmental transitions.

Cooperative multi-agent tasks are commonly formalized as a decentralized partially observable Markov decision process (Dec-POMDP) [6, 7]. A Dec-POMDP is defined by the tuple $\langle S, U, P, r, Z, O, N, \rho, \gamma \rangle$, where $s \in S$ denotes the true environment state, U represents the action space (which may be discrete or continuous), $P : S \times U \times S \rightarrow [0, 1]$ is the state transition function, $N = \{1, \dots, n\}$ denotes the set of n agents, ρ is the initial state distribution, $\gamma \in [0, 1)$ is the discount factor determining the importance of future rewards, set Z denotes the observation space, and $O : S \times N \rightarrow Z$ represents the observation function that defines the observation received by each agent given the underlying state. At each time step t , every agent $a \in \{1, \dots, n\}$ receives a local observation $z_t^a \in Z$, selects an action $u_t^a \in U$, forming the joint action $u_t = \{u_t^a\}_{a=1}^n$ [7]. The environment then transitions according to the stochastic transition function $P(s'_t | s_t, u_t)$, and agents

receive a shared team reward $r_t = r(s_t, u_t)$. This formulation captures the core challenges of cooperative MARL, including decentralized execution, partial observability and joint action coordination.

In order to enable effective coordination under this Dec-POMDP framework, wide range of algorithmic paradigm have been developed. One these widely used paradigm is centralized training with decentralized execution (CTDE), where agents utilize shared information during training while maintaining independent policies at execution time. Value factorization methods such as Value-Decomposition Network (VDN) [8] and Monotonic Value Mixing Network (QMIX) [9] decompose a global value function into individual components to facilitate coordination. Scalability and credit assignment have been further addressed by gradient-based approaches, including Independent Proximal Policy Optimization (IPPO) [10] and Counterfactual Multi-Agent Policy Gradients (COMA) [11]. Regardless of the effectiveness of these methods, their central objective is the maximization of the cumulative team reward. This reflects a utilitarian bias where the pursuit of total efficiency often comes at the expense of how those rewards are distributed among agents.

As MARL systems are increasingly considered for deployment in socially sensitive and resource-constrained domains, attention has shifted from aggregate performance to the distribution of outcomes among agents. In many practical systems, maximizing total reward does not necessarily ensure balanced or equitable outcome distribution and may comes at the expense of equity among individual agents, particularly when agents represent heterogeneous users or stakeholders. This phenomenon has motivated a growing research interest in fairness-aware reinforcement learning, where learning objectives incorporate distributive considerations such as fairness in addition to efficiency of the system.

Following this, researchers incorporated fairness consideration into the learning objective using different approaches. One of them draw inspiration from social sciences, such as utilizing the Homo Egualis model [12] to create agents that mimic human fairness behaviors in their interactions. Others incorporated it through hierarchical control architecture where a controller dynamically selects sub-policies to maximize a fair-efficient reward [13]. An other approach formalize fairness by modifying the underlying optimization problem through explicit fairness constraints into Markov Decision Processes (MDPs) using on-line convex optimization [14]. Subsequent research also incorporated protected attributes directly into the learning process to ensure that policies account for fairness-sensitive information [15]. More recent efforts embed fairness regularization within scalable Proximal Policy Optimization (PPO) algorithm [16], while welfare-driven cooperative approaches employ generalized Gini welfare function [4], which assigns proportionally larger influence to the utilities of less advantaged agents to promote equity among agents in decentralized settings [7]. Overall the gradual improvement of these frameworks reflects that the

Marl research paradigm is evolving rapidly, in which the traditional focus on efficiency is shifting towards considering fairness integration and equitable reward distribution.

1.2 Motivation

The motivation of this research is driven from the aim of harmonizing the equity-efficiency tradeoff by exploring how fairness can seamlessly be incorporated into cooperative MARL frameworks while preserving or even improving the collective performance. While MARL systems primarily aim to maximize the cumulative reward, this focus on cumulative system reward often comes at the expense of fair reward distribution among agents. In other word fewer agents may monopolize the system reward, leaving the rest of the agents severely disadvantaged or even starved. For example, agents who initially gain advantages may continue to accumulate rewards and get disproportionate benefits, showing Matthew Effect scenario, while the rest of the agents are left behind, resulting in inequitable outcomes under cooperation. Previous approaches, such as Fairness-aware Proximal Policy Optimization (FAPPO) [7], addressed this by optimizing the generalized gini welfare function, but they relied on a static, manually tuned weight vector (ω). This non-adaptive constraint places a rigid trade-off, as the policy’s sensitivity to inequity remains constant regardless of the environment’s current state or the severity of disparities among agents.

Based on the inspiration taken from the law of diminishing returns in economics, the central motivation for this study is therefore to overcome this rigidity by learning a context-aware fairness objective. By introducing a fairness critic neural network, the system can strategically modulate the welfare weights ($\mathbf{w}_t$) based on the environment state and agent reward, allowing it to maximize collective return when agents are fairly rewarded, yet instantly enforce equity when disparity occurs, thereby achieving state-dependent equilibrium that outperforms the capabilities of any static approach.

1.3 Statement of the Problem

In cooperative MARL, agents learn policies that collectively maximize a global objective, typically performance with respect to the total (or average) of per-user efficiency measure (e.g., waiting times of cars in traffic light control, safety of different sites in patrolling, or throughput of users in internet congestion) [17]. However, this unconstrained maximization of mutual rewards empirically yields unfair outcomes for individual team members [18], where some agents contribute disproportionately by shouldering more responsibility while others are left with limited opportunities to benefit.

This imbalance not only undermines fairness but may also cause resentment or demotivation among agents with little to no rewards, leading to inefficiencies in long-term

cooperation and undermining the ethical principles of fairness. For instance, when one agent performs a rewarding action while another act unhelpfully, the shared reward signal cannot distinguish which agent’s action contributed positively or negatively to the outcome [19]. Consider what happens when one agent takes a rewarding action and accumulate rewards while another agent act unhelpfully and starve for rewards [19].

Existing approaches that aims to address this issue, such as FAPPO and its attention-based variant (AT-FAPPO) [7], incorporate fairness by optimizing generalized gini welfare (GGW) function which is formulated in their study as:

$$\phi_{\omega}(x) = \sum_{i=1}^n \omega_i x_i^{\uparrow} \quad (1.1)$$

where $x \in \mathbb{R}^n$ represents the vector of individual utilities and $\omega \in \Delta_n$ is a fixed positive weight vector with strictly decreasing components ($\omega_1 > \omega_2 > \dots > \omega_n > 0$). For prioritizing the worse-off agents by assigning larger weights to them, [7] optimized Equation (1.1) [7, 20]. In their method [7], however, the welfare weights (ω) used for the function are fixed and predefined as shown in Equation (1.2) [7], representing static fairness parameters that remain unchanged throughout training.

$$\omega_i = \frac{1}{2^i}, \quad i = 0, \dots, N - 1 \quad (1.2)$$

Both variants of [7] which are FAPPO and AT-FAPPO update the policy network by incorporating those fixed fairness weights ω_i (1.2) into the clipped PPO surrogate objective. This was achieved by computing the clipped PPO surrogate objective for each agent and multiplying it by its corresponding fairness weight. Mathematically [7] formulated it in their implementation as:

$$\mathcal{L}_{\text{actor}} = - \sum_{a=1}^N \omega_a \min \left(r_a A_a, \text{clip}(r_a, 1 - \epsilon, 1 + \epsilon) A_a \right) \quad (1.3)$$

where N is the number of agents, ω_a is the sorted version of (1.2) assigned to agent a , $r_a = \frac{\pi_{\theta}}{\pi_{\theta_{\text{old}}}}$ is the PPO probability ratio for agent a , A_a is the advantage of agent a , and ϵ is the PPO clipping parameter. Although [7] introduced some level of fairness using formulation in (1.3), both variants of their method lacks adaptability to the changing agents reward distribution or environmental states due to the static nature of the weight ω_i in (1.3) [7]. Therefore, the core problem addressed by this research lies in the inherent rigidity and lack of adaptability within [7] method. The critical methodological bottlenecks are:

1. Fixed Fairness Constraint: These method rely on a static, manually tuned weight vector (ω), where the weights are fixed before training begins which is (1.2)

2. Lack of Contextual Sensitivity: This fixed constraint imposes a constant marginal penalty for inequity across all episodes and all environment states. Following this the policy’s trade-off between maximizing system wide efficiency and enforcing equity remains static.

Although [7] enforce fairness through welfare-based optimization to sustain the involvement of all agents in dynamic environments, however they typically relied on static fairness objectives that do not adapt to different environment state or varying agent reward distribution. This static reliance directly affects how the policy balances the trade-off between maximizing overall system efficiency and enforcing equity among agents. Because the weights used in the optimization of GGW function [7] (1.2) does not adapt to changing reward distributions or environmental dynamics, the policy may either enforce equity in a way that limits overall efficiency or overemphasize collective performance at the expense of fairness. Based on these fundamental limitation, we propose a Weight Adaptive Fairness Integrated PPO (AW-FIPPO) and its attention based variant AT-(AW-FIPPO) framework inspired from the law of diminishing return in economics, through which fairness is treated as an adaptive, state-dependent priority rather than a fixed constraint.

1.4 Research Question

Based on the established problem, the research question driving this study is formulated as follows:

- RQ1: How can an auxiliary fairness critic neural network be designed to observe global state transitions and reward distributions to produce real-time weight vectors?
- RQ2: How effectively does the designed AW-FIPPO and AT(AW-FIPPO) artifact perform against existing static-constraint baselines in terms of utility and equity?

1.5 Objective

1.5.1 General objective

The primary general objective is to design, implement, and validate a novel adaptive fairness mechanism that resolves the rigidity of static welfare constraints in cooperative MARL.

1.5.2 Specific objective

- To conduct a comprehensive review of existing approaches in fairness-aware MARL.

- To design and implement the auxiliary Fairness Critic Network ($\mathcal{F}$).
- To develop the AW-FIPPO algorithm by incorporating the auxiliary network into the actor-critic structure.
- To extend the mechanism to the communication-enabled variant, AT(AW-FIPPO).
- To empirically evaluate AW-FIPPO and AT(AW-FIPPO) algorithms against standard MARL baselines.

1.6 Significance

The significance of this research lies in its dual contribution of advancing the theoretical integrity of MARL fairness and enabling the practical deployment of fairness integrated multi-agent systems. From the theoretical aspect, our approach challenges traditional methods that rely on static constraints for integrating fairness considerations into the system. This static reliance often affects how the policy balances the trade-off between maximizing overall system efficiency and enforcing equity among agents. By introducing the Fairness Critic, this thesis provides a mechanism that treats fairness as a dynamic and learnable welfare weight instead of a static constraint. This allows the policy to maximize the system objective as well as consider fairness within the system based on the learned welfare weight. Specifically, through adaptive Gini weight integrated into the policy objective to condition the learning, we provide a mathematically grounded method that balances collective efficiency with distributive equity.

From the practicality aspect, our approach provides solutions for complex and resource-constrained environments where unequal access to rewards or learning opportunities is a systemic challenge. In these contexts, structural inequity often forces a choice between high efficiency and fair distribution. Therefore, for problems with inherent systemic inequality, for instance, for applications like traffic management, resource allocation, and decentralized coordination, our approach provides a solution by detecting the system state and amending the learning based on it. By mitigating agent neglect while preserving collective efficiency, our research paves the way for cooperative systems that are both high-performing and socially sustainable for complex, real-world deployment.

1.7 Scope and Limitations

The scope of this study is constrained to replacing the static fairness integration into cooperative MARL, specifically within the robust, decentralized Independent Proximal Policy Optimization (IPPO) framework by a context aware approach. These includes designing, implementing, and evaluating the proposed approach that introduce Fairness

Critic Network as an extension of [7]’s algorithm. Our approach exclusively focus on optimizing weight adaptive welfare function in fully cooperative settings to impose an outcome-based definition of fairness (equity in rewards). The empirical evaluation is constrained to synthetic cooperative MARL environments designed by [7] to highlight the tradeoff between efficiency and equity.

The study’s primary limitations include the high computational cost of the extended PPO architecture due to multiple optimization passes and the inclusion of the third neural network, currently requires extensive training time. Furthermore, the reliance of the current Fairness Critic Network on episodic returns introduces a temporal lag in the fairness objective, meaning it only reacts to the episodic disadvantages rather than instantaneous per-step unfairness.

1.8 Contribution

This study present a mechanism that learns welfare function weights dynamically through a dedicated Fairness Critic Network. This advancement replaces the reliance on pre-defined, static parameters like the one used by [7]’s fixed gini weights, thereby resolving a fundamental limitation and providing the first significant approach for an adaptive, state-dependent fairness objective in PPO-based MARL. By allowing fairness weights to evolve according to the environment and agent states, the model captures context-specific trade-offs between efficiency and equity in real time. The proposed method, Weight Adaptive Fairness Integrated PPO (AW-FIPPO) and its attention based variant AT-(AW-FIPPO) algorithm integrates the robust training stability of PPO with the proposed dynamic welfare objective. This creates a fairness-aware optimization process that adapts continuously during learning. This study offers several key contributions to the advancement of fairness-integrated cooperative MARL, which are summarized as follows:

- We formulate fairness-aware multi-agent reinforcement learning as a dynamic welfare optimization problem and introduce a Fairness Critic Network ($\mathcal{F}$) that adaptively learns welfare function weights. In contrast to state-of-the-art approaches such as [7], which employ fixed Generalized Gini parameters, our formulation allows welfare weights to be adjusted based on the state of the environment and the observed reward distribution during training.
- Based on this formulation, we propose the Weight Adaptive Fairness Integrated PPO (AW-FIPPO) algorithm, which integrates adaptive welfare weighting into the PPO objective. Unlike state-of-the-art fairness-aware methods with static objectives like FAPPO, AW-FIPPO enables fairness sensitivity to evolve over time while maintaining PPO’s empirical stability and convergence behavior.

- We further extend the framework to an attention-based variant, AT(AW-FIPPO), which incorporates self-attention to model inter-agent dependencies and coordination. Compared to state-of-the-art attention-based methods, the proposed approach replace the dependence on static welfare parameters and supports a more flexible fairness–efficiency tradeoff.
- Experimental results under the considered environments and evaluation settings indicate that the proposed approach achieves consistent improvements over a representative state-of-the-art baselines [7–11]. Specifically, AW-FIPPO achieves 122.1 % increase in Total Reward, 24.5 % reduction in Gini Index, and 30.6 % reduction in Coefficient of Variation (CV), and also AT(AW-FIPPO) achieves 79.5% gain in total reward, reduce the Gini index by 27.5 % and CV by 29.8 % over the baselines.

1.9 Thesis Structure

This study is organized into five chapters. Following the introductory chapter, Chapter 2.1 reviews the necessary background on MARL, existing fairness literature, and challenges associated with them. Chapter 3 describes the methodology of the original FAPPO algorithm and formally introduces WA-FIPPO architecture, including the structure and optimization of the Fairness Critic Network, hyperparameter configurations used, detailing the environments, as well as the experimental setup. Chapter 4 presents the comprehensive experimental results, provides an in-depth analysis of the performance gains in both efficiency and equity. Finally, Chapter 5 concludes the thesis with the summary of key findings, contributions of the thesis, and suggestions for future research directions.

Chapter 2

Literature Review

2.1 MARL Paradigms

Multi-Agent Reinforcement Learning (MARL) builds upon deep reinforcement learning (DRL) and generalizes the standard Markov Decision Process (MDP) into Stochastic Games (or Markov Games) in which reward signals and state transitions are functions of the joint action space of all participating agents [21]. By providing a framework in which multiple agents interact within a shared environment, MARL aims to optimize agents utility or a joint objective [17]. In other words, MARL provide a learning framework where multiple agents learn policies through interactions with the environment and each other. The agents aim to optimize their respective rewards under cooperative, competitive, or mixed settings.

Since each agent is learning and optimizing its own policy, the environment appears to be in constant flux from the perspective of any single agent. This constant flux within the environment introduce a core difficulties which stems from non-stationarity. The challenge becomes worse when considering the partially observable situation and the fact that agents must coordinate their actions without central control, which further complicates the problem space [6]. With all these challenges, MARL strength is placed on its ability to represent collaborative and competitive behaviors within a unified framework. This versatility has made it an invaluable tool for addressing a wide range of distributed system challenges including decentralized resource allocation, coordinating teams of autonomous robots, and infrastructure control [1, 4].

The learning approaches in MARL typically classified into three main categories, each defined by how agents share information and make decisions during training and execution [22]. The categories are centralized training with centralized execution (CTCE), decentralized training with decentralized execution (DTDE), and centralized training with decentralized execution (CTDE) [22]. All of these paradigms have their own limitations, as they try to achieve one objective at the expense of an other objectives. This

objectives includes coordination, scalability, communication overhead, and suitability for real-world applications. The choice of the learning paradigm directly impact the algorithmic strategies.

Multi-Agent Reinforcement Learning (MARL)

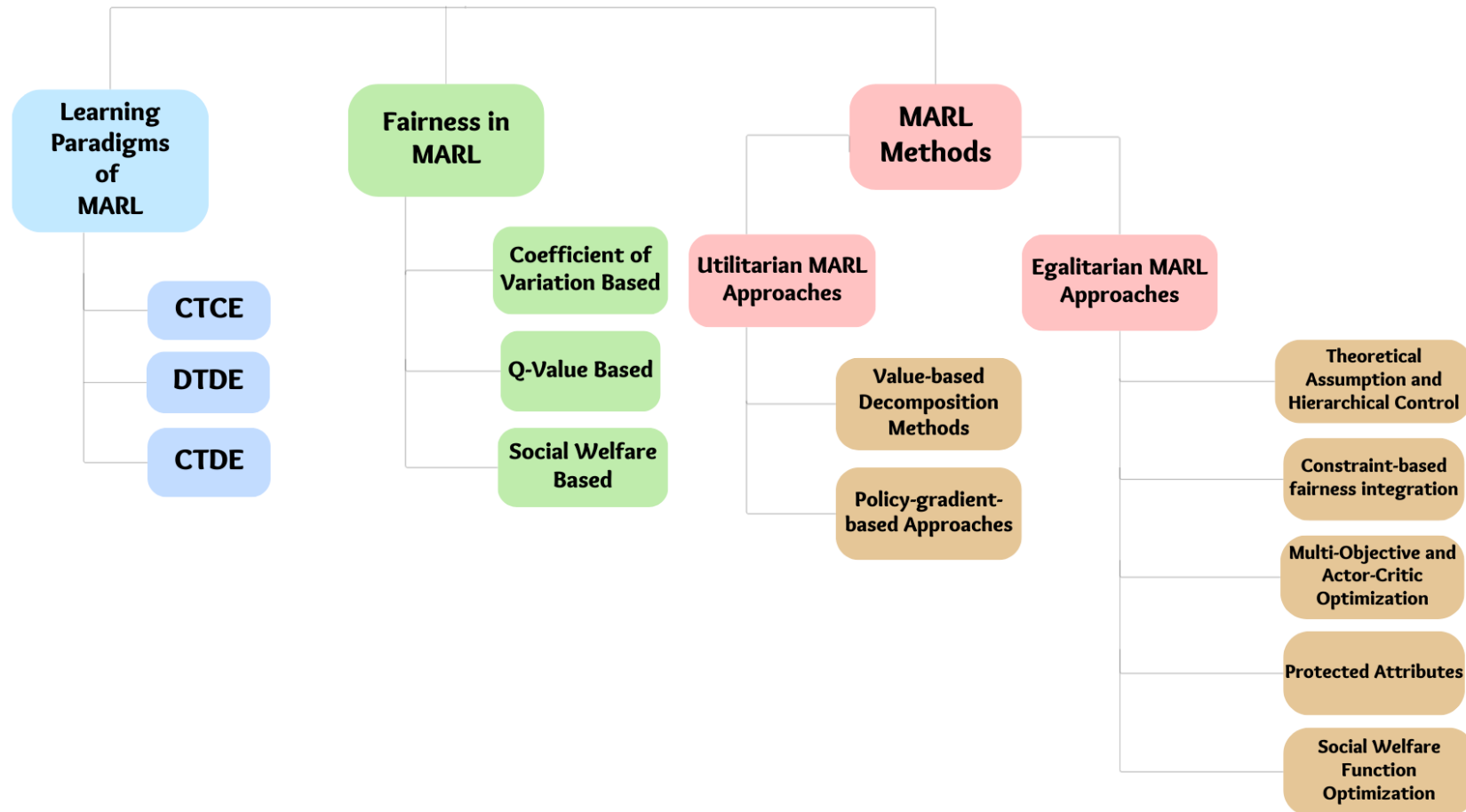


Figure 2.1: Multi-Agent Reinforcement Learning (MARL) Overview

As illustrated in the Figure 2.1, our literature review provides a hierarchical taxonomy that stems from the Main MARL root and branches into sub fundamental pillars. These pillars include learning paradigms of MARL, fairness in MARL and approaches or methods in MARL. The learning Paradigms branch classify the learning by execution and training regimes (into three classes as CTCE, DTDE, and CTDE). The second pillar fairness in MARL identifies the specific metrics studies used to assess equity and analyzes fairness perspectives through the lenses of Q-Value, Social Welfare, and Coefficient of Variation. The third pillar methods or MARL approach divide the research approach into egalitarian category which focuses on distributive justice and utilitarian category which focuses on collective reward maximization. These whole structured pillar maps the progression of MARL from high-level theoretical frameworks to the specialized methodologies and optimization criteria that define the current state of the art in agent coordination and equitable resource allocation.

2.1.1 Centralized Training with Centralized Execution (CTCE)

Under the Centralized Training and Centralized Execution (CTCE) paradigm, centralization is maintained across both training and execution [22]. This approach simplifies the complexities of multi-agent interactions by reducing the environment to a single-agent framework; however, this reduction comes at the cost of an exponentially expanded state-action space that can hinder scalability. This approach allows a direct application of classical RL algorithms such as policy gradients, actor-critic methods, or value iteration regardless of additional decentralization constraints [6, 21, 22].

In small cooperative environments with flawless communication and manageable computational burdens, CTCE potentially outperforms decentralized execution methods, which incorporate centralized control. However, as the number of agents increases, CTCE becomes less scalable due to the exponential growth of centralized action and observation spaces alongside the number of agents, leading to an extensive degree of computational overhead and sample complexity. This bound the CTCE’s applicability to theoretical assumptions and small cooperative MARL problems [23].

Despite the theoretical simplicity CTCE offers for coordination, it relies on assumptions that rarely hold under physical constraints. The assumption of constant, synchronized global observability serves more as a computational abstraction than a reflection of real-world environments [24]. For systems like robotic swarms or wireless networks, where processes are inherently decentralized, centralized control is often impractical. Practical deployment reveals significant vulnerabilities, where factors such as communication latency and limited bandwidth render a centralized global observation space nearly impossible to maintain [24].

As a result, the CTCE paradigm fails to scale effectively in dynamic settings. In such

environments, the reliance on global synchronization becomes a liability, necessitating a shift toward decentralized strategies where agents can operate reliably based on localized information. This restricts CTCE to the role of a theoretical benchmark of a best-case scenario, which is used to measure the performance of more flexible architectures. These limitations of CTCE have motivated the introduction of Centralized Training with Decentralized Execution (CTDE), a framework that attempts to address the need for high-level coordination with the practical realities of large-scale, distributed deployment [11].

2.1.2 Decentralized Training with Decentralized Execution (DTDE)

Decentralized Training with Decentralized Execution tries to mimic real-world multi-agent systems by making each agent independently learn and execute its own policy based solely on local observations and individual reward signals. Since there is no centralized system here, access to the global state information or central coordination is not allowed during either training or execution. In such paradigm agents operate under limited communication, strict autonomy constraints, and partial observability. Applications of this system includes traffic systems, distributed sensor networks, and robotic swarms [6, 24].

Although DTDE brings robustness, scalability, and deployability through its distributed approach however it introduce challenges including difficulty in coordinating agents implicitly, non-stationarity from simultaneous policy updates, and the credit assignment problem. Credit assignment problem is a situation in which the system struggle to determine how much each agent’s action contributed to a shared outcome [23, 25]. Due to the independent training and executing process happening in DTDE, an agent treat other agents as part of the environment. This assumption often leads to instability and suboptimal convergence in cooperative tasks [26].

Moreover, integrating global objectives such as fairness and efficiency into the system is difficult since agents lack access to the centralized information and they optimize their own local rewards which may conflict with system level objectives. Although DTDE is attractive from a systems perspective, its limitations have motivated the development of the hybrid paradigm. By retaining the practicality of decentralized execution while mitigating the coordination and fairness challenges inherent in fully decentralized learning, CTDE becomes an acceptable paradigm among the others [11, 27].

2.1.3 Centralized Training with Decentralized Execution (CTDE)

Due to its ability to balance coordination efficiency and practical deployability, CTDE emerged as the dominant paradigm in cooperative MARL. In this learning paradigm, agents are trained through centralized information, such as the global state, joint obser-

vations, or joint actions. During execution each agent operates independently based only on its local observations, without reliance on centralized control or communication [6, 24]. This framework directly addresses the scalability and communication limitations of centralized execution while mitigating the instability and coordination challenges inherent in fully decentralized learning.

CTDE enables agents to reason about the collective impact of their actions during training through their ability of leveraging centralized critics, joint value functions, or shared training signals which helped them to mitigate non-stationarity and credit assignment problems [11]. The common CTDE based algorithms are: Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [27], Counterfactual Multi-Agent Policy Gradients (COMA) [11], Monotonic Value Mixing Network (QMIX) [9], and Value-Decomposition Network (VDN) [8]. These algorithms employ centralized critic or value decomposition techniques to preserve coordination while enforcing decentralized policies at execution time [8, 9, 27].

Due to the centralized training property presented within the system, CTDE provides a flexible foundation for incorporating global objectives, including fairness, efficiency, and robustness. Optimizing social welfare functions within the centralized training of CTDE is one approach for incorporating global objectives like fairness while maintaining autonomous agent behavior at deployment. The centralized training property of CTDE makes it a standard paradigm for large-scale cooperative MARL and the underlying framework for modern extensions. An approach such as Fairness-Aware PPO (FAPPO) and its attention based variant AT-FAPPO [7] are a proof for CTDE’s efficient policy optimization without sacrificing decentralized execution constraints.

2.2 Fairness in MARL

As multi-agent reinforcement learning approaches practical applications, particularly where the consequences of decision making are important, fairness stops being a topic for discussion and becomes something we must contend with on a practical basis. For example, if multi-agent reinforcement learning systems impact multiple end users, higher power output for some users or agents can lead to interference to nearby terminals in a wireless network; as a result, if we want to build trust and encourage adoption with users, we need to factor in fairness as a critical component in the successful deployment and operation of these systems. Earlier work in the field of multi-agent reinforcement learning (e.g., through overviews such as [23]) has focused on aggregate return or overall efficiency. If performance is the only concern, then that makes sense. But the picture becomes more nuanced when we consider who gains and who does not.

According to [13] there are recurring trends that take the form of a Matthew effect. An advantage gives the agent an increase in reward leading to further rewards based

off what they start with. Eventually, they will no longer just be growing their overall amount of reward, but they will also create instability and a lack of diversity within their composition of agents or classes. This skewed reward maximization toward a dominant minority results in systemic inequity that compromises the robustness of the entire multi-agent population. This distributive imbalance not only presents ethical challenges; it also diminishes the motivation for agent participation and threatens the long-term stability and cooperation in the framework.

In order to address this imbalance, recent MARL approaches incorporate fairness into the learning objective. Rather than solely optimizing raw returns, these approaches incorporate social welfare functions or explicit disparity penalties into the learning objective. Methods such as Fairness-Aware PPO and its attention based extension [7] are an example of this approach in which fairness is considered by integrating inequality-sensitive terms into the learning process. Metrics such as the coefficient of variation, Gini coefficient, and minimum-reward criteria are utilized to measure inequality within the system [4, 7].

Fairness itself is inherently multi-dimensional and context-dependent. It encompasses a wide range of definitions and perspectives across the literature, which makes its formalization and implementation particularly challenging [4, 13]. Different authors emphasize different aspects based their views of what systems should achieve which is seen as a mathematical problem. Fairness tends to match some of criteria, such as equal or equitable allocation, representation, or error rates, for a particular or problem. In cooperative settings, fairness may rewards across agents, because group fairness measures no discrepancy on certain attributes. In other contexts, it prevents extreme disadvantage rather than equality, which asserts that any two individuals who are similar with respect to a particular task be treated similarly. The chosen metric shapes outcomes, though principled definitions require an unbiased test to assess whether two individuals are indeed similar.

An example use is the coefficient of variation (CV) which measure dispersion in agents' utilities [13]. A lower variation means that the rewards cluster more tightly around the mean, the outcome is more balanced. A similar viewpoint is shown in [4], which draws inspiration from economics about equity and considers fairness, as it should not be taken as strict equality. In their case, A policy is said to be fair if it does not achieve a highly concentrated distribution of rewards but still achieves a strong return on the system. They consider statistical measures like the Gini coefficient and the CV as a summary of the reward dispersion at the population level. Fairness is evaluated globally and not enforced through fixed per-agent caps. The choice indicates a comprehension of equity as a characteristic of the overall distribution rather than of a hard constraint on each individual.

Fairness is also defined and incorporated into the system as Q-function-based con-

sideration [28]. [28] considers a policy fair if and only if it doesn’t systematically prefer actions that have lower long-term value over actions that have higher long-term value. In this sense, the algorithms prioritize long-term action value over immediate rewards to maintain a merit-based decision-making approach, effectively mitigating the influence of short-term biases. Under this framework, decisions are made in accordance with the reward structure’s inherent logic. By only selecting actions that offer equal or greater long-term utility—as represented by the Q -values [29]—the system ensures that its strategic evolution remains consistent with the true value landscape.

While Q -values represent individual utility, they often overlook the complex relational concerns that are vital to multi-agent stability. As a result, relying strictly on Q -functions can ignore the inter-dependencies that govern agent interactions. As noted by [23], fairness should instead be viewed as a deterministic structural property essential for enduring cooperation. Persistent imbalances do not merely result in unfairness; they actively drive non-stationarity and foster systemic dominance. Without explicit constraints for equity, the learning process may stall in sub-optimal states. Therefore, integrating fairness is critical for enhancing algorithmic resilience, as it stabilizes the learning environment and prevents catastrophic system failure during long-term deployment.

Another core strategy for integrating fairness is the direct optimization of social welfare functions within the algorithm’s objective [17, 30]. Drawing from the lens of distributive justice [31], these works define a fair outcome through a triad of properties. First, efficiency requires the solution to reside on the Pareto frontier. Second, impartiality ensures horizontal equity, meaning the algorithm treats identical agents consistently regardless of their identity. Third, equity addresses the balance of reward distributions to prevent extreme disparities. As suggested by [17], this perspective treats more uniform benefit distributions as inherently superior to those that prioritize a few dominant participants, thus aligning mathematical optimization with social equity.

Building upon these foundations, our approach adopts a welfare-centric perspective that evaluates fairness through both individual reward distributions and aggregate system outcomes. Rather than enforcing strict equality—which can impose rigid constraints on system efficiency—we aim to prevent the disproportionate accumulation of rewards that undermines cooperative equilibrium. Our framework is guided by the principle that no agent should maximize its utility at the cost of marginalizing others. To this end, we incorporate the concept of envy-freeness [32], ensuring each agent perceives its own allocation as at least as desirable as any other. By positioning fairness and efficiency as parallel objectives, we ensure that the pursuit of equity remains consistent with high-performance metrics. This approach facilitates a learning environment where distributional balance and operational effectiveness are co-optimized, ensuring the long-term sustainability of the cooperative system.

2.3 Existing Approaches in Cooperative MARL

2.3.1 Utilitarian Approaches

Initial approaches to cooperative MARL were predominantly governed by utilitarianism, where the central goal was the maximization of the collective return. While this simplified performance metrics to a single aggregate value, it neglected the distributional equity between agents. Despite a succession of architectural and algorithmic advancements addressing challenges such as coordination, credit assignment, scalability, and representational capacity, the core objective of total reward maximization persisted as the standard, leaving the issue of fairness largely unaddressed.

Among the foundational utilitarian approaches, Value-Decomposition Networks (VDN) [8] introduced a linear factorization of the joint action-value function into a summation of individual agent utilities. This architectural choice facilitates decentralized execution within a centralized training paradigm driven by shared team rewards. While this additive decomposition mitigates the exponential expansion of the joint action space—a common bottleneck in traditional centralized Q -learning—it still imposes a restrictive linear assumption. This limitation reduces the model’s representational capacity. Consequently, VDN often fails to capture the intricate, nonlinear inter-dependencies between agents inherent in complex real-world environments. Moreover, the exclusive focus on maximizing collective return means VDN remains indifferent to individual reward distributions. This omission can lead to systemic imbalances where certain agents are marginalized, ultimately compromising the stability and fairness of the cooperative task.

Building on the value-decomposition paradigm, QMIX [9] was proposed to address the representational limitations of VDN [8]. It introduces a nonlinear monotonic mixing network that captures complex cooperative interactions by enforcing a monotonic relationship between individual utilities and the joint action-value function. Unlike the simple additive structure of VDN, the monotonic constraint in QMIX allows for the representation of a significantly richer class of value functions, facilitating the modeling of complex cooperative behaviors while preserving decentralized execution. The authors demonstrated the efficacy of QMIX through extensive evaluations on the StarCraft Multi-Agent Challenge (SMAC) [33], where it consistently outperformed VDN and other baselines. However, this monotonicity constraint inherently restricts the model’s ability to capture non-monotonic or adversarial interactions. Additionally, just like VDN, its objective is still to maximize the aggregate team return regardless of how fairly reward is distributed among team members.

In order to address the credit assignment problem where strategies struggle to assign rewards to individual agents effectively, Counterfactual Multi-Agent Policy Gradients (COMA) [11] presents a method that leverages a centralized critic to compute counter-

factual baselines that address the multi-agent credit assignment problem. Through the use of a counterfactual baseline—achieved by marginalizing out individual agent actions while keeping the actions of the collective fixed, COMA provides a more stable and accurate learning signal for decentralized policies. This approach effectively addresses the multi-agent credit assignment problem, bridging the gap between centralized training and decentralized execution. While this allows agents to learn coordinated strategies without requiring global observability during deployment, a significant limitation persists: the reliance on a centralized critic during training. As the number of agents increases, the computational complexity of training such critics grows substantially, potentially hindering the scalability of COMA in large-scale multi-agent systems.

Recent findings by [10] have challenged the necessity of centralized critics for achieving competitive performance. Through rigorous empirical investigation, [10] demonstrated that Independent PPO (IPPO)—a decentralized learning framework—can outperform established joint learning methods such as QMIX [9] on the complex SMAC [33] benchmark with minimal hyperparameter tuning. IPPO’s robustness is largely attributed to the policy clipping mechanism of PPO [34], which stabilizes the learning signal and mitigates the non-stationarity inherent in independent learning. Notably, IPPO achieves state-of-the-art results without the need for centralized value functions or intricate decomposition techniques, suggesting that the benefits of centralized critics for certain high-dimensional environments are outweighed by the simplicity and robustness of well-tuned independent agents.

Although IPPO [10] demonstrates the power of decentralized learning, it simultaneously highlights the need to identify environments where global information provides a decisive advantage. Integrating fairness or resource-aware constraints remains difficult because local policy optimization does not naturally converge toward global welfare goals. In decentralized settings, these trade-offs between individual rewards and collective equity present a significant barrier to the development of socially responsible and practically sustainable MARL systems. These issues highlight that, despite the promising results of independent learning methods, integrating global objectives remains an unresolved and complex problem in cooperative multi-agent reinforcement learning. All of these subsequent approaches share a fundamental assumption that maximizing total team reward is sufficient to achieve desirable outcomes which does not guarantee equitable reward distribution among agents.

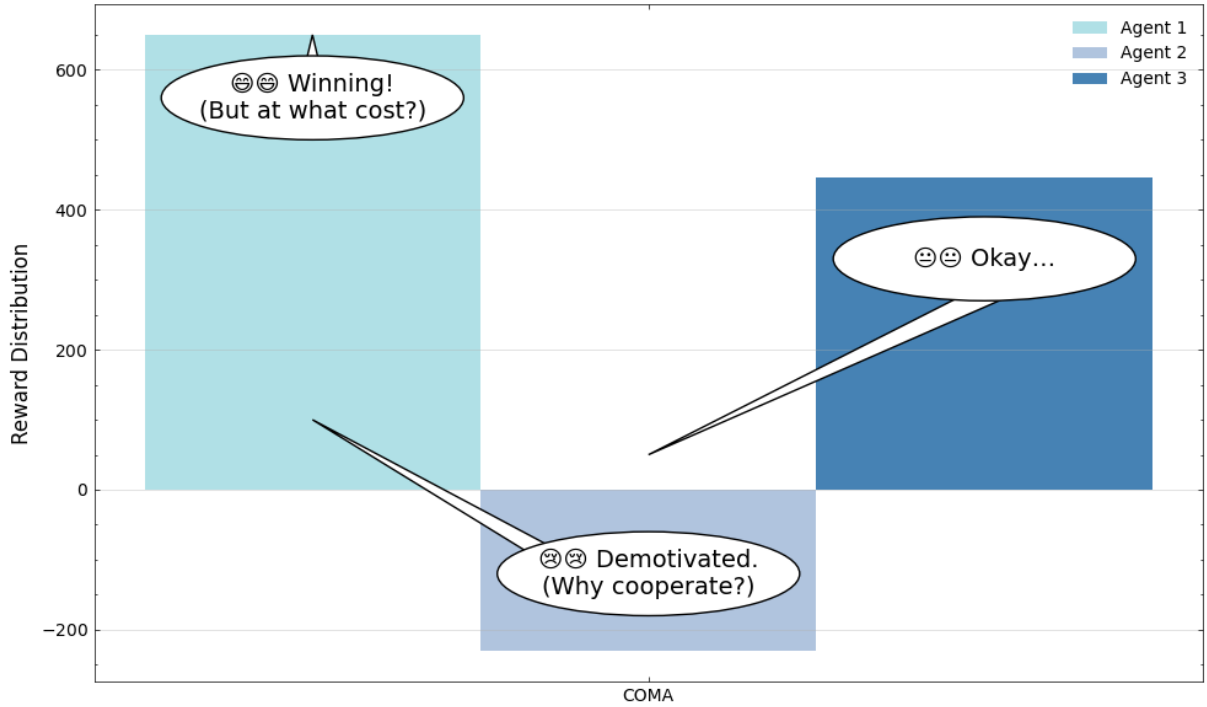


Figure 2.2: Reward accumulation vs agent starvation and de-motivation.

Figure 2.2 is the experimental result of COMA [11] under a random MDP environment [7]. The figure provides clear empirical evidence of unfair reward distribution among agents in the cooperative multi-agent system, reflecting a Matthew effect where minor stochastic advantages obtained during the early phases of training are gradually amplified through the learning dynamics. This effect leads to persistent structural dominance by a subset of agents and increased reward inequity among agents over time. The agent positioned on the left tends to dominate the flow of environmental rewards, which results in starvation for the agent positioned in the middle. Of course, the system’s aggregate performance appears as stable or increasing; however, the accumulated rewards of the agent positioned on the left indicate the high Gini coefficient of the system and the fundamental imbalance in resource allocation. When agents are marginalized, the resulting lack of meaningful feedback renders their policy gradients to become noisy and less reliable. This degradation of the gradient quality leads to a state of permanent sub-optimality, undermining long-term cooperation within the system.

2.3.2 Fairness Integrated Approaches

Conventional MARL frameworks prioritize the maximization of aggregate discounted returns, often at the expense of distributional equity [29, 35]. However, an exclusive focus on return maximization that neglects fairness considerations introduces significant systemic risks. Such a narrow optimization lens often masks underlying distributive flaws, leading to fragile cooperative structures that are prone to instability in complex environments.

The systemic risks include the production of discriminatory outcomes that are not only unethical but actively harmful because of the unconstrained optimization. In sensitive applications, these algorithms risk automating and disseminating societal biases [29, 36]. Moreover, fairness is increasingly becoming a legal and regulatory requirement in many jurisdictions; explicit mandates now exist to ensure algorithmic equity and accountability [29, 37].

This shift toward fairness as a critical design constraint has motivated a growing body of research to focus on designing algorithms that balance collective performance with equitable outcomes for individual agents. Below we tried to discuss existing fairness-integrated approaches by broadly categorizing them by how their method utilized fairness.

Fairness through Theoretical Assumption and Hierarchical Control

The foundational understandings of fairness constraints in RL, primarily in a single-agent environment, are established by [28]. They emphasize fairness in long-term reward and analyzing computational challenges. By defining fairness based on the true expected long-term rewards, [28] designs the Fair-E3 algorithm that achieves near-optimal performance with polynomial sample complexity under the defined fairness constraint. Crucially, their work identifies a complexity ceiling, proving that exact fairness requirements can lead to exponential computational demands. By adopting a fairness notion that permits conditional discrimination favoring actions with demonstrably higher returns, they align RL objectives with the pragmatic constraints of high-stakes domains like finance and human resources.

While [28] provides a foundational pillar for fair RL, their framework is constrained by several critical operational overheads. Central to these is the dependency on the true optimal Q-function (Q^*), an idealized metric that is notoriously difficult to approximate in stochastic or high-dimensional settings. Furthermore, their polynomial sample complexity bounds, while theoretically significant, often translate to prohibitive data requirements in practice, particularly within expansive state-action spaces. Perhaps most notably, their focus is exclusively single-agent; the framework does not account for the competitive inter-agent dynamics of resource allocation. This results in a fairness gap: while their approach ensures individual reward ordering, it fails to provide a mechanism for managing the collective reward disparities that define systemic inequality in MARL.

Expanding into hierarchical architectural design, [13] proposes the Fair-Efficient Network (FEN), which utilizes a multi-level reinforcement learning paradigm to reconcile the often competing objectives of global efficiency and local fairness. By targeting the stochastic resource allocation challenges typical of wireless networks, FEN addresses the structural bias of traditional MARL approaches optimized solely for aggregate utility, which tends to catalyze winner-take-all dynamics. The FEN model addresses this by

decomposing fairness for each agent and employing a hierarchical structure that includes a controller and multiple sub-policies, allowing agents to learn policies that consider both their own rewards and the collective fairness of the system.

The controller in the hierarchical reinforcement learning model FEN [13] is responsible for selecting among the various sub-policies based on the current observation. Specifically, the controller outputs a probability distribution $p_\theta(z \mid o_t)$ over sub-policy indices z , given the current observation o_t . At each decision point, the controller samples a sub-policy index z_t from this distribution, which determines the sub-policy to be activated for the next T time steps. In order to measure fairness among agents, [13] utilized coefficients of variation (CV).

Despite their achievement, their approach assumes that agents must share individual utility values with neighboring agents to achieve fairness, which can be impractical in environments where such information is private or sensitive. It also relies on a hierarchical structure in which individual policies are first learned and then adjusted for fairness through a leader or team policy, limiting its scalability and suitability for fully decentralized systems. Moreover, the assumption of additive rewards and equal resource access reduces the method’s realism and applicability in dynamic, real-world settings where agents often operate under unequal opportunities and constraints.

Constraint-based Fairness Integration

The effect of unconstrained maximization of team reward motivated [18] to introduce team fairness as a group-based fairness measure. [18] propose a learning-based techniques such as Fairness through Equivariance (Fair-E) and its soft-constraint variant Fair-ER, which can be directly integrated into decentralized deep RL settings. These methods allow agents to learn fair coordination strategies through shared reward signals without the need for complex equilibrium computation or centralized critics. [18] presents task-agnostic methods that promote fairness in a broader sense, independent of any specific fairness model. Rather than that, these techniques integrate fairness through constraints on agent policies—either hard constraints that policies must adhere to strictly, or soft constraints through regularization that introduces penalties related to fairness in the learning objective. Instead of designing new algorithms for each specific fairness principle, this framework provides enhanced flexibility because it can be adjusted to different concepts of fairness through a mere modification of constraints.

Despite its contributions, the framework proposed by [18] exhibits critical limitations regarding temporal adaptability and structural homogeneity. Their approach is predicated on a baseline of functional symmetry, which precludes its effectiveness in heterogeneous environments where agents possess divergent capabilities or reward scales. Furthermore, the reliance on static training-time constraints creates a blind spot for

emergent, real-time biases. Without an integrated mechanism for online self-monitoring or individual fairness correction, the system remains unable to modulate policy updates in response to the dynamic behavioral shifts that characterize complex, competitive MARL trajectories.

Building upon the foundational constraints of [18], [38] introduces the Fairness-Sensitive Policy Gradient Reinforcement Learning (FSPGRL) framework. This method shifts the focus from abstract regularization to sensory-grounded bias detection, utilizing Principal Component Analysis (PCA) and logistic regression to model human-informed bias dimensions—specifically willingness, quality, priority, and risk. While this approach enables robots to identify subtle discriminatory patterns in real-time interaction, its corrective mechanism remains fundamentally reactive. By adjusting rewards post hoc to penalize detected misalignment, FSPGRL addresses bias after its emergence, contrasting with the proactive, state-integrated adjustment mechanisms required for complex, autonomous MARL.

A distinctive contribution of [38] is the introduction of hierarchical penalization, a mechanism that allows for the granular differentiation of fairness violations based on their ethical severity. By categorizing behaviors—such as distinguishing between immediate risk and long-term neglect—the framework enables a nuanced prioritization of ethical constraints. This is achieved by integrating these penalties directly into the policy gradient, ensuring that the learning objective is explicitly modulated by the specific nature of the biased behavior rather than a uniform penalty. This promotes more equitable and unbiased assistance for individual users. This shifts the focus from group-based fairness to individual fairness and bias mitigation, filling a key gap in applying fairness principles to human-AI interactions.

Fairness through Actor-Critic and Multi-Objective Optimization

In order to balance the incorporation of multiple conflicting objectives such as efficiency, safety, and fairness, [39] proposes a joint optimization framework that handles multiple reward functions in reinforcement learning by offering a mathematically grounded formulation. Their approach formulates the learning process as a constrained optimization problem where each reward corresponds to a constraint or objective to be satisfied simultaneously. This enables their method to be proactive rather than reactive in handling fairness. Instead of waiting for bias to emerge and correct it, in [39] agents learn policies that inherently balance fairness-sensitive rewards with other performance goals during training.

However, although [39] balance conflicting trade-offs of reward signals using Lagrangian relaxation and adaptive learning rates, their method suffers from extremely slow convergence due to its reliance on the Monte Carlo method for gradient evaluation.

Plus, their approach still remains quite general and abstract, without fully addressing the challenges of stable and efficient learning in continuous control settings, particularly in actor-critic architectures. Due to this focus on a generalized formulation, [39] overlooks the structural complexities of continuous action spaces and the non-stationary dynamics of shared environments. Moreover, their framework stops short of providing convergence guarantees for actor-critic architectures—a critical omission for ensuring the stability and reliability of fairness-integrated policies in practical deployment.

In contrast to heuristic-based models, [40] introduces the Fairness-aware Actor-Critic (FA2C) algorithm, which formally embeds α -fairness objectives within the actor-critic framework. By grounding fairness in a rigorous utility function and providing provable convergence guarantees—a significant advancement over the more abstract frameworks like [39]—FA2C offers a theoretically sound solution for network resource allocation. Unlike the single-agent focus of FA2C, multi-agent fairness requires navigating a landscape of interactive complexity, where agents possess independent policies and private reward signals. Consequently, the centralized assumptions of FA2C fail to address the emergent social challenges of MARL, such as individual dominance and the coordination failures that arise in the absence of a global coordinator.

To overcome the single-agent constraints of previous actor-critic models, [14] proposes a framework specifically engineered for fairness integration in Multi-Agent Markov Decision Processes (MAMDPs). Moving beyond the pursuit of individual utility, their methodology mandates that agents optimize a joint fairness objective. By utilizing α -fairness—a versatile utility function that encapsulates both max-min fairness and social welfare maximization—the framework effectively addresses the challenges of agent heterogeneity and partial observability. This approach ensures that decentralized policies remain aligned with collective equity targets, even within information-constrained environments.

By leveraging multi-agent policy gradient methods combined with fairness regularization, [14] establishes a robust framework for cooperative learning under strict convergence guarantees. This methodology significantly extends the single-agent utility paradigm of [40] into the realm of multi-agent coordination, effectively modeling the complexities of resource sharing and cooperative robotics. Yet, while their convergence proofs provide high-level stability, the framework’s reliance on abstract utility functions often results in a contextual disconnect. It fails to account for the physical and morphological shifts—such as agent growth or spatial bottlenecks—that transform an equitable policy into an unfair one during real-time deployment.

Fairness Integration through Protected Attributes

Expanding the scope of fairness to include demographic and categorical variables, [15] introduces a framework that explicitly integrates protected attributes into the multi-agent learning objective. By formulating attribute-aware fairness criteria, they ensure that policy optimization remains equitable across agent groups defined by sensitive properties. Furthermore, [15] proposes methods for representing protected attributes within the agent’s state or observation space, enabling RL algorithms to condition policies on fairness-sensitive information and apply fairness constraints or penalties in a way that explicitly guards against discrimination. The system’s consideration of fairness-sensitive protected attributes to reason about fairness at the intersection of utility and identity is essential for trustworthy AI deployments in sensitive domains such as healthcare, finance, or public services. However, incorporating protected attributes introduces challenges related to privacy, ethical data use, and possible policy complexity, which [15] discusses as important areas for future research.

Moreover, [15] focused primarily on conceptual modeling and constrained optimization formulations and did not fully explore the algorithmic mechanisms for balancing fairness with learning stability and sample efficiency in modern RL. It leaves open the question of how to effectively integrate these fairness constraints into scalable, high-performance RL algorithms that can handle complex, non-stationary environments and continuous action spaces. By proposing a method that incorporates fairness constraints and protected attribute considerations within the widely-used and scalable PPO [34] framework, [16] addresses their previous gap presented in [15]. Unlike their earlier work [15], which focuses on problem formulation and attribute-aware fairness definitions, [16] implements fairness-aware learning by augmenting the PPO objective with fairness regularizers or constraints, enabling policies to learn fair behaviors. By considering protected attributes and integrating equity considerations through penalties within PPO’s clipped surrogate objective [34], [16] effectively introduce fairness into their implementation.

However, their approach still assumes full access to protected attributes such as race and gender during both training and evaluation. This assumption is often unrealistic due to privacy constraints or incomplete data presence in real-world systems. Plus, the fairness constraints that they applied are static and group-based, such as demographic parity, regardless of individual-level fairness considerations. Due to the static fairness constraints that [16] utilized, their methods’ effectiveness heavily depends on carefully tuning of penalty parameters α and β , which can be environment-specific and limit the approach’s scalability and generalization.

Furthermore, their algorithm is developed for a single-agent, centralized setting and lacks the scalability or coordination mechanisms required for multi-agent environments, where fairness must emerge from decentralized interactions. Their experimentation pro-

vides a limited empirical evaluation, as it is primarily evaluated in small, synthetic environments that do not reflect the complexity of fairness-sensitive real-world tasks. And also, they didn't mention a principled mechanism to control the trade-off between fairness and utility, leaving practitioners to rely on heuristic tuning of regularization parameters.

Fairness through Social Welfare Function Optimization

Fairness integration through Social Welfare Function optimization has recently emerged as a principled approach for embedding equity considerations directly into multi-agent reinforcement learning objectives. [17] establishes a fairness framework grounded in social welfare functions, which can incorporate multiple fairness definitions beyond additive rewards. They formulate a cooperative MARL as a fair optimization problem, where the goal is to maximize a fair social welfare function (SWF) that aggregates the utilities of all agents. Mathematically, this objective is expressed in [17] as:

$$J(\theta) = \phi(J_1^I(\theta_1), \dots, J_N^I(\theta_N)) \quad (2.1)$$

where, θ represents the policy parameters for each agent, and $J_i^I(\theta_i)$ denotes the individual utility of agent i . To address the conflicting interests among agents, the authors proposed the Self-Oriented Team-Oriented (SOTO) network architecture, consisting of two sub-networks:

- the self-oriented policy, which focuses on maximizing individual utility $J_i^I(\theta_i) = \mathbb{E}_{\tau \sim \pi_i}[R_i(\tau)]$.
- the team-oriented policy, which optimizes the overall social welfare function $J(\theta)$.

The learning process in [17] employs policy gradient methods, enabling agents to update their policies based on the gradients of the social welfare function, which is expressed as

$$\nabla_{\theta_i} J(\theta) = \nabla_{\theta_i} \phi(J_1^I(\theta_1), \dots, J_N^I(\theta_N)) \quad (2.2)$$

Even though [17] define fairness through social welfare optimization and propose principled fairness using equilibrium concepts such as lexicographic or Gini-weighted objectives, their use of pre-defined social welfare functions to represent specific fairness principles is task-dependent, implying that the fairness objectives are customized directly within the algorithm based on the chosen welfare function.

Another line of research addresses the challenge of learning fair policies in cooperative multi-agent reinforcement learning (MARL) by explicitly incorporating welfare optimization into the policy update process [7]. [7] identify a key limitation of standard MARL algorithms which typically focus solely on maximizing expected cumulative rewards in

that it may produce inequitable outcomes among agents even in fully cooperative tasks. To overcome this, [7] formulates the fairness objective via generalized Gini welfare function grounded in principles of distributive justice, emphasizing desirable properties such as Pareto dominance, equity, and impartiality, which are encoded through a generalized welfare aggregation of agent utilities.

By introducing the Fairness-Aware Multi-Agent PPO (FAPPO) algorithm, which adopts the IPPO [10] framework for optimizing each agent’s local observation individually, [7] ensures that the policy update considers fair reward distribution with the aim of treating all agents equally. Particularly, FAPPO maintains fair independent policy learning by optimizing the generalized Gini welfare function in the policy objective of PPO [34], thus preserving stability and flexibility in decentralized environments. Furthermore, they introduced a communication-enhanced variant, AT-FAPPO, that incorporates a self-attention mechanism to allow agents to share information relevant to decision making during training [7]. This attention module facilitates coordination by enabling agents to weigh incoming messages from peers according to their relevance for achieving equitable outcomes.

However, even though [7] optimized the generalized Gini welfare function as formulated in Equation (1.1) for prioritizing the worse-off agents, they explicitly utilized static, manually tuned weight vectors ω_i (1.2) as a welfare weight for the welfare function optimization in their implementation. While this static, geometrically decreasing sequence assigns larger weights for agents with small utility values, the static nature of the welfare weights imposes severe limitations in [7]’s method. The reliance of [7]’s methods on a predetermined, fixed welfare weight restricted the algorithm’s marginal sensitivity to inequality to be constant across all environmental states, agent rewards, and time steps. Thus, the system is unable to dynamically adjust the welfare weights based on the actual severity of the resource disparity or the specific demands of a complex state, limiting the robustness of the system in diverse and complex real-world MARL environments.

Although the above works share conceptual similarities with our study in addressing fairness, our method is closely related with FAPPO framework [7]. Instead of applying fixed welfare weights and static penalty parameters like them, our study present principled mechanism from a new perspective inspired by the law of diminishing return and propose a weight adaptive AW-FIPPO and AT-(AW-FIPPO) that continuously adapts to environment state and reward of agents (based on the current reward disparities), enabling more effective and nuanced fairness enforcement in complex, real-world MARL systems.

Our method is designed to integrates a Fairness Critic Network that generates adaptive weights based on which the policy penalize individual agents (who begin to accumulate rewards) through logarithmic transformation to the PPO surrogate to enforce concave, fairness-oriented gradient updates. Therefore agents can adjust their behavior

towards those holding minimal to none reward. In a balanced state, the system maintains utility mode, utilizing the linear surrogate to maximize cumulative team rewards.

Our framework employs state-dependent weighting to establish a dynamic equilibrium between efficiency and equity, ensuring policy robustness across diverse operational regimes. To validate this approach, we utilized two specialized benchmarks: a Random MDP for assessing stability under transition noise, and a Matthew Effect environment designed to catalyze rich-get-richer dynamics. Inspired by the economic Law of Diminishing Returns, our weighting mechanism prevents reward monopolization by ensuring that the marginal utility of additional gains decreases for dominant agents. This encourages a distributive parity that mimics natural ecological balances, maintaining systemic welfare without compromising aggregate utility.

2.4 Challenges in Addressing Fairness in MARL

Addressing fairness in MARL is fundamentally challenging because integrating fairness constraints into the standard MARL frameworks by itself introduces complex, multi-dimensional optimization challenges. This challenge basically comes from the inherent trade-off between system efficiency and equity because cooperative MARL algorithms are typically designed to maximize the total expected return of the system regardless of how reward is being distributed among agents [4, 7]. As agents are designed to prioritize selfish goals to maximize their own rewards to achieve the team goal, underperforming agents within the team may be neglected due to the reward monopolization by a few, creating conflict with the overall team’s need for fair task allocation. Therefore, achieving high system-level efficiency often comes at the expense of fair reward distribution among agents within the team, where certain agents consistently dominate while others are marginalized, leading to systemic inequity and potentially reducing cooperation and long-term stability of the system [13, 23].

The trade-off is especially pronounced in heterogeneous environments where agents differ in capabilities or initial conditions. On the other hand, enforcing strict fairness also prevents some agents from achieving more reward because they get constrained by the system to consider the disadvantaged agents [41]. Prioritizing the improvement of low-performing agents requires redistributing resources or actions away from higher-performing agents, which can reduce total cumulative reward in tasks such as resource allocation. Therefore, considering exclusive focus on either of these objectives creates an inherent tradeoff between them, remaining a common challenge in designing and implementing fairness-aware MARL algorithms that balance fairness and efficiency within the system. The trade-off can further be affected by technical limitations existed in the algorithms such as static objective functions that struggle to navigate the non-stationary dynamics of multi-agent environments.

Another critical challenge in fairness-aware MARL is the scalability of the system. As the number of agents in a system increases, maintaining fairness in resource allocation, task assignment, and reward distribution becomes exponentially more complex [21, 24]. Many fairness-aware algorithms, such as those using social welfare functions, Gini-based metrics, or variance-based reward adjustments, require either centralized computation or global knowledge of all agents’ rewards. In small-scale systems, this centralized computation is tractable, but in large-scale multi-agent systems, the dimensionality of the joint state-action space and the combinatorial growth of possible reward distributions make fairness evaluation and enforcement computationally expensive.

Scalability constraints in fairness-aware MARL manifest as a tripartite challenge. Primarily, the calculation of global distributive metrics—such as the Gini index or generalized welfare functions—imposes a centralized information bottleneck, requiring real-time access to high-dimensional joint reward states. Secondly, the coordination overhead scales super-linearly with the agent population, particularly when fairness enforcement necessitates inter-agent communication. Finally, even when decentralized, the competing pressures of individual reward maximization and collective fairness can lead to learning instability, effectively capping the maximum viable population size for equitable MARL frameworks [23].

Furthermore, as [42] stated, as the size of the state-action space increases the computational demands required for exploration and exploitation increase as well. In summary, although the above discussed challenges are inherent to MARL, this research specifically addresses the efficiency and equity trade-off. To balance these competing objectives, we propose an algorithmic framework that integrates fairness into the learning process based on adaptive weights generated from an auxiliary network. This network generates state-dependent weight vectors by dynamically analyzing the environment’s state and the evolving reward distribution among agents.

2.5 Related work

A wide range of cooperative MARL methods that aim to achieve multiple objectives is introduced. These methods differ in the kind of objective that they aim to address and how they address those objectives. Most of these methods primarily focus on maximizing collective returns or efficiency as a main objective. Some others also tried to incorporate explicit fairness and address reward imbalance among agents or take fairness as a main objective as well. In order to clearly show our desired properties and the gap we aim to fill, we categorize related work into value-decomposition methods, actor-critic-based policy-gradient approaches, and fairness-incorporated extensions, highlighting their respective strengths and weaknesses in balancing efficiency and equity.

Value-decomposition methods are the first category of related work we put forth for discussion, and it include Value-Decomposition Networks (VDN) [8] and Monotonic Value Function Factorization (QMIX) [9]. These frameworks stand as the foundational frameworks for value-function decomposition, enabling the derivation of individual agent utilities from a centralized joint Q-value function throughout the learning process. Both VDN and QMIX facilitate coordination among agents while allowing for independent action selection during execution by learning a factored representation of the global value.

While both of these approaches [8, 9] offer convergence stability, they are fundamentally limited by their reliance on discrete action space and off-policy learning strategies. Additionally, these frameworks are purely utilitarian, and they lack mechanisms for fairness consideration in reward distribution. Both of these approaches prioritize the aggregate sum of rewards regardless of how the reward is being distributed. This strategy is highly affected by the Matthew effect in biased environments, where the system may achieve high total utility by essentially sacrificing the learning trajectories of sub-optimal agents.

The second category of our related work is methods that adopt an actor-critic architecture to directly optimize agent policies, which includes methods like Independent Proximal Policy Optimization (IPPO) [10] and Counterfactual Multi-Agent (COMA) [11]. These frameworks move beyond the value-based decomposition by employing actor-critic architectures, which allow for a more natural handling of continuous action spaces and provide stronger theoretical guarantees for policy convergence in stochastic environments.

Although IPPO facilitates decentralized learning and COMA enhances credit assignment via counterfactual baselines, both frameworks' optimization objectives remain strictly efficiency-centric, treating reward distribution as an incidental byproduct rather than a core systemic constraint. By prioritizing aggregate cumulative returns, these methods lack the internal mechanisms required to regulate reward distribution. Without an integrated fairness-aware objective or a mechanism to penalize utility variance, these frameworks are susceptible to the Matthew Effect, where structural advantages lead to

winner-take-all dynamics that marginalize underperforming agents.

Fairness-incorporated MARL methods are the last category of our related work including methods like SOTO [17], FEN [13], and FAPPO [7]. By incorporating fairness objectives and inter-agent communication mechanisms, methods in this category move beyond raw utility maximization and mitigate reward disparity among agents.

While FEN [13] introduces a decentralized fairness objective based on coefficient of variation fairness criteria, FAPPO [7], on the other hand, optimizes a generalized Gini welfare function to utilize fair policy updates. Even though these frameworks address fairness in reward distribution among agents through the concept of distributive justice, both of the frameworks are fundamentally constrained by their reliance on static, manually tuned fairness parameters. For instance, FAPPO [7] used static Gini weights ($\omega_i = 1/2^i$), which imposes a rigid constraint that does not account for the dynamically changing environment or the evolving agent reward, struggling to maintain robustness when faced with sudden reward distribution or complex environmental shifts. This creates a fairness-efficiency tradeoff, where the system remains constrained by fixed parameters that may be too restrictive in balanced states or insufficient in highly biased scenarios.

While FEN [13] and FAPPO [7] advance distributive justice through Coefficient of Variation and Generalized Gini Welfare metrics, both are architecturally limited by their reliance on static regularization. FAPPO, for instance, employs fixed Gini weights ($\omega_i = 1/2^i$) that impose a monotonic penalty structure regardless of real-time environment dynamics. This lack of parametric plasticity or adaptability prevents the system from distinguishing between temporary performance fluctuations and entrenched systemic bias, ultimately inducing a fragile fairness-efficiency trade-off where fixed constraints are either unnecessarily restrictive or insufficiently robust. To the best of our knowledge, there is currently no mechanism capable of dynamically modulating the fairness weight in response to real-time environmental context and evolving reward disparities among agents. This limitation motivates the need for an adaptive mechanism that can continuously adjust the fairness weight based on contextual signals such as the degree of reward inequality and current environmental state.

Table 2.1 provides a complete summary of our related work, identifying approaches by their key characteristics. We compare the approaches using properties like actor-critic structure, policy-gradient usage, credit assignment mechanisms, ability to support continuous actions, convergence stability, on-policy learning, communication capabilities, and fairness awareness. In the table, a check mark (✓) indicates that a method performs well in the corresponding aspect, whereas a cross mark (✗) denotes the limitation of that capability in the corresponding aspect.

Table 2.1: Comprehensive summary of multi-agent algorithms

Characteristic	VDN [8]	QMIX [9]	IPPO [10]	COMA [11]	SOTO [17]	FEN [13]	FAPPO [7]
Policy gradient	✗	✗	✓	✓	✓	✓	✓
Actor-critic	✗	✗	✓	✓	✓	✓	✓
Continuous actions	✗	✗	✓	✓	✓	✓	✓
On-policy	✗	✗	✓	✓	✓	✓	✓
Convergence stability	✓	✓	✓	✗	✗	✗	✓
Communication mech.	✗	✗	✗	✓	✓	✓	✓
Fairness-aware	✗	✗	✗	✗	✓	✓	✓
Adaptive gini weights	✗	✗	✗	✗	✗	✗	✗
Adaptivity to reward changes	✗	✗	✗	✗	✗	✗	✗
Adaptivity to env.t changes	✗	✗	✗	✗	✗	✗	✗

Chapter 3

Methodology

3.1 Research Methodology Framework

In order to develop our novel weight-adaptive approach, we adopt the Design Science Research Process (DSRP) methodology, which is a paradigm specifically suited for the development and validation of innovative artifacts aimed at solving complex socio-technical problems [43]. The DSRP methodology is a systematic and iterative framework used to design, implement, and evaluate innovative solutions [43]. DSRP contains iterative phases or cycles such as problem identification, objective definition, design and development, demonstration, evaluation, and communication. Therefore, we utilize these progressive stages for the creation of our weight-adaptive solution that balances the tradeoff between efficiency and fairness. We balance this tradeoff by dynamically navigating between maximizing the system efficiency and minimizing reward monopolization by a few agents based on the actual disparity level reflected within the system.

As illustrated in Figure 3.1, our study proceeds through the structured and progressive sequence of DSRP phases that begin from the abstract identification of the fairness-efficiency tradeoff to the production and dissemination of weight adaptive MARL framework. This methodology ensures that the produced artifact is both theoretically sound and technically robust by providing a replicable framework that strikes a balance between the general need for collective utility maximization and the equitable reward distribution among individual agents in cooperative MARL.

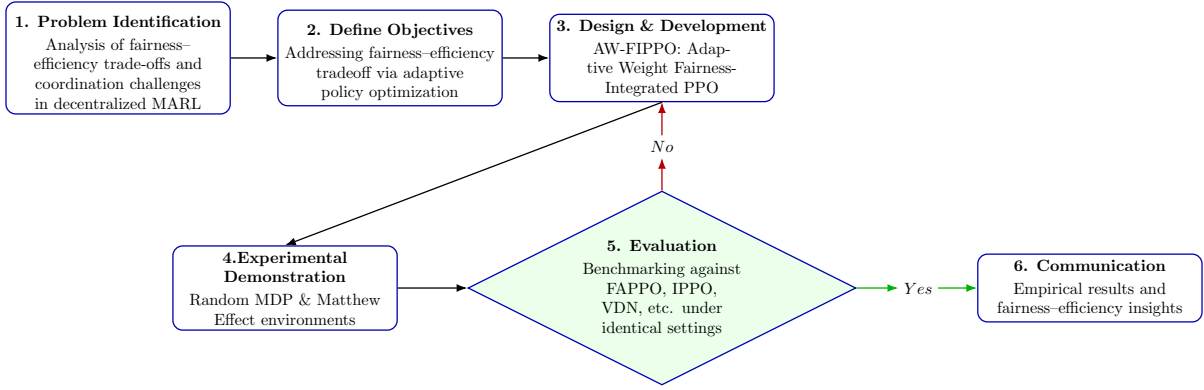


Figure 3.1: Design science research approach

Problem Identification and Motivation

As an initial phase, through a comprehensive literature review we analyze that traditional cooperative MARL frameworks are driven by the primary objective to maximize the total joint returns. However, this singular objective often leads to systemic inequity that happens when a few agents monopolize the system’s reward at the expense of the fairness of the system. This systemic inequity is reflected by a common phenomenon known as the Matthew Effect, where agents that gain an early advantage continue to monopolize rewards while others are starved.

Recent advancements like FAPPO [7] attempted to mitigate this systemic inequity by optimizing the Generalized Gini Welfare (GGW) function; they heavily relied on static, manually tuned fairness weights. Specifically, they optimized GGW using agent-specific weights $w_i = \frac{1}{2^i}$ for $i = 0, \dots, n - 1$. Even though these static and geometrically decreasing weight vectors correctly assign higher influence to worse-off or low-performing agents and introduce some level of fairness into the system as discussed in Section 1.3 and the formulation in (1.3), the weight’s static nature presents a significant bottleneck. Because these weights remain constant throughout the entire training process, the system lacks adaptivity to the evolving nature of the environment state and the changing reward imbalance that occurs as the agents learn. And then, having identified this lack of adaptability as a core research gap, we move to the next phase to define the requirements for a more dynamic solution.

Definition of Objectives for a Solution

In the second phase, guided by the research questions formulated in the Section 1.4, we translate the identified challenges into concrete design objectives. Specifically, these research questions seek to understand how collective efficiency can be preserved while preventing unfair reward accumulation in decentralized multi-agent systems. Accordingly, the primary objective as stated in Section 1.5, is to engineer a learning framework ca-

pable of autonomously navigating the fairness–efficiency trade-off through dynamic and adaptive mechanisms in order to resolves the rigidity of static welfare constraints in cooperative MARL. Unlike fixed or manually tuned constraints employed in approaches such as FAPPO, the proposed solution aims to replace static fairness parameters with an adaptive process that responds to the evolving environmental state and the changing performance disparities among agents.

Specifically, the objective is to design a system that can detect disparity thresholds—points at which inequity adversely affects the group —and shift its optimization strategy between utilitarian (maximizing total return) and egalitarian (balancing distribution) goals. This requires the artifact to be computationally efficient and compatible with on-policy training, ensuring that fairness integration does not compromise the convergence stability or the joint performance of the multi-agent system.

Design and Development

Based up on the objective in Section 1.5 that we aim to achieve, we introduce AW-FIPPO method that incorporated a dedicated additional Fairness Critic Network. This network is designed to dynamically learn and adjust welfare function weights by observing the environment state and the current reward distribution among agents. Unlike previous static method such as FAPPO that relied on fixed, pre-determined weights to impose fairness, our approach uses the added network to produce a context aware weight vectors as its primary output. This weight vectors acts as a dynamic foundation to balance fairness efficiency tradeoff, blending the standard PPO surrogate loss with a concave logarithmic utility transformation.

The standard PPO surrogate loss used for total reward maximization in a balanced state while the concave logarithmic transformation used to integrate fairness in the state where unfair distribution is detected. The log concavity integrate fairness by lowering the importance of additional gains for agents who have already accumulated significant rewards, while increasing the importance of rewards for those who have minimal to none. Instead of optimizing welfare function with a rigid fairness constraints, our network learns to detect the distribution gap between agents and shifts the learning accordingly. Therefore through iterative design and empirical evaluation, we developed a method that remains both efficient and socially equitable.

Demonstration

In the fourth phase, we conducted the experiment on two distinct fairness sensitive environments: the Random MDP and the Matthew Effect scenario. These test-beds are intentionally used because they showcase how cooperative MARL tasks can inherently produce unfair outcomes, where certain agents become more advantageous than others.

Random MDP environment showcase a scenario where rewards are distributed randomly. In contrast, the Matthew Effect scenario simulates a more complex phenomenon where the rich get richer and the poor get poorer. In this environment, agents that starts with a slight advantage or happen to receive more rewards during early exploration tend to accumulate resources at an accelerating rate, while others are pushed towards reward starvation.

Because our framework utilizes on-policy training, there is no separate data-gathering stage, the agents generate training samples through live interaction within these environments. As agents interact within the Random MDP and Matthew Effect scenarios, the fairness critic dynamically detects the agents who begins to monopolize resources. This cause the policy to adjust unfairness through log transformation for prioritizing the learning progress of the under-performing agents. This phase serves as formal validation of the proposed dynamic fairness mechanism in achieving a more balanced and efficient outcome.

Evaluation

In this phase, we evaluate the performance of our AW-FIPPO framework against standard MARL baselines, including FAPPO, IPPO, VDN, QMIX, and COMA. Under identical experimental conditions, we analyze their performance in terms of efficiency and equity. We measure efficiency using total cumulative returns and equity using Gini coefficient, Coefficient of Variation (CV), Minimum reward, and Social Welfare Score.

This phase is critical for validating whether our approach successfully navigates the fairness-efficiency trade-off more effectively than the baselines or not. We analyze the improvements in total return, Gini coefficient and CV values compared to the baselines. Furthermore, we analyze the stability of the on-policy learning curves to ensure that the dynamic weighting does not prevent the agents from reaching a stable, high-performing policy. We used the result gathered from these experiments to draw final conclusion about the artifact and its impact on cooperative MARL systems.

Communication

In the communication phase, our research findings are formally documented and submitted for publication to peer-reviewed research publication, articulating AW-FIPPO’s contribution to the broader MARL community. We analyze the shift from static fairness weight configuration to a learned and dynamic one, demonstrating a learned fairness integration while preserving high performance. We provide a detailed analysis of how the system dynamically transition between utilitarian and egalitarian objectives. This communication phase completes the design science cycle by providing a replicable, theoretically grounded and empirically validated solution for creating more equitable multi-agent

systems.

3.2 Proposed Approach Description

IPPO framework [10] is employed as a methodological foundation of this research in which each agent learns a decentralized policy. This framework used as the base algorithm for training both the baseline FAPPO [7] and this study. FAPPO modifies the standard optimization objective of the framework by focusing on maximizing the generalized gini welfare function ϕ_ω , rather than simply maximizing the discounted sum of rewards, thereby promoting equitable outcomes. Specifically, FAPPO employs the core mechanics of IPPO for learning individual policies based solely on each agent’s local observations z_t^a [7].

In this independent learning setting, FAPPO update the policy by computing the gradient of the welfare function using a variant of the policy gradient theorem: $\nabla_\theta \phi_\omega(J(\pi_\theta)) = w_\sigma^\top \nabla_\theta J(\pi_\theta)$ [7], where w_σ is the fixed welfare weight vector sorted by agent utility $J(\pi_\theta)$. Their method further refines the IPPO training by using the team reward $r_t(s_t, u_t)$ as an approximation for the per-time-step reward and estimates the advantage A^a for each agent separately. They also incorporates individual advantage normalization and separate policy and critic loss clipping for each agent, a PPO-specific technique used to stabilize learning and prevent the advantage scale from being dominated by other agents.

Building up on these foundational components—RL stability from PPO [34], independent learning setting from IPPO [10], and fairness-oriented optimization from FAPPO (and it’s variant ATFAPPO for coordination efficiency from attention-based communication) [7], this study propose AW-FIPPO algorithm that introduce a novel auxiliary network, the Fairness Critic ($\mathbf{F}_{\text{Critic}}$), and integrates its output into the PPO objective to achieve a state-dependent trade-off between efficiency and equity.

The input to the $\mathbf{F}_{\text{Critic}}$ is defined as a concatenation of the environment state ($\mathbf{s}_t$) and the flattened episode returns ($\mathbf{R}_t$) for all N agents (simply the input size dynamically calculated as $\text{state_dim} + n_{\text{agents}}$). This allows the critic to learn a weight that is conditioned not just on the current environment state, but also on the rewards received by the agents.

$$\mathbf{F}_{\text{Critic}}^{\text{Input}} = [\mathbf{s}_t \parallel \mathbf{R}_t] \quad (3.1)$$

AW-FIPPO algorithm requires the simultaneous training of three distinct neural networks, all optimized within a unified framework using a single Adam optimizer. The first two networks constitute the base PPO architecture [7] and FAPPO [7], responsible for the core policy and value updates:

1. Actor (Policy) Network (θ_π): This is the decentralized component responsible for

action selection. It receives local observations (z_a) and generates the action probabilities $\pi(u_a|z_a)$ for each agent a .

2. Standard Value Critic Network (θ_v): This centralized component estimates the expected return for a given global state ($\mathbf{s}$). It is used to compute the Generalized Advantage Estimate (A_a), which is essential for training both the Actor and the Critic itself via the standard PPO critic loss.
3. Novel Component: Fairness Critic Network (θ_{fc}): This is the novel, auxiliary network that implements the adaptive weighting mechanism. It is trained to predict the required level of fairness intervention, generating the adaptive scaling factor ω Equation (3.10) which controls the policy loss.

The parameters of all three networks—the Policy Network (θ_π), the Standard Critic (θ_v), and the Fairness Critic (θ_{fc})—are combined ($\theta = [\theta_\pi, \theta_v, \theta_{fc}]$) and optimized simultaneously by a single Adam optimizer. This enable the learned fairness weight (ω) to seamlessly integrated into the overall multi-agent policy optimization process. This fairness weight helps AW-FIPPO to know how and when to shift between fairness prioritization and naive reward maximization. Instead of relying on static, hard-coded fairness weight like FAPPO and AT-FAPPO [7], AW-FIPPO’s additional network dynamically learns and updates the weights in real time based on the evolving environment state and agent reward distribution. This will enable the system to navigate between fairness-efficiency tradeoffs based on the learned weights, ensuring a more context-aware and fair-efficient policy learning process.

3.2.1 Fairness Formulation: Generalized Gini Welfare Function (GGW) in MARL

The fairness formulation that this study grounded is through the use of social welfare functions [44], particularly using the generalized Gini welfare function (GGW), where fairness is directly integrated into the policy objective. GGW functions incorporate inequality-sensitive weight vectors in order to promote equitable outcomes among agents. Formally, the GGW function ϕ_ω (3.2) as formulated in [7], aggregates the individual utilities $x \in \mathbb{R}^n$ (representing the utilities of all agents) as follows:

$$\phi_\omega(x) = \sum_{i=1}^n \omega_i x_i^\uparrow, \quad (3.2)$$

where $x_i^\uparrow$ denotes the utilities sorted in non-decreasing order, and $\omega \in \Delta^n$ is a fixed weight vector with positive and non-increasing components as:

$$\omega_1 \geq \omega_2 \geq \dots \geq \omega_n > 0. \quad (3.3)$$

The decreasing structure of ω (3.3) ensures that larger weights are assigned to lower utility values, thereby prioritizing disadvantaged agents and reducing disparities in outcomes. In the learning setting, the welfare function ϕ_ω (3.2) is incorporated directly into the optimization objective. The goal is to learn a policy π_θ that maximizes social welfare [7]:

$$\max_{\pi_\theta} \phi_\omega(J(\pi_\theta)), \quad (3.4)$$

where $J(\pi_\theta) \in \mathbb{R}^n$ represents the expected joint return vector of all agents under policy π_θ .

To optimize this objective, [7] compute the policy gradient as a weighted sum of the individual policy gradients:

$$\nabla_\theta \phi_\omega(J(\pi_\theta)) = w_\sigma^\top \nabla_\theta J(\pi_\theta), \quad (3.5)$$

where $\nabla_\theta J(\pi_\theta)$ is the joint gradient of the agents' returns, and w_σ is the weight vector reordered according to the sorted values of $J(\pi_\theta)$. The reordering allow agents with lower utility values to receive higher weight w_σ . This formulation leads the policy update towards improving the utilities of disadvantaged agents.

Finally, the optimization of individual policy gradient of (3.2) is implemented via the standard PPO surrogate objective [34] (3.6), which uses the ratio ρ_θ and its clipped version $\bar{\rho}_\theta$ to ensure stable updates:

$$L_a^{\text{PPO}}(\theta_a) = \mathbb{E} \left[\min \left(\rho_a A^a, \text{clip}(\rho_a, 1 - \epsilon, 1 + \epsilon) A^a \right) \right], \quad (3.6)$$

where the importance sampling ratio is defined as

$$\rho_a = \frac{\pi_{\theta_a}(u_a \mid z_a)}{\pi_{\theta_a^{\text{old}}}(u_a \mid z_a)}, \quad (3.7)$$

and A^a denotes the advantage estimate for agent a . Therefore, the resulting policy gradient of FAPPO becomes:

$$L_a^{\text{FAPPO}}(\theta_a) = \omega_a \cdot \mathbb{E} \left[\min \left(\rho_a A^a, \text{clip}(\rho_a, 1 - \epsilon, 1 + \epsilon) A^a \right) \right] \quad (3.8)$$

This modification scales each agent's policy update by its welfare weight ω_a , leading agents assigned higher importance under the welfare function contribute proportionally more to the optimization objective. However, in [7] implementation, the authors utilized a fixed, static weight vector for the welfare function as:

$$\omega_i = \frac{1}{2^i}, \quad i = 0, \dots, N - 1 \quad (3.9)$$

While this (3.9) choice ensures that agents with lower utilities are given higher impor-

tance, it lacks adaptability to the dynamic nature of the environment state or the agents' utility distribution during training. The inability to dynamically adjust to shifting reward distributions or evolving environment state limits the method's effectiveness in real-time. Consequently, the reliance on static parameters like the one formulated in (3.9) restricts the framework's capacity to maintain fairness across diverse and evolving operational scenarios. Adaptive weights that respond to the current distribution of utilities could potentially enhance the fairness objectives, but such flexibility is not explored in [7].

3.2.2 The Novel Approach: Weight Adaptive Fairness Integrated PPO (AW-FIPPO)

The AW-FIPPO method introduces a state-dependent, learned weights $\omega(\mathbf{s}_t)$ that replaces the fixed weights used in the GGW policy objective. The Fairness Critic ($\mathbf{F}_{\text{Critic}}$), an auxiliary network predicts the necessary level of fairness intervention. The critic maps the environment state $\mathbf{s}_t$ and total episode returns $\mathbf{r}_t$ to a vector of adaptive weights:

$$\omega_t = \mathbf{F}_{\text{Critic}}(s_t, r_t) \in [0, 1] \quad (3.10)$$

The $\mathbf{F}_{\text{Critic}}$ is trained using a Mean Squared Error (MSE) loss to predict the True Gini Index ($\mathbf{G}_t$) calculated from the batch returns:

$$\mathcal{L}_{\text{FC}} = \mathbb{E}_{\text{Batch}} \left[\left(\hat{\mathbf{G}}(s_t, r_t) - \mathbf{G}_t \right)^2 \right] \quad (3.11)$$

This ensures ω is a learned indicator of the environmental inequity.

3.2.3 The Adaptive Blended Policy Objective

FAPPO policy objective in Equation (3.8) is replaced by an Adaptive Blended Objective ($\mathcal{L}_\pi$) (3.8) which dynamically combines the standard PPO [34] (3.6) Utility Term ($\mathbf{L}_{\text{Utility}}$) with a Concave Equity Term ($\mathbf{L}_{\text{Equity}}$).

The per-agent objective $\mathcal{L}_a^{\text{AW}}$ at time t is a convex combination:

$$\mathcal{L}_a^{\text{AW}}(\pi_\theta) = \mathbb{E}_t \left[(1 - \omega_t) \cdot \mathbf{L}_{\text{Utility}} + \omega_t \cdot \mathbf{L}_{\text{Equity}} \right] \quad (3.12)$$

Here:

1. Utility Term ($\mathbf{L}_{\text{Utility}}$): is the standard PPO clipped objective as shown in Equation (3.6):

$$\mathbf{L}_{\text{Utility}} = \min(\rho_\theta A_a^t, \bar{\rho}_\theta A_a^t), \quad \bar{\rho}_\theta = \text{clip}(\rho_\theta, 1 - \epsilon, 1 + \epsilon) \quad (3.13)$$

2. Equity Term ($\mathbf{L}_{\text{Equity}}$): is the concave utility function (log of the clipped advantage) that drives policy towards equity:

$$\mathbf{L}_{\text{Equity}} = \log (\min (\rho_{\theta} A_a^t, \bar{\rho}_{\theta} A_a^t) + \epsilon_{\log}) \quad (3.14)$$

The dynamic weight ω_t (3.10) controls the trade-off: when the $\mathbf{F}_{\text{Critic}}$ indicates a high need for fairness ($\omega_t \approx 1$), the policy update maximizes the equity term in Equation (3.14); otherwise, it maximizes the utility term in Equation (3.13):.

3.2.4 The Unified Optimization Objective

The final optimization problem is the minimization of the total loss, unifying the Policy, Critic, Entropy, and Fairness Critic objectives:

$$\arg \min_{\theta_{\pi}, \theta_v, \theta_{fc}} \mathcal{L}_{\text{AW-FIPPO}} = -\mathcal{L}_{\pi}(\theta_{\pi}) + c_V \mathcal{L}_V(\theta_v) - c_H \mathcal{L}_{\text{Entropy}}(\theta_{\pi}) + c_{\text{FC}} \mathcal{L}_{\text{FC}}(\theta_{fc}) \quad (3.15)$$

where, $\mathcal{L}_{\pi}$ is the adaptive policy loss, $\mathcal{L}_V$ is the PPO critic loss, $\mathcal{L}_{\text{Entropy}}$ is the entropy regularization, $\mathcal{L}_{\text{FC}}$ is the fairness critic loss, and c_V , c_H , and c_{FC} are hyper-parameters.

3.2.5 Attention-Based Fairness PPO (AT-FAPPO)

Attention-based FAPPO (AT-FAPPO) is an additional extension of FAPPO [7] that integrates the attention-based communication. This variant is introduced specifically to address the lack of inherent communication in the base FAPPO model, which relies on independent policy learning, recognizing that inter-agent communication is crucial for better coordination and enhancing fairness [7]. AT-FAPPO achieves this by incorporating a multi-head self-attention mechanism within the agent’s policy network. During training, this mechanism allows agents to learn what information to share. Each agent a computes its query (q^a), key (k^a), and value (v^a) vectors based on its local observation z_t^a [7]:

$$q_a = W_q z_t^a, \quad k_a = W_k z_t^a, \quad v_a = W_v z_t^a \quad (3.16)$$

An agent then computes attention scores by taking the inner product of its query vector with the key vectors from all neighboring agents (including itself), which quantifies the compatibility or relevance between observations. These scores are scaled and normalized using a softmax function to obtain attention weights α' [7, 45]. The final output vector b^a for the agent is a weighted sum of the value vectors, $b^a = \sum \alpha'_{a,j} v^j$ [7, 45]. This communication mechanism allows agents to coordinate their actions more effectively, which helps resolve non-stationarity issues and leads to more fair solutions by incorporating relevant information sharing. This resulting vector b^a is then fed into the policy

network to determine the agent’s final action. This mechanism resolves non-stationarity and mitigates unfairness by allowing agents to coordinate effectively. However, attention weights are completely different from the fairness weights used in the welfare function. The attention weights are about optimizing communication and cooperation by highlighting relevant information, while the fairness weights are used to ensure equitable rewards across agents. Although the attention mechanism in AT-FAPPO facilitates communication and information sharing among agents during training, the fairness weights ω within the welfare function is still fixed and predefined the same way in FAPPO.

Therefore, we replaced these static fairness weights ω with dynamic, learned weights that are updated throughout training in both variants as explained in Sub section 3.2.2. This means that instead of assigning a rigid decreasing order of weight for fairness prioritization, we use a learned weights to decide the learning flow of the policy network based on the environment state and the current utilities of agents. Even though the attention mechanism enables agents to learn what information to share and how to coordinate effectively, its effectiveness is still limited due to its reliance on a static weights for fairness integration. our fairness weighting mechanism enhance this limited performance by integrating a dedicated network that is responsive to the change in environment state and evolving agent utilities. The learned weights enable the welfare function to emphasize agents’ needs more accurately in real-time, promoting fairness more effectively as the system encounters different states and utility distributions. This integration of learned weights with the attention-based communication allows for a more nuanced approach where both the sharing of information and the fairness assessment are adaptable, thus enhancing the overall fairness and performance of the policy learning process in AT-FAPPO.

The summary in Table 3.1 further highlights the mathematical and structural distinctions between the baseline FAPPO and our adaptive framework (AW-FIPPO). While FAPPO incorporates fairness objective, it optimized it through a fixed welfare weight that may prematurely sacrifice total reward for marginal gains in equity. Our adaptive method, however, leverages the Fairness Critic to generate the welfare weights dynamically. These dynamic weights, therefore, are used as a base to selectively apply log concave utility transformation for fairness integration only when necessary. The transition logic encoded in the adaptive objective allows the agents to naively maximize reward in a balanced state while applying the log transformation when the distribution is unfair. The evaluation demonstrated that our adaptive method consistently achieves higher collective returns than FAPPO while maintaining superior Gini Index scores.

Table 3.1: Summary: Comparative Analysis of FAPPO and Weight Adaptive PPO (AW-FIPPO)

Characteristic	Fairness-aware PPO (FAPPO)	Weight Adaptive fairness integrated PPO (AW-FIPPO)
Core Concept	Static fairness constraint.	Dynamic, State dependent fairness objective.
Gini Weight Vector ($\mathbf{w}$)	Fixed ($\mathbf{w}$); manually tuned (e.g., $w_i = 1/2^i$).	Adaptive ($\mathbf{w}_t$); learned online as a function of the state.
Weight Determination	Hand-crafted, chosen heuristically based on worst-off agent preference.	Learned by a dedicated Fairness Critic Network ($\mathcal{F}$) parameterized by θ_{fc} .
Objective Function	Fixed Generalized Gini Welfare: $J = \phi_{\mathbf{w}}(\mathbf{J})$.	Adaptive Welfare Weighting: $J_t = \phi_{\mathbf{w}_t}(\mathbf{J})$, where $\mathbf{w}_t = \mathcal{F}(s_t, r_t; \theta_{fc})$.
Sensitivity to Inequality	Constant across the environment states (s_t) and reward distribution (r_t).	Dynamic, adjusted in real-time based on the environment state s_t and reward distribution r_t to modulate the efficiency-equity trade-off.
Algorithm Architecture	Two networks: Policy (θ_π), Critic (θ_v).	Three Networks: Policy (θ_π), Critic (θ_v), Fairness Critic (θ_{fc}).

3.3 The Proposed AW-FIPPO Architecture Detail

The proposed architecture adopt the CTDE learning paradigm [6]. And also it introduce an auxiliary Fairness Critic Network alongside the standard PPO actor-critic modules. As illustrated in the Figure 3.2, the auxiliary network receives a concatenated input vector containing the global state s_t and the individual return distribution r_t of all agents. This module is trained to approximate the Gini Index of the system, outputting a dynamic fairness weight (ω_t). This weight acts as a high-level coordination signal that modulates the policy objective $\mathcal{L}_\pi$. In an unbalanced state where the auxiliary network detects reward disparity ($\omega_t \approx 1$), the system prioritizes Equity Mode by applying a logarithmic transformation to the PPO surrogate to enforce fairness-oriented gradient updates. Alternatively, in balanced states, the system maintains Utility Mode, utilizing the linear surrogate to maximize cumulative team rewards. This adaptive feedback loop ensures that the global policy remains sensitive to agent inequality without compromising the overall efficiency of the multi-agent system.

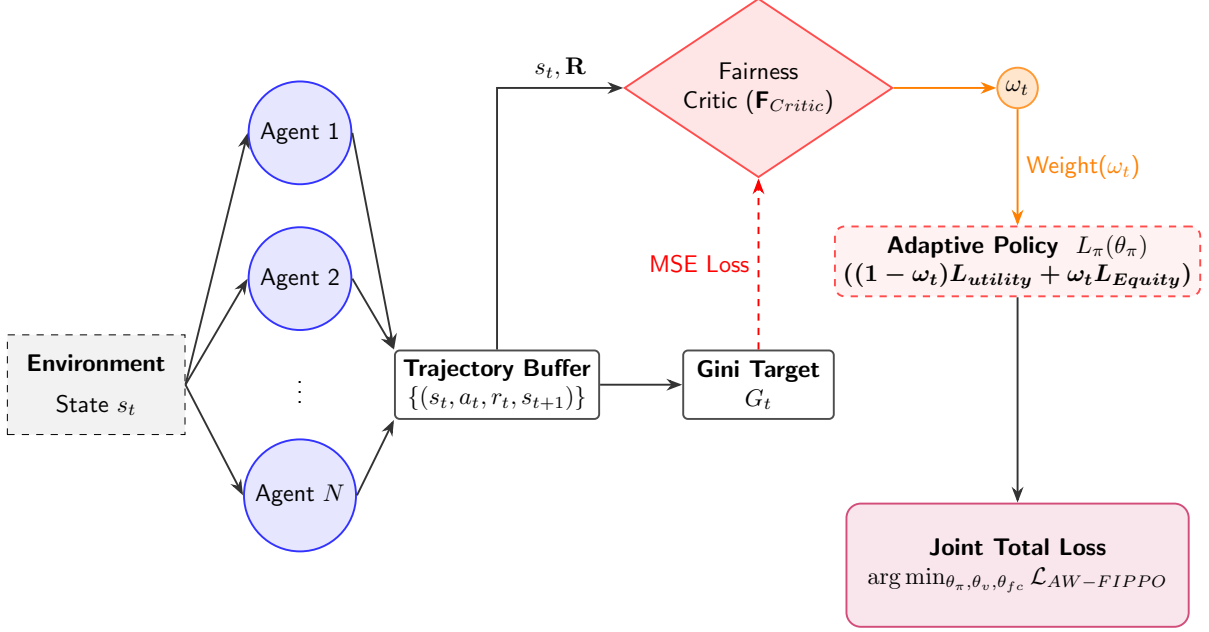


Figure 3.2: Architecture of our AW-FIPPO framework with decentralized policies and adaptive welfare-based optimization objective

3.4 Fairness Objectives

Fairness plays a central role here, as the goal is not only to achieve high collective returns but also to ensure an equitable distribution of rewards among agents. The distinct way fairness weights are derived and integrated into the optimization process distinguishes this study from the baseline FAPPO approach. In this section, two fairness formulations are described: the baseline FAPPO approach that uses a fixed, strictly decreasing positive weight vectors in the generalized gini welfare function and the proposed method that replaces this fixed vectors with learned fairness weights generated by a neural network.

3.4.1 Fairness Formulations of Baseline Paper

In the baseline FAPPO framework, fairness is enforced using the generalized Gini welfare function, which is mathematically expressed as:

$$\phi_{\omega}(\mathbf{x}) = \sum_{i=1}^n \omega_i x_i^{\uparrow} \quad (3.17)$$

where $\mathbf{x} \in \mathbb{R}^n$ represents the vector of utilities of the agents in a sorted non-decreasing order ($x_1^{\uparrow} \leq \dots \leq x_n^{\uparrow}$), the weight vector $\omega \in \Delta_n$ contains positive, strictly decreasing weights ($\omega_1 > \dots > \omega_n > 0$), larger weights on smaller utility emphasize the utilities of less advantaged agents, thus encouraging fairness and equity.

In FAPPO experiments, the weights for the welfare function are determined using a

fixed, deterministic mathematical formula, which results in a simple weight vector. This vector assigns a specific importance to each agent’s utility based on its position when utilities are ordered from smallest to largest. The formula employed is:

$$w_i = \frac{1}{2^i}, \quad i = 0, 1, \dots, n - 1 \quad (3.18)$$

where n , denotes the total number of agents in the given environment. This means that the weight for the i^{th} utility decreases exponentially as i increases, effectively assigning higher weights to agents with lower utilities — that is, those who are less advantaged. This way of equity mechanism of the baseline FAPPO framework is adapted from the Pigou-Dalton principle of fairness [31], which emphasizes transferring utility towards the worse-off agents from a more advantaged one to promote equity.

Although this way the baseline FAPPO framework improves the outcome of the least advantaged agents within the environment, the weights (3.18) used for the adjustment are rigid and static throughout the entire training. This predetermined weight usage lacks the necessary adaptability and flexibility to respond to the changing environment state or to the evolving reward disparity among agents, leading the system to fail to adequately capture the nuances or dynamic nature of real-world settings.

3.4.2 Fairness Formulations of This Study

The fairness formulation of this study is grounded in replacing the static weighting strategy we discussed above and aimed at employing an adaptive mechanism that dynamically produces fairness weights, addressing the static weighting schemes utilized in the baseline FAPPO framework. Therefore, by incorporating context-dependent fairness weights, we seek to overcome the rigidity of the heuristic approach in Equation (3.18), allowing for a more dynamic integration of fairness metrics.

For that we introduce a dedicated Fairness Network $\mathcal{F}$, parameterized by $\theta_{\mathcal{F}}$. Unlike traditional regularizers that apply a static penalty for inequality, $\mathcal{F}$ acts as an estimator of the weights for balancing equity and efficiency. The network takes the environment state s_t and the vector of individual agents return r_t as an input, the output of this network is a normalized vector of fairness weights:

$$\omega_t = \mathcal{F}(s_t, r_t; \theta_{\mathcal{F}}) \in [0, 1] \quad (3.19)$$

These fairness weights are then used to modulate the importance of each agent’s utility in real-time. By conditioning on (s_t) and (r_t) , the network generate fairness weights that represent the state of the reward distribution among agents. This contextual awareness enables the policy to dynamically shift between Utilitarian efficiency and Egalitarian equity in real-time. Instead of relaying on a fixed, predefined weights, the system utilizes

those dynamic fairness weights ω_t (3.19) for the transition. The per-agent objective $\mathcal{L}_a^{\text{AW}}$ at time t is defined as a convex combination:

$$\mathcal{L}_a^{\text{AW}}(\pi_\theta) = \mathbb{E}_t [(1 - \omega_t) \cdot \mathbf{L}_{\text{Utility}} + \omega_t \cdot \mathbf{L}_{\text{Equity}}] \quad (3.20)$$

In this formulation, $\mathbf{L}_{\text{Utility}}$ corresponds to the standard PPO surrogate $\min(\rho_t A_t, \text{clip}(\rho_t) A_t)$, while $\mathbf{L}_{\text{Equity}}$ represents the concave log-transformation $\log(\mathbf{L}_{\text{Utility}} + \epsilon)$.

The weight ω_t determines the operational mode for each agent. When $\omega_t \approx 1$ (indicating the agent is part of a high-disparity distribution), the system prioritizes Equity Mode by applying the concave log-transformation to the advantage-based surrogate, forcing gradients to favor the performance of disadvantaged agents. When $\omega_t \approx 0$ (indicating a fair reward distribution among agents), the system maintains Utility Mode by utilizing the standard PPO objective to maximize cumulative rewards. This ensures that the global policy remains sensitive to agent inequity without compromising the overall efficiency of the multi-agent system.

The parameters of the Fairness Network θ_{fc} are jointly optimized alongside the Actor (θ_π) and Critic (θ_V) parameters of the PPO agents. The resulting parameter set $\Theta = \{\theta_\pi, \theta_v, \theta_{fc}\}$ is updated via gradient descent to ensure that the Fairness Switch—the transition between prioritizing total reward and prioritizing the performance of lagging agents—is an emergent, learned property of the system.

The input utilized by the fairness critic network is derived from the episodic rewards r_t and the global state s_t , ensuring that the weighting mechanism remains sensitive to the long-term trajectory of the multi-agent system rather than just immediate, local rewards.

3.5 Benchmark Environment

The study bench-marked the proposed approach against key MARL baselines including value-based methods such as Value-Decomposition Networks (VDN) and Monotonic Value Mixing Network (QMIX), policy-based methods such as Counterfactual Multi-Agent Policy Gradients (COMA), as an independent learning baseline such as IPPO, and FAPPO (with its communication-enabled variant AT-FAPPO) to ensure a comprehensive comparison of both performance and fairness.

The experimental evaluation was conducted in two fairness-sensitive environments: the Random MDP and the Matthew Effect scenario. These environments were initially designed and employed by [7, 13]. To ensure a direct comparison and maintain rigorous testing criteria, we benchmarked our proposed Weight Adaptive Fairness Integrated PPO (AW-FIPPO) algorithm against established baselines within the same multi-agent environments. Both environments represent cooperative multi-agent domains designed to assess resource allocation and fair coordination under conditions that naturally give

rise to resource accumulation tendencies and unfair reward distribution.

3.5.1 Random MDP

We employ a Random MDP experimental setting, a controlled yet challenging testbed designed by [7] to intentionally introduce randomness by assigning rewards randomly across agents, thereby creating potential disparities in outcomes. The environment is configured as a constrained 5×2 grid-world, facilitating a discrete state-space for three cooperative agents [7]. To ensure computational tractability, states are represented via one-hot encoding of agent coordinates. The action space $\mathcal{A}$ is consistent across the population, comprising cardinal movements and a stationary (idle) command. System dynamics are defined by a shared transition matrix $\mathcal{P}$, stochastically initialized from a uniform distribution. This shared transition logic implies that the environmental laws such as the probability of a move succeeding or failing are identical for all agents, creating a baseline of operational symmetry.

To elevate the environmental complexity, the environment adopts a partially observable setting where global state transparency is withheld. Each agent i is constrained to a local observation $o_i \in \mathcal{O}$, precluding direct access to the joint state $\mathcal{S}$. Despite this partial observability, the environment remains cooperative since the total reward used for training is computed as the sum of each individual agent’s rewards [7]. This design challenges fairness-aware MARL systems, as it introduces an inherent trade-off. The system’s policies are expected not only to maximize the overall cumulative reward but also to ensure equitable reward distribution among agents. The goal is to strike a balance between efficiency (maximizing total return) and equity (achieving fairness in distribution). Therefore, the experiments in this environment are aimed at assessing the capability of fairness-aware methods, including our weight-adaptive approach, to maximize the system-wide efficiency while minimizing disparities in reward distribution among agents.

3.5.2 Matthew Effect

We utilize a Matthew Effect experimental setting, which is designed by [13] to model a social and economic phenomenon often summarized as “the rich get richer and the poor get poorer.” This phenomenon reflects a setting where an agent that begins with an advantage or happens to obtain early rewards can easily accumulate additional rewards compared to those who have no initial opportunity, thereby widening distribution inequity among agents [7]. In the environment, there are ten Pac-Man (agents) and three stationary ghosts that interact within a shared space. The agents are initialized with heterogeneous attributes, including differing positions, sizes, and speeds, while the ghosts are initialized randomly.

Each agent’s observation is restricted to other three nearest neighboring agents and one closest ghost, reflecting realistic limitations in real world. The state representation contains each agent’s position, speed, and size, as well as the positions of the ghosts. Each agent $i \in \mathcal{N}$ is modeled to operate within a discrete action space including moving up, down, left, right, and remaining idle. The state transition function is governed by the relative spatial proximity between agents and ghosts. A successful transition occurs when the Euclidean distance $d(a_i, g_j)$ falls below the agent’s collision radius, triggering a sparse reward signal $r_i = 1$. This event induces a state change in the agent, specifically through an incremental increase in its morphological size and velocity parameters bounded by a predefined threshold while the consumed ghost is stochastically redistributed within the environment [13].

The environment inherently fosters a self-reinforcing bias, as high-performing agents’ size and velocity increase, resulting in a reduced relative difficulty in capturing subsequent ghosts. This effect creates a positive feedback loop that lowers the marginal cost of reward acquisition for dominant agents, marginalizing unsuccessful agents and leading to a ‘winner-take-all’ convergence pathology. Therefore, the environment challenges the fairness-aware approach’s ability to maximize the number of ghosts consumed collectively and, at the same time, their ability to avoid runaway inequity in consumption. In other words, fair and efficient approaches should ensure that the consumptions are evenly distributed across agents while maintaining the maximum collective consumption.

3.6 Evaluation Metrics

In order to assess the effectiveness of our weight-adaptive method in balancing the fairness and efficiency tradeoff, we employed complementary quantitative metrics for each side of the tradeoff. Because a higher system return does not guarantee equitable reward distribution within the system, it is insufficient to rely solely on efficiency metrics as the only metrics of the system. Therefore, we utilize structured efficiency and disparity metrics to evaluate our approach’s capability compared to the baselines. While efficiency metrics measure the system’s ability to maximize aggregate utility gains, our disparity metrics audit the distribution of those gains, ensuring that this gain is not achieved by suppressing low-utility agents to inflate the performance of a dominant minority. These structured assessments ensure that improvements in aggregate utility are not achieved through reward concentration by few agents and that fairness incorporation into the system does not come at the expense of substantial efficiency degradation.

3.6.1 Efficiency and Utility Metrics

Even though we aim to maintain distributive equity, it is also essential to ensure that our fairness constraint does not compromise the system’s fundamental productivity. To evaluate this balance, the total cumulative reward is used as our primary indicator of the system’s efficiency, as it directly reflects the overall performance achieved by the multi-agent system under cooperative learning policy. In this setting, efficiency is defined as the aggregate return accumulated across all agents throughout training, thereby capturing the system’s capacity to maximize collective utility.

Total Reward (R_{total})

It is calculated as a summation of individual agent’s cumulative reward r_i across a population of n agents:

$$R_{total} = \sum_{i=1}^n r_i \quad (3.21)$$

where n denotes the total number of agents and r_i represents the cumulative reward of agent i . Higher values indicate that agents collectively achieve more in the environment. To assess whether our adaptive fairness constraint introduces any unintended performance degradation, we explicitly compare the total reward trajectory of our method against baseline approaches that do not incorporate dynamic welfare adjustment.

While R_{total} effectively captures the aggregate performance level of the system, it is insensitive to the distribution of rewards among agents. It may hide severe disparities or even the systematic failure of individual agents by providing high team reward which could not guarantee high performance of each individual.

Maximum Reward (r_{max})

Defined as $r_{max} = \max\{r_1, \dots, r_n\}$, this metric identifies the upper bound of agent performance within the system. While achieving a high r_{max} indicates the environment’s potential for high utility, a large discrepancy between r_{max} and r_{min} is a primary indicator of winner take all dynamics and agent exploitation.

Minimum Reward (r_{min})

Defined as $r_{min} = \min\{r_1, \dots, r_n\}$, this represents the Rawlsian Maximin Criterion. In a fair system, the quality of the policy is judged by the status of the least-advantaged agent. A higher r_{min} indicates a robust coordination strategy that prevents agent starvation or permanent role-marginalization.

Even though the maximum and minimum rewards are straightforward absolute values, their value provides a foundational disparity analysis. A widening gap between the best-off agents or highest $r_{\max}$ value and worst-off agents or low $r_{\min}$ value indicates a divergent learning path, where the system is likely sacrificing the progress of low-performing agents to maximize the performance of a few high-performing outliers. In contrast a narrow gap between $r_{\min}$ and $r_{\max}$ values indicates that the system has achieved a more equitable distribution. This metric is especially relevant in environments such as the Matthew effect, where inequity can compound over time.

3.6.2 Disparity and Distribution Metrics

As our aim is to maintain a balanced tradeoff between fairness and efficiency, it is essential to evaluate our approach using disparity metrics in addition to the above-discussed efficiency metrics. Achieving the highest total reward alone as an efficiency metric does not guarantee the fairness of the system because this highest result may be gained by the policy that incentivizes a few dominant agents to accumulate rewards at the expense of the system’s equity and collective stability. Therefore, in order to assess the fairness of the system in distributing rewards among agents, we utilized structural disparity metrics: the Gini Index and Coefficient of Variation (CV) as measures to quantify disparity in reward distribution.

Gini Index (G)

We utilize the Gini index as a primary disparity metrics to quantify the degree of reward imbalance among agents. The Gini index measures the average pairwise difference between the rewards of all the agents, normalized by the mean to capture return disparity within the group. A Gini index with a value of zero indicates perfect equality, where all agents receive identical rewards, whereas a Gini index with a value closer to 1 reflects a higher disparity in reward distribution. In the context of MARL, where agents often compete for finite resources, the Gini Index identifies the emergence of dominance effects by identifying the specific gap dynamics between top-tier and bottom-tier agents. This allows for a more granular assessment of whether our proposed adaptive framework achieves a Pareto-optimal balance between aggregate utility and an egalitarian reward distribution.

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |r_i - r_j|}{2n^2\mu} \quad (3.22)$$

Coefficient of Variation (CV)

To move beyond purely distributive metrics, we incorporate the Coefficient of Variation (CV) as a rigorous, scale-invariant measure of relative dispersion. By expressing systemic variability as a ratio of the group mean, CV provides a dimensionless lens that remains robust against the natural inflation of rewards observed during MARL convergence. This dual-metric approach utilizing Gini for social welfare and CV for statistical dispersion allows us to detect deep structural inequalities that might otherwise be masked by high aggregate utility. Ultimately, the CV enables us to verify if our agents are truly converging toward a stable, equitable equilibrium or if collective gains are being driven by a subset of dominant agents.

$$CV = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (r_i - \mu)^2}}{\mu} \quad (3.23)$$

By analyzing all these metrics, we aim to determine whether our weight adaptive method successfully mitigates disparities without undermining overall system efficiency.

Welfare Score (W)

As an addition, we employed the below welfare score metric to jointly capture efficiency and equity within the multi-agent system. we define a welfare score as a metric that integrates both the average group performance and the degree of reward inequality.

$$W = \mu \cdot (1 - G) \quad (3.24)$$

where μ is the mean reward ($\frac{R_{total}}{n}$) and G is the Gini Index. This provides a single, scalarized Efficiency-Fairness score, ensuring that an algorithm cannot achieve a high W through high R_{total} alone if the Gini Index is also high. Together, these metrics provide a comprehensive evaluation framework. Total reward ensures the system achieves its operational goals, while the relationship between maximum and minimum rewards highlights the severity of internal difference.

Chapter 4

Experimentation

4.1 Implementation Details

The implementation of the proposed method builds up on the IPPO framework, designed for multi-agent cooperative environments. Each agent independently learns its own stochastic policy. Instead of optimizing the traditional sum of discounted rewards, recent approaches like FAPPO focus on maximizing a carefully designed welfare function particularly generalized Gini welfare function that promotes fairness among agents. In FAPPO the weights in the generalized Gini welfare function were originally computed as an Equation (1.1) a fixed vector and applied directly within the PPO loss function.

However, in our proposed method, these explicit computations are replaced by calls to a dedicated Fairness Network, which dynamically infers and generates agent specific weighting coefficients at each policy update step. Therefore the core novelty of this study is the introduction of this dedicated Fairness Network, which replaces these static Gini weights with a learned, dynamic fairness weight vectors. By conditioning the weights on the evolving utility profiles of the agents, the Fairness Network enables adaptive modulation of fairness constraints throughout training, allowing the learning process to respond to changing disparities rather than relying on fixed, pre-defined weighting schemes.

The Fairness Critic Network is implemented as a sequential MLP: it takes an input vector consisting of the environment state concatenated with the episodic reward collected by each agent. It processes this input through a lightweight feedforward neural network that feature a 64-unit hidden layer followed by Tanh activation to capture non-linear correlation between state and reward. The output layer utilizes a Sigmoid activation function to strictly constrain the generated fairness weights within the range of $(0, 1)$. Therefore the policy network is updated based upon these weights. If the Fairness Critic Network detects fair reward distribution (when the weights approach to 0), the policy moves towards the standard utility maximization, else if it detects high inequality (the

weights approach to 1) the policy prioritize concave logarithmic transformation. The Fairness Critic parameters (θ_{FC}) are updated to minimize the Mean Squared Error (MSE) between the predicted fairness weight and the actual observed distribution outcome of the environment, leading to accurate fairness weight prediction. The relative simplicity of the architecture ensures that the additional computational overhead introduced by the fairness network is minimal which is important for maintaining the scalability of the training process.

During training, advantages are normalized separately for each agent to prevent any single agent’s advantage scale from dominating the optimization process. PPO clipping is applied to the policy losses to ensure stable and consistent updates. And also normalization is used to mitigates issues of non-stationarity and varying reward scales, which are particularly relevant in MARL settings where agents can experience highly heterogeneous reward distributions. Training stability is ensured by adopting the Adam optimizer, which is widely used in deep reinforcement learning due to its adaptive learning rate capabilities and robustness to noisy gradient estimates. Therefore the policy, the value estimation, and the fairness network are optimized jointly using this optimizer, wherein the system seeks to learn policy parameters θ that maximize a state-dependent social welfare function.

The training hyper-parameters are selected in accordance with the specifications reported in [7]. Our implementation utilizes the PYMARL2 framework [46], whose source code is publicly available at <https://github.com/hijkzzz/pymar12>. Users of this codebase are encouraged to also cite the original PYMARL2 work. We extend the baseline implementation with additional fairness integrated components. The complete set of algorithm-specific components and training hyper-parameters is summarized in Table 4.1.

Parameter	Value	Description
Total Timesteps ($t_{\max}$)	6,050,000	Maximum number of environment steps for training.
Learning Rate (lr)	0.0005	Used for the Adam optimizer across all three networks.
Discount Factor (γ)	0.99	Used for calculating discounted returns.
GAE Factor (λ_{GAE})	0.95	Used for calculating the Generalized Advantage Estimates.
PPO Clipping (ϵ_{clip})	0.2	Clipping range for the probability ratio in the PPO loss.
Episodes per Collection	8	Number of episodes or optimization updates performed on each collected batch of data.
Critic Coefficient (c_1)	0.5	Weighting factor for the value function (critic) loss.
Entropy Coefficient (c_2)	0.01	Weighting factor for the entropy loss, encouraging exploration.
Fairness Critic Coef. (c_3)	0.1	Weighting factor for the Fairness Critic (FC) loss.
Agent Input	obs_agent_id: True	Agents receive their ID as part of their observation.

Table 4.1: Hyper-parameters and Configuration Settings for Training

When inter-agent communication is introduced, AW-FIPPO extended to the Attention-based AW-FIPPO (AT-AWFIPPO) variant the same way as AT-FAPPO [7]. It incorporates a multi-head self-attention mechanism inspired by the transformer architecture [7, 45]. This variant is engineered to be fully compatible with the dynamic fairness weight objective of this study, meaning it optimizes the adaptive welfare objective using the learned, context-dependent weight vector ω_t generated by the Fairness Critic Network. The self-attention module is integrated into the policy network, allowing agents to selectively focus on and exchange the most relevant information (encoded as value vectors) with others, thereby quantifying the importance of neighbor observations via learned attention weights. The combination of learned attention weights from the self-attention module for relevant information exchange and fairness weights from the Fairness Critic Network for reward distribution detection among agents together brings optimal equilibrium where high-throughput performance and systemic fairness are mutually achieved.

4.2 Experimental Results

4.2.1 Performance Evaluation in the Random MDP Environment

The experimental results in this environment indicate that dynamically learning fairness weights enables the policy to achieve higher overall welfare while maintaining lower inequality levels among agents.

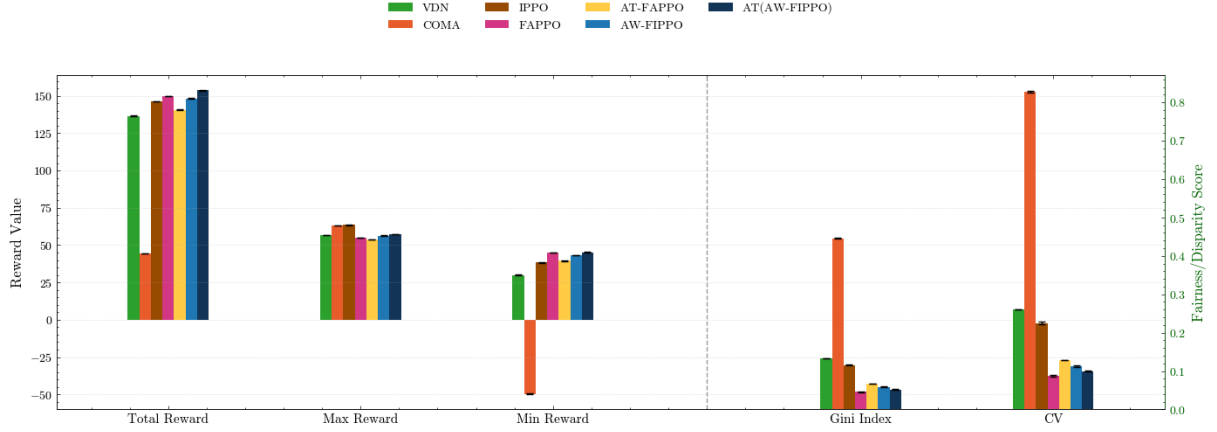


Figure 4.1: Random MDP aggregated performance result illustrating efficiency (total reward, minimum reward, maximum reward, and welfare score) and disparity (coefficient of variation and Gini index).

As shown in Figure 4.1, the proposed method, particularly AT-AWFIPPO, consistently yields the highest average total reward, outperforming all other methods. And also it achieves the highest Min reward (the strongest indicator of preventing agent starvation) and the highest welfare scores over the entire training duration as shown in Figure 4.2.

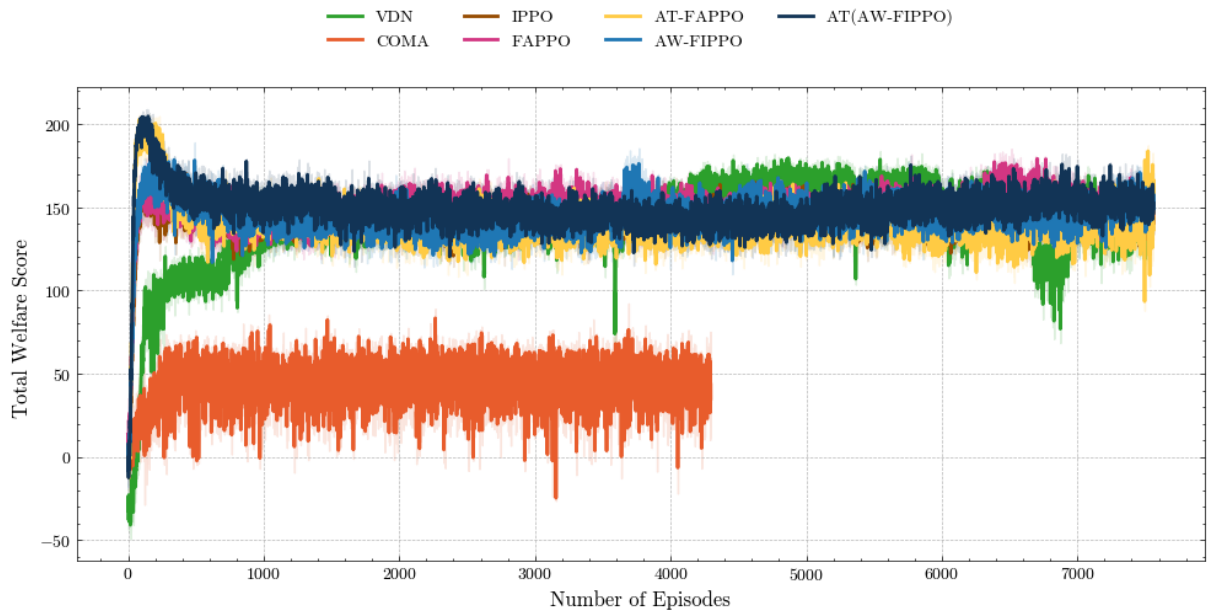


Figure 4.2: Random MDP welfare scores.

AT-AWFIPPO maintains a stable peak welfare score, demonstrating the best performance by effectively combining attention based communication with the learned dynamic fairness objective. The baseline FAPPO and AT-FAPPO perform well but are clearly outperformed by the proposed approach. This demonstrates that learning a dynamic fairness objective is not just compatible with maximizing efficiency but can enhance it, particularly when combined with inter-agent communication. The Fairness Critic Network identifies agents with low returns and triggers the Equity Mode (log-loss). This ensures that the policy gradient prioritizes the Min Rewarded agent. Unlike COMA, which can ignore a single failing agent if the team total is high, our adaptive loss prevents any agent’s reward from collapsing, leading to a higher system-wide Welfare Score.

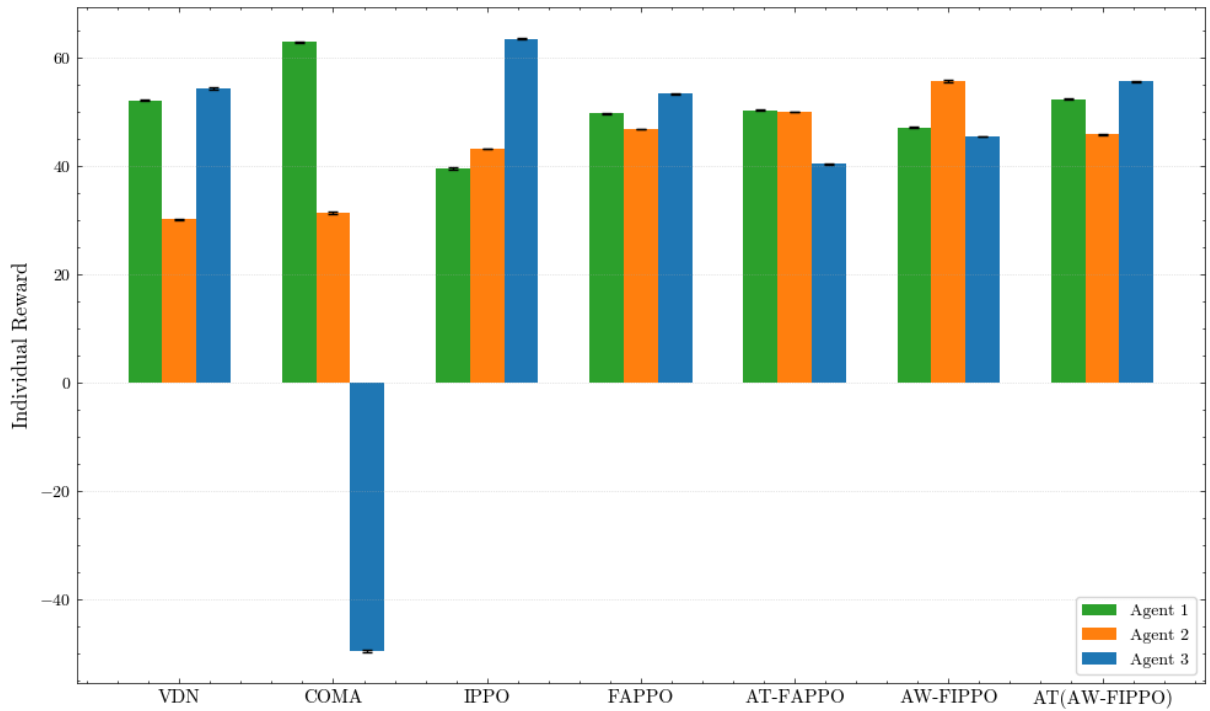


Figure 4.3: Individual rewards across the baselines.

Table 4.2: The mean Reward distribution among individual agents

Algorithm	Agent 1	Agent 2	Agent 3
VDN	52.10	30.16	54.28
COMA	62.89	31.34	-49.64
IPPO	39.60	43.19	63.52
FAPPO	49.68	46.81	53.31
AT-FAPPO	50.28	50.04	40.38
AW-FIPPO	47.11	55.67	45.49
AT(AW-FIPPO)	52.33	45.86	55.58

A comparative analysis of the individual agent rewards across the baselines as shown in Figure 4.3 and in Table 4.2 reveals that AW-FIPPO and AT(AW-FIPPO) achieve a balance between collective efficiency and individual agent contribution by directly elevating the performance of the weakest participants among baselines.

While baselines like COMA and VDN exhibit severe marginalization—with the least performing agents dropping as low as -49.64 and 30.16 respectively, FAPPO and AT-FAPPO variant’s aim to maintain equality resulted in lower individual outcomes, as shown in AT-FAPPO’s minimum agent reward of 40.38 . This suggests that the agents stay relatively close to one another, but the system fails to unlock the full potential of the environment, resulting in a balanced but lower performance compared to AW variants. AW-FIPPO and AT(AW-FIPPO) methods ensure that the lowest-earning agent is significantly more productive, with AT(AW-FIPPO) reaching a minimum of 45.86 . This indicates that among all tested methods, our approach provides the highest performance floor, ensuring that every agent maintains a high reward level.

This equitable yet high-performance result is the direct consequence of the Adaptive Objective Switch. By applying standard PPO to the policy loss during balanced states, our approach enables each individual to secure larger rewards compared to the least successful agents in other methods. Simultaneously, in an unbalanced state the application

of the concave log-transformation reduces the marginal utility of the dominant agent’s actions, incentivizing the system to prioritize cooperative strategies that disproportionately benefit low-performing agents. This confirms that the dynamically learned weights successfully steer the policy toward a Pareto-superior outcome, ensuring the most equitable distribution while maintaining high performance across the entire agent population.

Additionally, we included the plot for total reward against fairness scores, to reveals how improvements in one dimension may influence the other, providing deeper insight into the balance between efficiency and equity in cooperative multi-agent reinforcement learning.

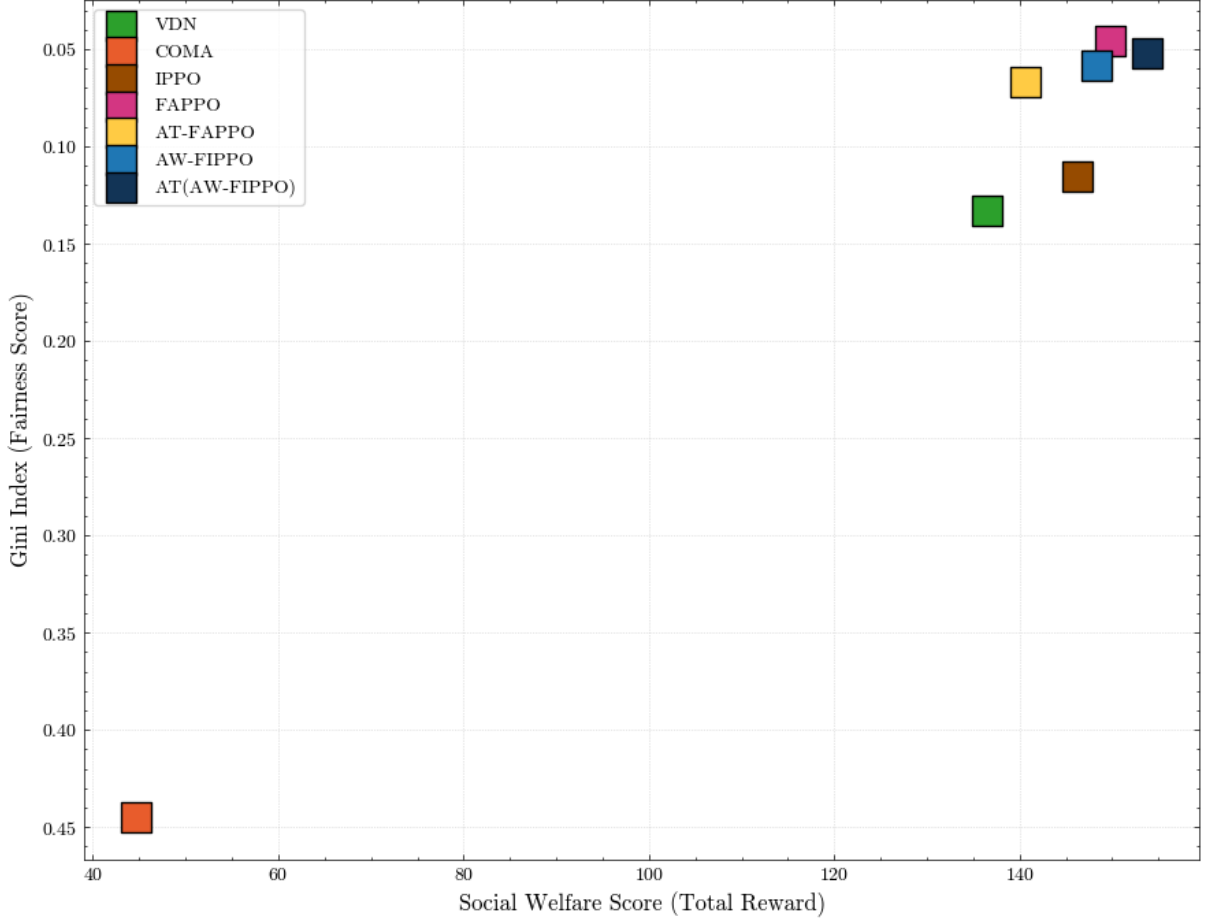


Figure 4.4: Efficiency-equity trade-off(Random MDP): Total Reward vs. Gini index

In the trade-off scenarios shown in Figure 4.4, our method clusters in the top-right quadrant (High Total Reward, Low Gini), whereas methods like VDN, COMA and IPPO are dispersed toward the high-disparity region. They depict a follow of linear trade-off where increasing total reward naturally increases the Gini index (inequality). By dynamically re-weighting the objective, our method achieves a more efficient frontier where high performance does not necessitate high inequality.

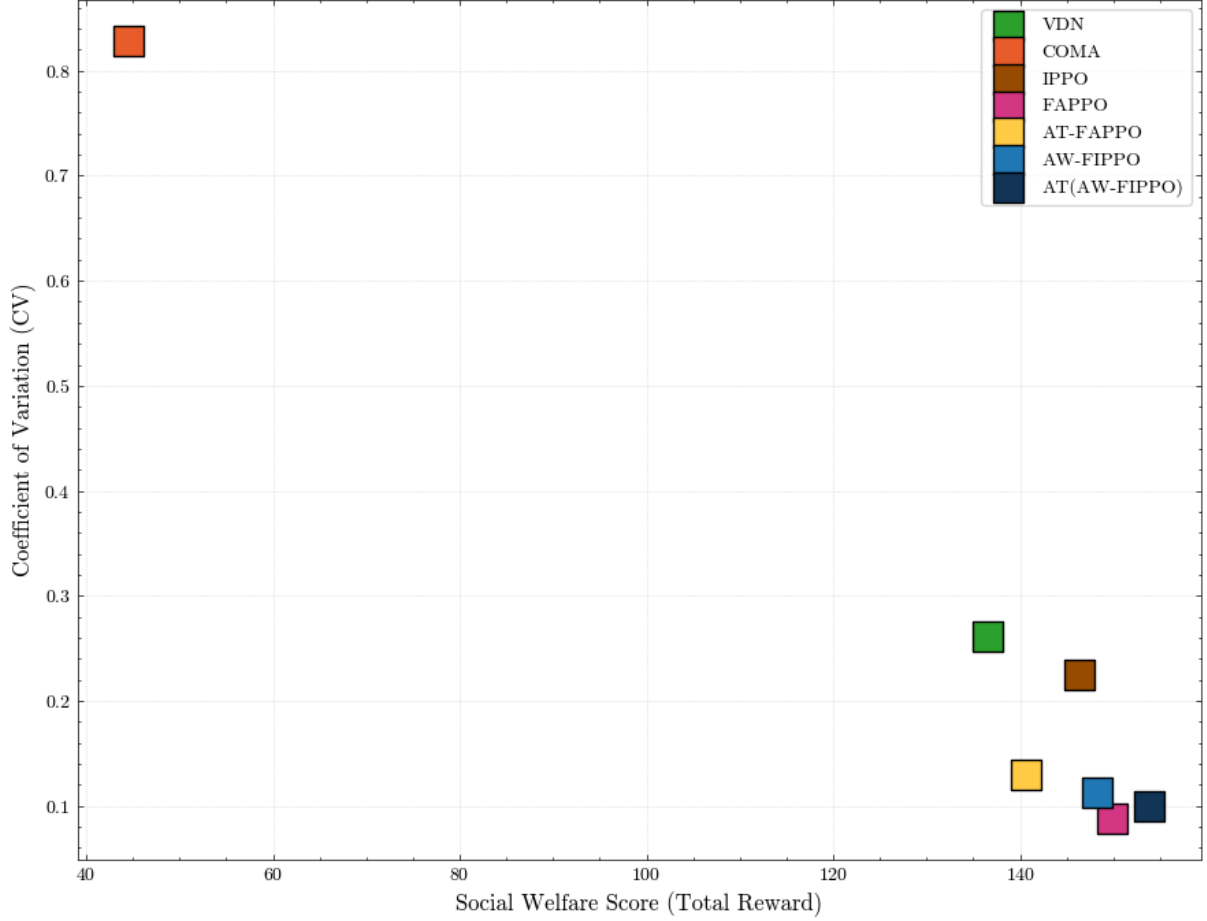


Figure 4.5: Efficiency-equity trade-off(Random MDP): Total Reward vs. CV

Regarding the total reward versus CV trade-off depicted in Figure 4.5, our method demonstrates a significantly smaller variance (tighter error bars) in the CV dimension when compared to FAPPO. The Fairness Critic acts as a Stabilization Mechanism. Because it continuously approximates the Gini Index ($\mathcal{G}(\mathbf{R})$), it can apply corrective weights as soon as the CV starts to spike. Standard PPO baselines have no such feedback loop, making them highly susceptible to environmental noise that might accidentally favor one agent over another during a single training seed.

4.2.2 Performance Evaluation in the Matthew Environment

The experimental results in the Matthew environment show a clear distinction between utilitarian-focused baselines and our equity integrated framework. The Matthew Effect—where agents that gain an initial advantage continue to accumulate disproportionate rewards is visibly mitigated across the following visualizations:

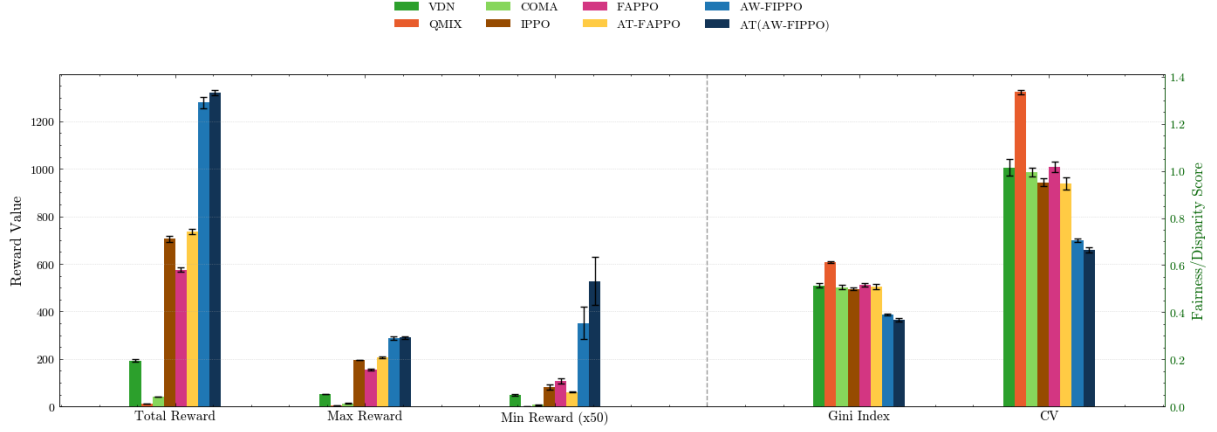


Figure 4.6: Matthew aggregated performance results illustrating efficiency (total reward, minimum reward, maximum reward, and welfare score) and disparity (coefficient of variation and Gini index).

As evidenced by Figure 4.6, the proposed method (AW-FIPPO and AT(AW-FIPPO)) consistently yields the highest average total reward (over 1000), max and min reward, outperforming all other baseline methods. This demonstrates that learning a dynamic fairness objective is not just compatible with maximizing efficiency but can enhance it. In the minimum reward, which is the strongest indicator of preventing agent starvation, our approach achieves the highest minimum reward, substantially surpassing the original FAPPO and AT-FAPPO, and dramatically exceeding baselines like IPPO. This validates the success of the learned objective in protecting the worst-off agent.

The CV and gini index measures reward disparity; lower is better. Our approach achieves the lowest CV and gini index across all methods effectively near-zero (below 1.0) as reflected in Figure 4.6. This result is significantly better than the FAPPO and AT-FAPPO result, confirming that dynamically adjusting the welfare weights is the most effective mechanism for minimizing inequality.

The core achievement of the adaptive weighting (AW-FIPPO and AT-AWFIPPO) lies in its ability to avoid sacrificing the highest-rewarding agents entirely while strongly focusing on raising the minimum reward. This careful balancing act is essential for maximizing the Welfare Score, demonstrating that the learned weights strategically manage the trade-off, preventing the high performers' rewards from collapsing, yet still ensuring a highly equitable outcome that dramatically outperforms the original, static-weighted FAPPO methods.

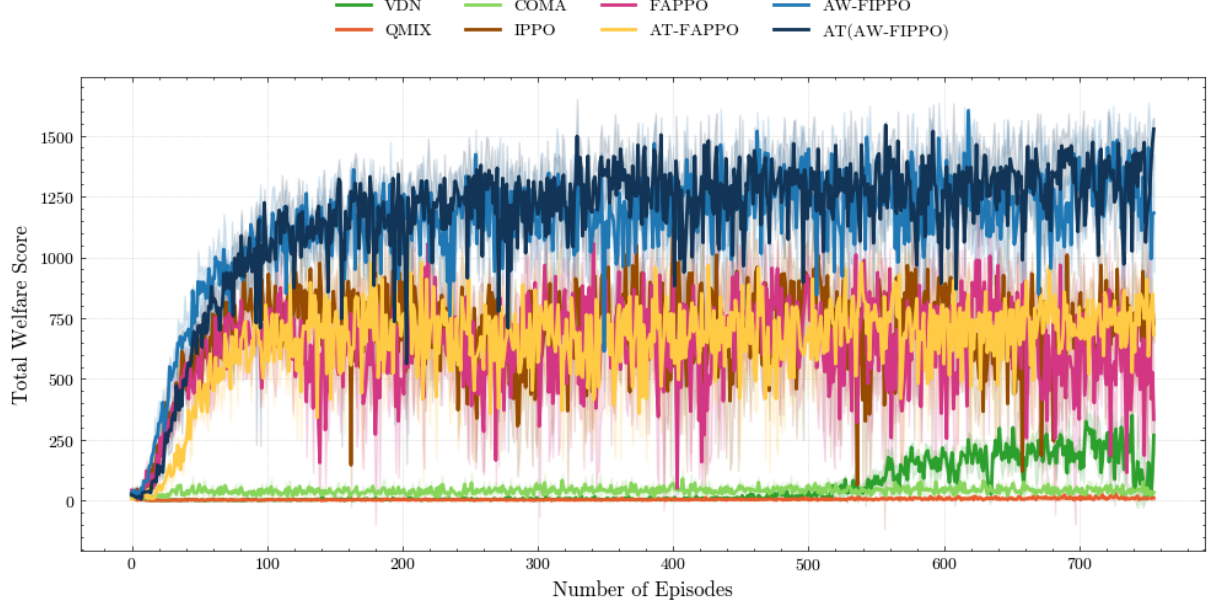


Figure 4.7: Matthew welfare scores.

As illustrated in Figure 4.7, the welfare score learning curve, which is the comprehensive metric that quantifies the balance between efficiency and equity, AW-FIPPO and AT-AWFIPPO, consistently achieve the highest welfare scores over the entire training duration. The baseline FAPPO and AT-FAPPO perform well but are clearly outperformed by our proposed approach.

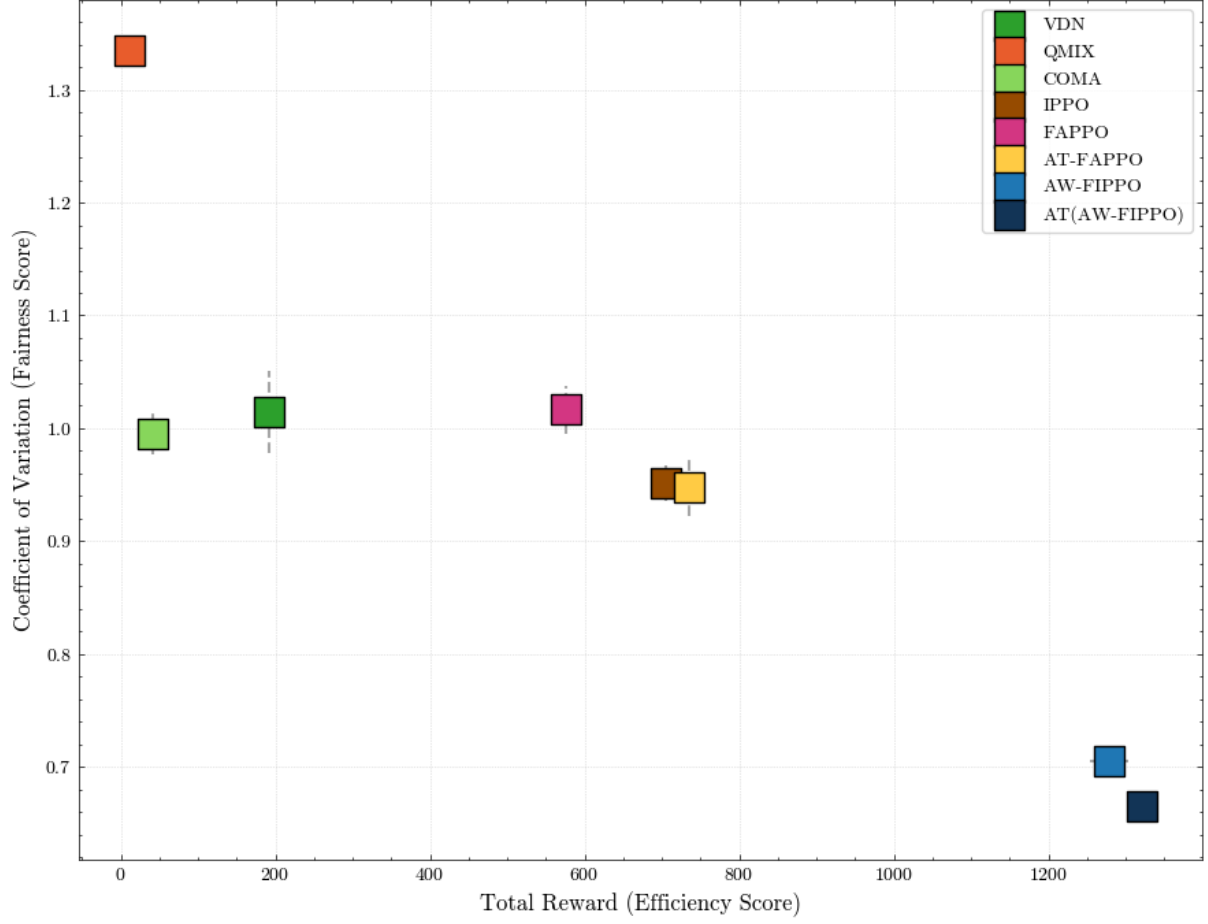


Figure 4.8: Efficiency-equity trade-off(Matthew): Total Reward vs. Gini index

The scatter plot in Figure 4.8, the trade-off between total reward vs. CV, demonstrates the superior Pareto-optimality of our proposed approach compared to established baselines. Our method occupies the desirable bottom-right quadrant, achieving significantly higher total reward (Efficiency) while maintaining a lower CV (Equity). As clearly shown in the plot, the baseline algorithms, for instance, IPPO and AT-FAPPO, exhibit a sharp trade-off where gains in total reward resulted reward disparity among agents. Our approach, on the other hand, effectively breaks this correlation and provides a more stable and equitable coordination strategy across all experimental seeds.

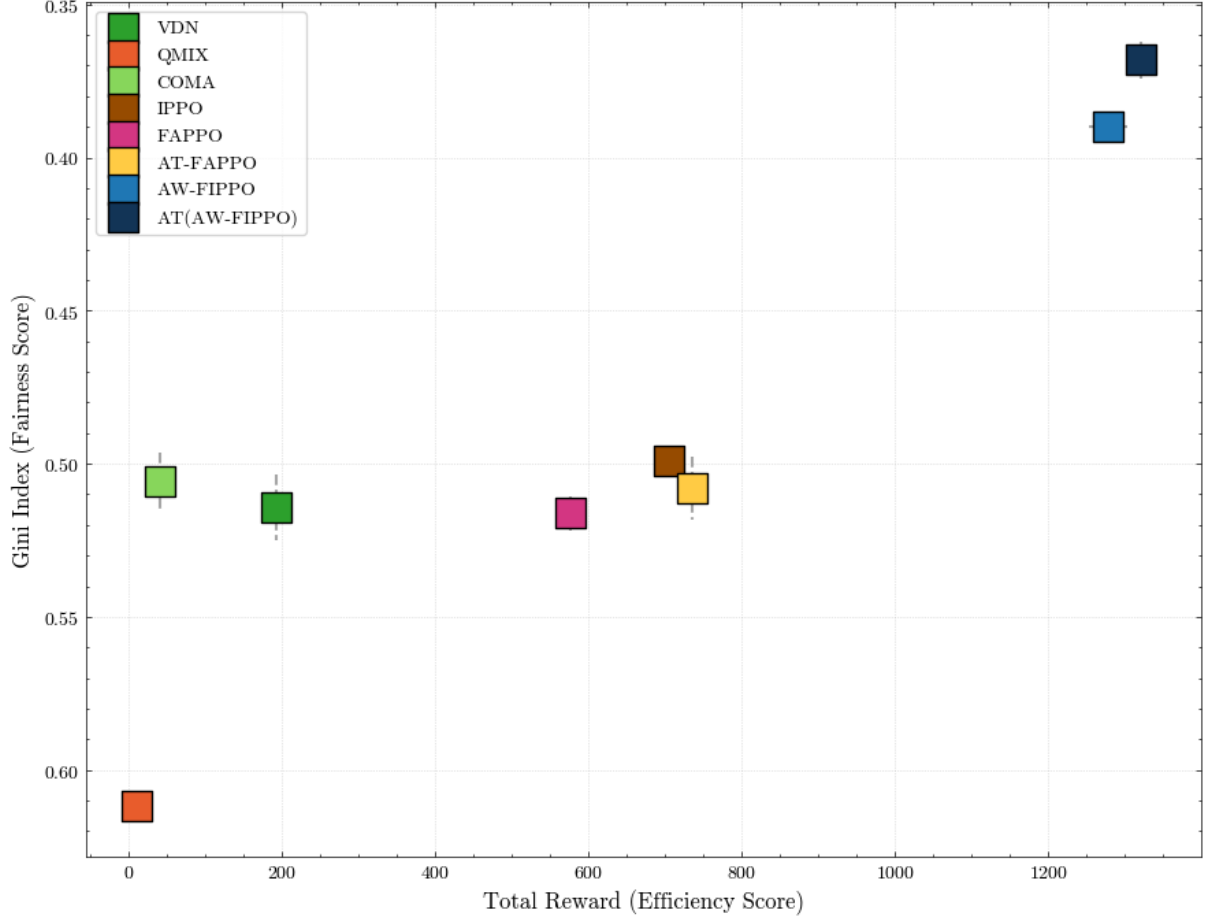


Figure 4.9: Efficiency-equity trade-off(Matthew): Total Reward vs. CV

The visualization in Figure 4.9, which maps total reward against the Gini index, further validates the effectiveness of our approach. The cluster representing our method is positioned as expected at the top-right of the coordinate plane, indicating a simultaneous maximization of collective utility and minimization of individual inequality. Unlike the baselines, which show higher variance and greater disparity (higher Gini index), our approach maintains a tight, highest-performance cluster, proving its robustness in fostering system-wide efficiency, and the lowest fairness cluster, showing its fairness in resource distribution among agents without sacrificing system-wide efficiency.

Overall the fairness-weight adaptiveness of our study helped the policy to navigate through fairness and efficiency, enabling the policy to keep maximizing the cumulative return and be highly sensitive to inequality only when the environment state necessitates intervention. Unlike the baseline FAPPO’s static $w_i = 1/2^i$ weights, which penalize unfairness with the same constraint across the entire training, our approach excels at maintaining the system’s robustness in both fairness and efficiency manner based on the current state of the environment. Our approach’s adaptivity is the main cause for the overall improvement in both stability and performance of the system across different environment types.

4.3 Discussion

Traditional MARL methods either enforce fairness constraints that typically affect the system performance or they purely maximize utility or system efficiency that often causes the Matthew Effect, where reward monopolization by a few individuals undermines long-term cooperation. To address this, we set an objective to develop a framework that can balance the equity-efficiency tradeoff by dynamically navigating between maximizing collective efficiency and fair reward distribution in cooperative MARL. In order to achieve our objective, we introduce the Weight Adaptive FIPPO (AW-FIPPO) algorithm, which incorporates a dedicated Fairness Critic Network ($\mathcal{F}$) onto the baseline FAPPO [7] to produce learned fairness weights.

The experimental result, particularly on the complex Mathew Effect environment 3.5.2, confirms that the learned fairness weights balance the tradeoff and proves that the static, manually tuned weight vectors used in the FAPPO method fundamentally limited the algorithm’s performance and robustness. By incorporating the fairness network to learn the context-aware weight (w_t) based on the current environmental state (s_t) and reward disparity among agents (r_t), AW-FIPPO move beyond rigid, fixed fairness constraints toward an adaptive, data-driven objective. The learned weight vector (w_t) is a reference to the state of reward distribution that determines whether the system is achieving maximum productivity without compromising equitable distribution or not.

The transition from static FAPPO to the weight adaptive AW-FIPPO represents a fundamental paradigm shift from relying on a fixed and static weight vectors to a learned and data-driven fairness weights for optimizing social welfare function. FAPPO’s reliance on fixed exponential weighting ($\omega_i = 1/2^i$) imposes a rigid hierarchy that can limit the system’s ability to achieve higher performance. This effect particularly shown in its performance in Matthew Effect environment as it resulted in a modest total return of 575.72, the system failed to gain highest performance compared to AW-FIPPO.

AW-FIPPO achieves a total return of 1278.61 a great difference which is 122% increase in collective productivity. This performance improvement is a direct result of the framework’s ability to dynamically shift the policy toward a standard PPO maximization objective when the environment is balanced. Unlike FAPPO which constrained the policy by a static fairness weight throughout the entire training, AW-FIPPO detect states of equitable distribution and relaxes its fairness constraints to prioritize pure efficiency (using standard PPO). This allows the agents to maximize their performance without being Pulled to the side by the static penalty when they actually interact fairly. From the fairness aspect, AW-FIPPO reduced the Gini index from 0.5162 to 0.3899. This indicates that by knowing when to operate using the standard PPO and when to prioritize fairness using log transformation, AW-FIPPO effectively break the Matthew Effect without undermining the global optimization.

On the other hand, while AT-FAPPO demonstrates that communication alone can improve coordination and raise efficiency to 735.54, it fails to fundamentally resolve the fairness problem, as the Gini index remains high at 0.5080. This proves that agents can learn to coordinate their movements and actions (tactical coordination) without necessarily learning to distribute rewards fairly (social cooperation). However, by integrating our adaptive weight approach, AT-(AW-FIPPO) achieves a massive improvement in the results: the highest efficiency (1320.50) and the lowest inequality (Gini = 0.3682). This indicates that while the Attention mechanism provides the agents with the eyes to see their peers, the Adaptive Weighting provides the means to ensure those peers succeed.

Based on the evaluation analysis we can conclude that Matthew Effect is not an inevitable consequence of maximizing performance in MARL, rather it is a limitation of inflexible objective functions. This is proven by our adaptive variants performance compared to their static counterparts. AW-FIPPO provides a 24.5% reduction in inequality (in Gini index) over FAPPO, while AT-(AW-FIPPO) provides a 27.5% reduction over AT-FAPPO. The critical insight derived from this analysis is that the equity-efficiency tradeoff can be addressed by conditioning the objective policy on adaptive welfare weights. The Fairness Critic’s ability to detect fairness state ensures that the pursuit of social welfare becomes a catalyst rather than a barrier to the global optimization.

4.4 Ablation Study

To evaluate the individual contributions of the structural enhancements within the architecture, we conducted a systematic ablation study consisting of four distinct experimental configurations. These configurations included the baseline FAPPO, an attention-only version (AT-FAPPO), a version featuring only the Adaptive Weighting mechanism (AW-FIPPO), and finally the complete architecture combining both features AT-(AW-FIPPO). By separately experimenting this configurations, we were able to isolate the influence of AW components from the configuration that do not incorporate AW (adaptive weight) component. This provides us a clear effect of the component in how each mechanism contributes to cooperative stability, performance efficiency, and the overall fairness of reward allocation.

Variant A: FAPPO (The Baseline)

FAPPO serves as the experimental baseline, representing the fundamental architecture upon which subsequent improvements are performed. This variant integrates fairness directly into the objective function via generalized Gini welfare function. In the configuration, the fairness weights incorporated in GGW function are pre-determined and static from the beginning of the training process. This formulation allows us to isolate

and quantify the impact of the dynamic, adaptive fairness weights introduced in more advanced variants.

Variant B: AW-FIPPO (Adaptive Weighting-Only)

In this configuration, we implement the Adaptive Weighting module that incorporate a layer of adaptive approach into the baseline FAPPO. Unlike the static approach where weights are fixed at initialization, AW-FIPPO employs an adaptive weight generated by a dedicated additional network in order to allow the policy to make an informed adjustments to either maximize performance or ensure fairness among agents when necessary. Using the adaptive weight, the model respond to the changing environment state and reward disparity, shifting its focus between maximizing total reward and minimizing disparity as the agents.

Variant C: AT-FAPPO (Attention-Only)

This variant introduces a multi-head self-attention mechanism to the baseline to facilitate inter-agent communication but omits the Adaptive Weighting (AW) module. By packing all agents' queries (Q), keys (K), and values (V) into matrices, AT-FAPPO allows each agent to compute attention scores through a compatibility function—the inner product of its query vector with the key vectors of itself and its neighbors. The attention scores are then scaled and normalized via softmax to produce attention weights (α'), which are then used to aggregate the corresponding value vectors into an informed output vector (b). The results allow agents to quantify relationships between their observations. Higher weights are assigned when agents are in related environment or their actions significantly influence one another.

Variant D: AT(AW-FIPPO) (The Complete Architecture Combining both Features)

The complete architecture integrates both the attention-based mechanism of Variant C and the adaptive fairness weighting approach of Variant B. From the attention-based mechanism, the model utilizes multi-head attention to extract high-fidelity relational features between agents observations. In order to dynamically modulate the trade-off between performance and equity, the model utilizes the adaptive weighting mechanism of variant B. These combined features of the two variants ensure that the agents not only possess the intelligence to understand complex environmental interactions but also the policy flexibility to prioritize fairness when reward disparity among agents is detected.

Table 4.3: The absence and presence of architectural components in each variants

Variant	Static Weight	Attention (AT)	Adaptive Weighting (AW)
FAPPO	✓	✗	✗
AT-FAPPO	✓	✓	✗
AW-FIPPO	✗	✗	✓
AT(AW-FIPPO)	✗	✓	✓

Table 4.4: The Performance comparison of each variant

Variant	Total	Max	Gini	CV
FAPPO	575.72	156.27	0.5162	1.0163
AT-FAPPO	735.54	206.65	0.5080	0.9478
AW-FIPPO	1278.61	288.06	0.3899	0.7048
AT(AW-FIPPO)	1320.50	289.45	0.3682	0.6651

The ablation study results, summarized in Table 4.4, evaluate the impact of Adaptive Weighting (AW) mechanisms on the baseline FAPPO and AT-FAPPO architectures. The empirical result suggests a clear improvement both in terms of efficiency and fairness. The baseline FAPPO exhibits the lowest mean performance (575.72) and the highest CV (1.0163), indicating that static feature weighting limits the agent’s capacity to generalize in dynamic environments. The attention mechanism in AT-FAPPO yields a moderate improvement compared to FAPPO yielding mean score to 735.54, suggesting that spatial feature prioritization facilitates better decision-making.

The performance improvements are well observed with the AW variant such as AW-FIPPO and AT(AW-FIPPO). AW-FIPPO achieved a massive improvement compared to both FAPPO and AT-FAPPO by scoring mean (1278.61) and Gini index value (0.3899). The combined AT(AW-FIPPO) even enhanced the results by a mean of 1320.50 compared to all variants, followed by the lowest Gini index value (0.3682) than each of them. This indicates that the Adaptive Weighting mechanism is the core point for maximizing the system performance alongside maintaining fairness. From the results suggest can conclude that the AW component are the primary driver for improved efficiency and a balanced reward distribution in both AW-FIPPO and AT(AW-FIPPO) variants.

4.5 Research Findings and Detailed Answers

In this section, we provide a detailed analysis of our findings as they relate to the core research questions established in Section 1.4.

RQ 1: How can an auxiliary fairness critic neural network be designed to observe global state transitions and reward distributions to produce real-time weight vectors?

We designed the auxiliary fairness critic network as a sequential MLP featuring a 64-unit hidden layer with tanh activation to capture non-linear correlations of state and reward. The purpose of the network is to map episodic agents' rewards and initial environment state to the learned fairness weights. In order to produce these fairness weights, the network processes a concatenated feature vector that combines the environment's initial state with the episodic performance outcomes as an input. Specifically, it extracts the state at the start of the episode ($batch["state"][:, 0, :]$) and concatenates it with the flattened, summed rewards of each agent across each time step of the episode.

The output layer of the network utilizes a sigmoid activation to constrain the learned fairness weights within the range $(0, 1)$. These constrained fairness weights are then used to blend the standard PPO objective with a concave log transformation. To avoid numerical instability for agents with zero advantages, PPO's surrogate objective ($\min_surr$) is clamped to a small positive value (10^{-6}) before the log transform. The final loss term is calculated as a weighted sum: $(1 - \omega_t) \min(surr_1, \bar{surr}_2) + \omega_t \log(\min(surr_1, \bar{surr}_2) + \epsilon_{log})$. In order for a fairness critic to have an accurate prediction of the learned fairness weights, we train it with a Gini-targeted objective. Therefore, the goal is to minimize the Mean Squared Error (MSE) between its predicted fairness weights and the actual Gini coefficient result, which measures the actual reward dispersion observed during agent trajectories.

The implementation of our design begins with preparing the inputs for the fairness critic, which involves extracting the initial states and the rewards obtained by each agent from the episode batch. These values are concatenated into a single input tensor, which the fairness critic uses to produce the fairness weights. The core of the update logic uses these fairness weight to modulate the actor loss. The standard surrogate loss is balanced against the logarithmic utility transformation using the fairness critic's weight as the interpolation factor. When the gradient is computed, the term $((1 - \omega_t) \min(surr_1, \bar{surr}_2))$ ensures that the system still pursues high cumulative rewards, maintaining competitive performance. The $(\omega_t \log(\min(surr_1, \bar{surr}_2) + \epsilon_{log}))$ term simultaneously applies a corrective force that push the policy towards fairness. When Fairness Critic detects a thresholds where collective gain starts to cause unfairness, the policy dynamically shift the update towards log transformation to keep the agents within a fairness corridor. Finally, this policy loss is aggregated with the fairness critic loss, standard critic loss and an entropy-regularized exploration term to compute the total batch loss.

RQ 2: How effectively does the designed AW-FIPPO and AT(AW-FIPPO) artifact

perform against existing baselines in terms of utility and equity?

The effectiveness of AW-FIPPO artifact is demonstrated through its performance in resource-scarce environments where reward accumulation bias is inherent. The evaluation compares our Adaptive Weighting FIPPO mechanism against static-constraint baselines (FAPPO) and purely utilitarian methods (COMA, VDN, QMIX, and IPPO). By replacing the static fairness weights with environment-aware preference shifts, our method bridge the gap between system efficiency and agent equity. It adjust its focus depending on the severity of resource imbalance detected within the environment. This approach allows the agent to navigate the Pareto frontier more effectively than static-constraint baseline, which often suffer from a fairness tax where productivity is sacrificed due to the static nature of fairness weights. The evaluation was conducted across two distinct environments to test the effectiveness of our adaptive approach. As shown in the empirical results in Section 4.2, our approach significantly outperformed the baselines, particularly in the Matthew Effect environment, where wealth concentration typically accelerates without intervention.

Efficiency Measure Analysis

In terms of overall system productivity, the our methods demonstrates that incorporating adaptive fairness weighting approach enhance the system performance and maintained a fair reward distribution among agents. Our empirical evaluation in the high-volatility Matthew Effect environment reveals that our AW-FIPPO method significantly outperforms the baseline FAPPO by achieving a 122.1% gain in total score and an 84.3% gain in the maximum score attained by any single agent. Compared to AT-FAPPO, our approach AT(AW-FIPPO) achieves a 79.5% gain in total score and a 40.1% gain in max score. This improvement in both weight adaptive variants indicates that the adaptive mechanism allows high-performing agents to continue their trajectories as long as systemic equity is maintained. Unlike baseline models that suffer from performance collapse due to inflexible constraints, AW-FIPPO and AT-(AW-FIPPO) maximize aggregate throughput. This proves that a balanced reward distribution can coexist with—and even stabilize—superior collective utility, effectively resolving the tradeoff between fairness and efficiency in MARL.

Equity Measure Analysis

The empirical effectiveness of our approach is most notably demonstrated by its capacity to obtain a significantly more equitable reward distribution across the agent population. When compared to the static weighting of FAPPO, the AW-FIPPO framework achieved a 24.5% reduction in the Gini Index and a 30.6% reduction in the Coefficient of Variation (CV).

Furthermore, the superiority of the adaptive mechanism is maintained when integrated with attention-based architectures. Our AT-(AW-FIPPO) variant outperformed AT-FAPPO, yielding a 27.5% reduction in the Gini Index and a 29.8% decrease in CV.

These results provide strong empirical evidence that transitioning from static to adaptive fairness integration enables the system to converge toward a more egalitarian equilibrium. Crucially, these gains in distributive equity do not cause the fairness tax which is typically observed in multi-agent systems; rather, the architecture maintains—and in some cases enhances—overall system efficiency. This confirms that a learned Fairness Critic can successfully navigate the tradeoff between individual agent performance and collective social welfare.

Chapter 5

Conclusion and future work

5.1 Conclusion

This study introduced fairness adaptive learning framework designed to address the trade-off between collective system efficiency and individual equity in cooperative MARL. By taking static fairness constraints in existing approach as a limitation, we formulated research questions centered on how we can design a mechanism that can learn fairness constraints dynamically in order to respond to the changing reward disparities. We used Design Science Research Process (DSRP) as our guide to convert this question into a design objective and developed AW-FIPPO architecture as a solution artifact. The core contribution of this work lies in the integration of an auxiliary fairness critic network for generating context-aware fairness weights ω_t . We used those fairness weights in the policy optimization objective for transitioning between utility-driven and equity-oriented update.

Empirical evaluation across two fairness-sensitive environments demonstrates that the proposed approach achieved a more balanced distribution of rewards compared to the baseline FAPPO and other standard methods, including QMIX, VDN, COMA, and IPPO methods while maintaining improved collective returns. The performance improvement of our approach is much more reflected in Matthew Effect environment, which reflect higher resource volatility, the generalizability of the adaptive weighting mechanism was even more evident.

AW-FIPPO achieves a 122.1% increase in Total Reward compared to FAPPO and also a 24.5 % reduction in Gini Index followed by a 30.6 % reduction in reward dispersion (CV). On the other hand, our approach AT(AW-FIPPO) achieves a 79.5% gain in total score while reducing the Gini index by 27.5% and CV by 29.8%. This indicates that our approach sustained a balanced agent participation.

These results of the proposed study conclusively demonstrate significant and quantifiable improvement in cumulative return and social welfare scores. The resulting reward

distributions exhibit the lowest CV, confirming that the adaptive weights successfully minimized reward disparity. By using the environment state and agents returns as input, the Fairness Critic made it possible for the policy to be highly sensitive to unfairness while maintaining aggressive pursuit of high returns when equity is secured. This validated the hypothesis that a fixed, manually tuned fairness constraint in FAPPO, which is static welfare weights is fundamentally limited in handling the diverse and dynamic challenges of MARL environments. The stability and superior performance of both AW-FIPPO and AT-AWFIPPO across all metrics mark a crucial step toward creating efficient and ethically-aware multi-agent system. Therefore, the experimental findings definitively validate the core objectives of this research.

5.2 Future Work

The experimental findings presented in this study demonstrate that utilizing state-dependent weight vectors in our adaptive approach is a driving factor for balancing fairness and efficiency tradeoffs. Building upon these results, we identify the following future directions to further strengthen the framework’s effectiveness, scalability, and generalization across diverse multi-agent scenarios.

- **Mid-Episode Feedback Integration for Temporal Adaptivity**

Instead of relying on the episodic returns, utilizing instantaneous or per-step returns as an input for the fairness critic network will enable the policy objective to respond quickly to unfairness that arises in the middle of the episode. Employing per-step returns as an input for the fairness critic network will further improve the system’s temporal adaptability in the face of emerging mid-episode unfairness. This improvement is particularly effective in highly dynamic environments where reward monopolization occurs in a fast rise.

- **Incorporating Multi-Dimensional Socio-Technical Features**

The current version of the framework utilizes the simple reward return and environment state as an input for fairness critic network. Incorporating richer dimensions of agent interaction, such as historical resource consumption patterns, agent effort, and opportunity costs associated with decisions made in constrained environments, will lead the Fairness Critics Network to assess fairness in a more holistic manner. The more accurately the Fairness Critics Network assesses unfairness and has a more nuanced and context-sensitive interpretation of fairness within the system, the better the framework adjusts the policy objective towards a fairer reward distribution within the system. This will lead the framework to align more closely

with principled and philosophically grounded notions like equity and need-based reward allocation.

References

- [1] M. A. Wiering *et al.*, “Multi-agent reinforcement learning for traffic light control,” in *Machine Learning: Proceedings of the Seventeenth International Conference (ICML’2000)*, 2000, pp. 1151–1158.
- [2] Y. Cao, W. Yu, W. Ren, and G. Chen, “An overview of recent progress in the study of distributed multi-agent coordination,” *IEEE Transactions on Industrial informatics*, vol. 9, no. 1, pp. 427–438, 2012.
- [3] U. Siddique, A. Sinha, and Y. Cao, “On deep reinforcement learning for target capture autonomous guidance,” in *AIAA SCITECH 2024 Forum*, 2024, p. 0957.
- [4] Z. Yao, Z. Ding, and T. Clausen, “Multi-agent reinforcement learning for network load balancing in data center,” in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3594–3603.
- [5] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, “Resource management in wireless networks via multi-agent deep reinforcement learning,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3507–3523, 2021.
- [6] F. A. Oliehoek, C. Amato *et al.*, *A concise introduction to decentralized POMDPs*. Springer, 2016, vol. 1.
- [7] U. Siddique, P. Li, and Y. Cao, “Towards fair and efficient policy learning in cooperative multi-agent reinforcement learning,” in *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, 2025, pp. 2744–2746.
- [8] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, “Value-decomposition networks for cooperative multi-agent learning based on team reward,” in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2085–2087.
- [9] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, “Qmix: Monotonic value function factorisation for deep multi-agent reinforcement

- learning,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.
- [10] C. S. De Witt, T. Gupta, D. Makoviichuk, V. Makoviyuchuk, P. H. Torr, M. Sun, and S. Whiteson, “Is independent learning all you need in the starcraft multi-agent challenge?” *arXiv preprint arXiv:2011.09533*, 2020.
 - [11] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
 - [12] s. d. Jong, k. tuyls, and k. verbeeck, “Fairness in multi-agent systems,” *The Knowledge Engineering Review*, vol. 23, no. 2, p. 153–180, 2008.
 - [13] J. Jiang and Z. Lu, “Learning fairness in multi-agent systems,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 13 854–13 865.
 - [14] P. Ju, A. Ghosh, and N. B. Shroff, “Achieving fairness in multi-agent markov decision processes using reinforcement learning,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.00324>
 - [15] G. La Malfa, J. M. Zhang, M. Luck, and E. Black, “Using protected attributes to consider fairness in multi-agent systems,” *CoRR*, 2024.
 - [16] —, “Fairness aware reinforcement learning via proximal policy optimization,” *CoRR*, 2025.
 - [17] M. Zimmer, C. Glanois, U. Siddique, and P. Weng, “Learning fair policies in decentralized cooperative multi-agent reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 967–12 978.
 - [18] N. A. Grupen, B. Selman, and D. D. Lee, “Cooperative multi-agent fairness and equivariant policies,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 9350–9359.
 - [19] Y. Du, J. Z. Leibo, U. Islam, R. Willis, and P. Sunehag, “A review of cooperation in multi-agent learning,” *CoRR*, 2023.
 - [20] C. d’Aspremont and L. Gevers, “Social welfare functionals and interpersonal comparability,” *Handbook of social choice and welfare*, vol. 1, pp. 459–541, 2002.
 - [21] L. Busoniu, R. Babuska, and B. De Schutter, *Comprehensive Survey of Multiagent Reinforcement Learning*. IEEE, 2008.

- [22] C. Amato, “An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning,” *arXiv preprint arXiv:2409.03052*, 2024.
- [23] K. Zhang, Z. Yang, and T. Başar, “Multi-agent reinforcement learning: A selective overview of theories and algorithms,” *Handbook of Reinforcement Learning and Control*, 2021.
- [24] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, “A survey of learning in multiagent environments: Dealing with non-stationarity.” *CoRR*, 2017.
- [25] M. L. Littman, “Markov games as a framework for multi-agent reinforcement learning,” in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.
- [26] M. Tan, “Multi-agent reinforcement learning: Independent vs. cooperative agents,” in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 330–337.
- [27] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.
- [28] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, “Fairness in reinforcement learning,” in *International conference on machine learning*. PMLR, 2017, pp. 1617–1626.
- [29] A. Reuel and D. Ma, “Fairness in reinforcement learning: A survey,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 1218–1230.
- [30] U. Siddique, P. Weng, and M. Zimmer, “Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 8905–8915.
- [31] H. Moulin, *Fair division and collective welfare*. MIT press, 2004.
- [32] M. Omid, “Fair and efficient resource scheduling in heterogeneous multi-agent systems,” Master’s thesis, University of Waterloo, 2024.
- [33] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser *et al.*, “Starcraft ii: A new challenge for reinforcement learning,” *arXiv preprint arXiv:1708.04782*, 2017.
- [34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

- [35] R. S. Sutton and A. G. Barto, “Reinforcement learning: an introduction 1998 cambridge, ma,” 2018.
- [36] M. Vlasceanu and D. M. Amodio, “Propagation of societal gender inequality by internet search algorithms,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 29, p. e2204529119, 2022.
- [37] P. Hacker, “Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law,” *Common market law review*, vol. 55, no. 4, 2018.
- [38] J. Zhu, M. Hu, A. Zhang, R. Jin, and R. Liu, “Fairness-sensitive policy-gradient reinforcement learning for reducing bias in robotic assistance,” in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 549–554.
- [39] M. Agarwal and V. Aggarwal, “Reinforcement learning for joint optimization of multiple rewards,” *Journal of Machine Learning Research*, vol. 24, no. 49, pp. 1–41, 2023.
- [40] J. Chen, Y. Wang, and T. Lan, “Bringing fairness to actor-critic reinforcement learning for network utility optimization,” in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.
- [41] Y. Ge, X. Zhao, L. Yu, S. Paul, D. Hu, C.-C. Hsieh, and Y. Zhang, “Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning,” in *Proceedings of the fifteenth ACM international conference on web search and data mining*, 2022, pp. 316–324.
- [42] K. Gogineni and P. Wei, “Scalability bottlenecks in multi-agent reinforcement learning systems,” in *FastPath 2023: International Workshop on Performance Analysis of Machine Learning Systems*, 2023.
- [43] K. Pfeffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, V. Virtanen, and J. Bragge, “The design science research process: A model for producing and presenting information systems research,” in *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006)*, Claremont, CA, USA, 2006, pp. 83–106.
- [44] W. Ogryczak, H. Luss, M. Pióro, D. Nace, and A. Tomaszewski, “Fair optimization and networks: A survey,” *Journal of Applied Mathematics*, vol. 2014, no. 1, p. 612018, 2014.

- [45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [46] J. Hu, S. Jiang, S. A. Harding, H. Wu, and S.-w. Liao, “Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning,” *arXiv preprint arXiv:2102.03479*, 2021.