Beyond Mandatory Federations: Balancing Egoism, Utilitarianism and Egalitarianism in Mixed-Motive Games

Shaokang Dong[1, 2], **Chao Li**[3], **Shangdong Yang**[3], **Hongye Cao**[2], **Wanqi Yang**[1*], **Yang Gao**[2]

[1] School of Computer and Electronic Information, Nanjing Normal University, Nanjing, China
[2] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[3] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
shaokangdong@gmail.com, yangwq@njnu.edu.cn, gaoy@nju.edu.cn.

Abstract

In the field of mixed-motive games, extensive multi-agent learning studies have explored the balance between egoism (individual interest), utilitarianism (collective interest), and egalitarianism (fairness). Traditional approaches often rely on manually designed reward functions, social norms, and alliance/federation mechanisms to transition agents from individualistic behaviors toward cooperative strategies. However, these methods typically require all agents to share private local information or to mandatorily participate in federations, which is impractical in real-world applications. To address these issues, this paper proposes a Flexible-Participation Federation (FPF) framework that allows agents to participate in the federation voluntarily. Furthermore, we extend the federation from a global to a Local Multi-Federation (LMF) framework, enabling agents to form multiple localized federations, thereby promoting more efficient and adaptive cooperation. Theoretical evidence demonstrates that the global FPF model, along with the discrepancy between decentralized egoistic policies and federated utilitarian policies, achieves an O(1/T) convergence rate. Agents in the LMF framework also reach consensus within a sublinear gap. Extensive experiments show that agents opting out of federation participation experience a reduction in egoism, and our approach outperforms multiple baselines in terms of both utilitarianism and egalitarianism.

Code — https://github.com/Shaokang-Agent/FPF-LMF.

Introduction

Egoism (individual interest), utilitarianism (collective interest) and egalitarianism (fairness) are three critical considerations in the field of mixed-motive games, including common resource allocation (Ostrom 2008; Perolat et al. 2017), task scheduling (Arunarani, Manjula, and Sugumaran 2019), and autonomous driving (Zhou et al. 2021; Vinitsky et al. 2022). For instance, in the coordination dilemma of autonomous driving at a crossroad, each agent seeks to reach its destination as quickly as possible. However, agents exclusively prioritizing egoism may cause severe traffic congestion and accidents, adversely affecting long-term individual returns. Simultaneously, the system must consider overall road travel time, potentially requiring some agents to prioritize others, which may compromise egalitarianism. A potential resolution to this paradox involves balancing egoism, utilitarianism and egalitarianism in mixed-motive games (Dong et al. 2024a).

Extensive Multi-Agent Reinforcement Learning (MARL) methods have been developed to balance these criteria, guiding agents from individualistic behaviors toward cooperation. Some approaches address inequity aversion (Hughes et al. 2018), causal influence (Jaques et al. 2019), or fairer allocation (Jiang and Lu 2019; Zimmer et al. 2021) among agents by manually designing reward functions. Other works adopt a systemic perspective, promoting social norms (Vinitsky et al. 2023; Köster et al. 2022; Anastassacos et al. 2021) and alliance/federation mechanisms (Baker 2020; Anastassacos, Hailes, and Musolesi 2020; Dong et al. 2024b).

However, these methods frequently overlook a critical aspect: *It is impractical to require all agents to share private local information or participate in alliances/federations*. One issue is that certain agents in the system may experience downtime at any given step. Additionally, sharing private information and mandating participation in a centralized federation disregards agent autonomy, potentially destabilizing cooperation and increasing the risk of free-riding (Hughes et al. 2018; Köster, et al. 2020) or exploitation by other agents (Chelarescu 2021).

In this paper, we first propose a Flexible-Participation Federation (FPF) framework where agent engagement in the global federation is voluntary. The federation integrates the egoistic policy parameters of participating agents and broadcasts the utilitarian and egalitarian policy parameters to all agents. Then the participating agents regularize their egoistic policy parameters, fostering a balance between egoism, utilitarianism and egalitarianism. Moreover, we extend the global federation to a Local Multi-Federation (LMF) framework, enabling agents to efficiently exchange local policy parameters or policy gradients within self-organized groups while retaining the option to operate independently. In summary, the contributions of this work are three-fold:

- We propose FPF and LMF frameworks as alternatives to mandatory federation to balance egoism, utilitarianism and egalitarianism in mixed-motive games.
- Theoretical analysis shows that the FPF model, as well as the discrepancy between decentralized egoistic policies and federated utilitarian policies, achieves an $\mathcal{O}(1/T)$

*Corresponding author: Wanqi Yang.

convergence rate. Furthermore, agents in the LMF framework can reach consensus with a gap of $\mathcal{O}(1/T^{\beta} + \bar{\epsilon}^{T})$.

- Extensive experiments conducted across various environments demonstrate that FPF and LMF outperform multiple baselines in balancing utilitarianism and egalitarianism. Additionally, agents who opt out of federation participation experience a decrease in individual interest.

Related Work

Reward Shaping. This type of approach promotes the emergence of prosocial cooperative behaviors or utilitarianism in mixed-motive games by modifying individual agent rewards. For example, Inequality Aversion (IA) (Hughes et al. 2018) and Individual Fairness Concern Utility Function (IFCUF) (Chen et al. 2023) address advantageous and disadvantageous inequities to achieve a more rational distribution of social welfare. Social Influence (Jaques et al. 2019) incentivizes actions that affect other agents' policies. The Learning to Incentivize Others (LIO) (Yang et al. 2020) provides additional internal rewards to other agents, thereby encouraging cooperative behavior. Social Value Orientation (SVO) (McKee et al. 2020) fosters cooperation by leveraging the discrepancy between observed reward tendencies and target SVO as internal incentives. Additionally, BAROCCO (Ivanov, Egorov, and Shpilman 2021) demonstrates that egalitarianism can emerge from egoism by incorporating both selfish and social motives through the long-term value function rather than immediate rewards. Finally, works like FEN (Jiang and Lu 2019) and SOTO (Zimmer et al. 2021), inspired by the fair social welfare function (Speicher et al. 2018; Heidari et al. 2018), introduce efficient and fair reward functions to achieve balanced policies in mixed-motive games.

Social Norms. Effective social norms can facilitate coordination in the absence of societal laws to constrain agent behaviors, significantly enhancing the performance of both individual agents and agent societies (Sen and Airiau 2007). Social norms can be underpinned through sanctioning, as exemplified by the Classifier Norm Model (CNM) (Vinitsky et al. 2023), which introduces a classifier to predict whether a given behavior will be approved or sanctioned by the system. Another work (Anastassacos et al. 2021) introduces mechanisms for collectively establishing social norms and assigning reputations, resulting in improved policy equilibria. Additionally, (Köster et al. 2022) explores how even spurious or trivial norms can positively impact the learnability of compliant behaviors, thereby enhancing overall performance in terms of utilitarianism.

Alliance/Federation. Several works have investigated the mechanisms for forming alliances or federations. For instance, the work (Anastassacos, Hailes, and Musolesi 2020) examines partner selection mechanisms that promote cooperation among agents with selfish objectives, ultimately leading to a Tit-for-Tat strategy. Randomized Uncertain Social Preferences (RUSP) (Baker 2020) demonstrates that training reinforcement learning agents with a randomized reward transformation matrix can foster both reciprocity and team formation. Another study (Dong et al. 2024b) defines egoism, utilitarianism, and egalitarianism criteria in mixed-motive games and proposes a mandatory participation Decentralized and Federated (D&F) framework to balance these criteria.

However, these methods often require access to other agents' private information to design reward-shaping mechanisms and social norms or compel agents to participate in alliances or federations, raising the potential for exploitation by other agents in mixed-motive games.

Preliminary

Problem Formulation

The mixed-motive games can be formalized as the following tuple $\langle \mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{O}}, \boldsymbol{\mathcal{A}}, \mathcal{T}, \boldsymbol{\mathcal{R}}, \gamma \rangle$, where

- $\mathcal{N}$ is the finite set of N agents.
- $\mathcal{S}$ represents the finite state space in the environment.
- $\boldsymbol{\mathcal{O}} = \times_{i \in \mathcal{N}} \mathcal{O}_i$ denotes a joint observation space, where $\mathcal{O}_i$ is the finite observation space of agent i.
- $\boldsymbol{\mathcal{A}} = \times_{i \in \mathcal{N}} \mathcal{A}_i$ denotes a joint action space, where $\mathcal{A}_i$ is the finite action space of agent i.
- $\mathcal{T} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0, 1]$ is the Markovian state transition probability, where $\mathcal{T}(s'|s, \boldsymbol{a})$ denotes the probability that taking joint action $\boldsymbol{a}$ in state s results in a transition to s'.
- $\boldsymbol{\mathcal{R}} = \times_{i \in \mathcal{N}} \mathcal{R}_i$, where $\mathcal{R}_i : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ is a reward function for agent i.
- $\gamma \in (0, 1)$ is a discount factor.

The **egoism**, **utilitarianism**, and **egalitarianism** criteria are defined as follows (Dong et al. 2024b):

$$\mathcal{J}_i^{\mathrm{ego}}(\theta_i) = \mathbb{E}_{\pi_i(\cdot;\theta_i), \pi_{-i}(\cdot;\theta_{-i})} \left[\sum_{t=0}^{\infty} \gamma^t r_i^t(s, a_i, a_{-i}) \right], \tag{1}$$

$$\mathcal{J}^{\mathrm{uti}}(\omega) = \mathbb{E}_{\boldsymbol{\pi}(\cdot;\omega)} \left[\sum_{t=0}^{\infty} \gamma^t r^{\mathrm{uti,t}}(s, \boldsymbol{a}) \right], \tag{2}$$

$$\mathcal{J}^{\mathrm{ega}}(\bar{\omega}) = \mathbb{E}_{\bar{\boldsymbol{\pi}}(\cdot;\bar{\omega})} \left[\sum_{t=0}^{\infty} \frac{1}{N} \sum_{i=1}^{N} -\gamma^t \left[r_i^t(s, \boldsymbol{a}) - r^{\mathrm{uti,t}}(s, \boldsymbol{a}) \right]^2 \right], \tag{3}$$

where $-i$ denotes all other agents except i and $r^{\mathrm{uti,t}}(s, \boldsymbol{a}) = \frac{1}{N} \sum_{i=1}^{N} r_i^t(s, \boldsymbol{a})$ is the average reward at the time step t. $\pi_i(a_i|o_i; \theta_i)$, $\boldsymbol{\pi}(\boldsymbol{a}|s; \omega)$ and $\bar{\boldsymbol{\pi}}(\boldsymbol{a}|s; \bar{\omega})$ with parameters θ_i, ω and $\bar{\omega}$ are the egoistic policy, the federated utilitarian policy, and the federated egalitarian policy, respectively.

Decentralized and Federated (D&F) Framework

In the D&F framework (Dong et al. 2024b), the overall objective $\mathcal{J}_i$ of each agent i is to find an efficient policy to balance egoism, utilitarianism, and egalitarianism criteria as

$$\mathcal{J}_i = \mathcal{J}_i^{\mathrm{ego}}(\theta_i) + \lambda_u \mathcal{J}^{\mathrm{uti}}(\omega) + \lambda_e \mathcal{J}^{\mathrm{ega}}(\bar{\omega}), \tag{4}$$

where coefficients $\lambda_u, \lambda_e \in [0, +\infty)$ control the balance.

The individual policy $\pi_i(\cdot; \theta_i)$ is optimized in a Decentralized manner ($\mathrm{D}_i^{\mathrm{ego}}$) with only local information, while the utilitarian policy $\boldsymbol{\pi}(\cdot; \omega)$ and the egalitarian policy $\bar{\boldsymbol{\pi}}(\cdot; \bar{\omega})$ are attained in the Federation ($\mathrm{F}^{\mathrm{uti}}$ and $\mathrm{F}^{\mathrm{ega}}$) with global

information. The optimization objective of each agent i can be represented as follows:

$$\mathrm{D}_i^{\mathrm{ego}}: \max_{\theta_i \in \mathbb{R}^d} F_i(\theta_i) := \mathcal{J}_i^{\mathrm{ego}}(\theta_i) - \frac{\lambda_u}{2}\|\theta_i - \omega\|^2 - \frac{\lambda_e}{2}\|\theta_i - \bar{\omega}\|^2, \tag{5}$$

where the constrained coefficients $\lambda_u, \lambda_e \in [0, +\infty)$ carry the same significance as in Equation (4). In the federation, the optimization objective of the utilitarian policy $\boldsymbol{\pi}(\cdot;\omega)$ and the egalitarian policy $\bar{\boldsymbol{\pi}}(\cdot;\bar{\omega})$ can be represented as follows:

$$\mathrm{F}^{\mathrm{uti}}: \max_{\omega \in \mathbb{R}^d} F(\omega) := \mathcal{J}^{\mathrm{uti}}(\omega) - \frac{\lambda_u}{2N}\sum_{i=1}^{N}\|\omega - \theta_i\|^2 - \frac{\lambda_e}{2}\|\omega - \bar{\omega}\|^2, \tag{6}$$

$$\mathrm{F}^{\mathrm{ega}}: \max_{\bar{\omega} \in \mathbb{R}^d} \bar{F}(\bar{\omega}) := \mathcal{J}^{\mathrm{ega}}(\bar{\omega}) + \lambda_h \mathcal{H}(\boldsymbol{\pi}(\cdot;\bar{\omega})), \tag{7}$$

where $\mathcal{H}(\cdot)$ is the entropy and $\lambda_h \in [0, +\infty)$.

Flexible-Participation Federation

To address the constraints imposed by mandatory participation in the D&F framework, this section proposes a Flexible-Participation Federation (FPF) framework, which allows agents to voluntarily participate in the global federation. Initially, the system identifies the set of agents opting to participate in the federation, denoted by $\mathcal{N}_p$ with a size of N_p. The global federation then broadcasts the utilitarian policy parameter ω and the egalitarian policy parameter $\bar{\omega}$ to all agents. Participating agents regularize their individual egoistic policy parameters θ and upload them to the federation. The federation subsequently updates the utilitarian and egalitarian policy parameters, iterating through the training process.

The core training procedure of FPF is specified in Algorithm 1. Specifically, for each agent $i \in \mathcal{N}$, we sample the data from mini-batch $\tilde{\mathcal{D}}_i \subseteq \mathcal{D}_i$ to optimize the policy or value function. The mini-batch provides an unbiased estimation, denoted as follows:

$$\nabla \mathcal{J}_i^{\mathrm{ego}}(\theta_i; \tilde{\mathcal{D}}_i) := \frac{1}{|\tilde{\mathcal{D}}_i|}\sum\nolimits_{\zeta_i \in \tilde{\mathcal{D}}_i} \nabla \mathcal{J}_i^{\mathrm{ego}}(\theta_i; \zeta_i). \tag{8}$$

The approximated parameter $\tilde{\theta}_i$ is then updated till $\|\nabla \tilde{F}_i(\tilde{\theta}_i, \tilde{\mathcal{D}}_i)\|^2 \leq \nu$, where the gradient is given by:

$$\begin{aligned}\nabla \tilde{F}_i(\tilde{\theta}_i, \tilde{\mathcal{D}}_i) = \mathbb{E}_{\tilde{\mathcal{D}}_i}[\nabla_{\tilde{\theta}_i} \mathcal{J}_i^{\mathrm{ego}}(\tilde{\theta}_i)] - \alpha\lambda_u(\tilde{\theta}_i - \omega_i) \\ - \alpha\lambda_e(\tilde{\theta}_i - \bar{\omega}_i).\end{aligned} \tag{9}$$

The federation then updates the global model by aggregating the local models from the participating agents set $\mathcal{N}_p$ as:

$$\omega_i^{up} = \omega_i - \alpha\lambda_u(\omega_i - \tilde{\theta}_i), \tag{10}$$

$$\omega = (1 - \eta)\omega + \eta \sum\nolimits_{i \in \mathcal{N}_p} \omega_i^{up}/N_p. \tag{11}$$

This leads to the following proposition regarding the gradient of the update process in Equation (11) as follows:

$$\nabla F_i(\omega) = -\lambda_u(\omega - \theta_i). \tag{12}$$

Algorithm 1: Flexible-Participation Federation (FPF)

Input: The finite set $\mathcal{N}$ of N agents, participating set $\mathcal{N}_p$ of N_p agents, learning rate α, constrained coefficients λ_u for utilitarianism and λ_e for egalitarianism, entropy coefficient λ_h, update rate η, gradient error ν, replay buffer $\mathcal{D} = \{\mathcal{D}_i\}_{i \in \mathcal{N}}$.
Initialize $\{\theta_i\} = \mathbf{0}, \omega = \bar{\omega} = \mathbf{0}$.
Federation sends ω and $\bar{\omega}$ to all agents.
foreach *agent* $i \in \mathcal{N}$ **do**
  $\omega_i = \omega, \bar{\omega}_i = \bar{\omega}$.
  if *decentralized training* **then**
    Sample the mini-batch $\tilde{\mathcal{D}}_i \subseteq \mathcal{D}_i$.
    Update local parameters till $\|\nabla \tilde{F}_i(\tilde{\theta}_i, \tilde{\mathcal{D}}_i)\|^2 \leq \nu$: $\tilde{\theta}_i = \theta_i + \alpha \nabla \tilde{F}_i(\tilde{\theta}_i, \tilde{\mathcal{D}}_i)$,
    if *agent* $i \in \mathcal{N}_p$ **then**
      Regularize and upload to the federation: $\omega_i^{up} = \omega_i - \alpha\lambda_u(\omega_i - \tilde{\theta}_i), \quad \theta_i = \tilde{\theta}_i$.
    end
  end
end
Federation updates the utilitarian model in Eq.(11).
if *federated training* **then**
  Sample the mini-batch $\tilde{\mathcal{D}} \subseteq \mathcal{D}$.
  Update the federated parameters:
  $\omega = \omega + \eta\alpha\mathbb{E}_{\tilde{\mathcal{D}}}[\mathcal{J}^{uti}(\omega)] - \eta\alpha\lambda_e(w - \bar{w})$,
  $\bar{\omega} = \bar{\omega} + \alpha\mathbb{E}_{\tilde{\mathcal{D}}}[\mathcal{J}^{ega}(\bar{\omega}) - \lambda_h\mathcal{H}(\boldsymbol{\pi}(\bar{\omega}))]$.
end

Theoretical Analysis

In this section, we demonstrate the theoretical convergence of FPF. First, we introduce assumptions regarding the L-smooth property of the egoistic objective function in Assumption 1. Then, the assumptions of bounded variance and diversity of local gradients are defined in Assumptions 2, 3, respectively.

Assumption 1 (L-smooth)**.** *The egoistic objective function $\mathcal{J}_i^{\mathrm{ego}}(\theta)$ with parameters θ are L-smooth:*

$$\|\nabla \mathcal{J}_i^{\mathrm{ego}}(\theta) - \nabla \mathcal{J}_i^{\mathrm{ego}}(\theta')\| \leq L\|\theta - \theta'\|, \forall \theta, \theta'.$$

Assumption 2 (Bounded variance)**.** *For the stochastic gradient $\mathcal{J}_i^{\mathrm{ego}}(\theta; \zeta_i)$ of each sample $x \sim \mathcal{D}^i$ (ζ_i is the distribution of sample x), the expected variance of stochastic gradients can be bounded as:*

$$\mathbb{E}_{\zeta_i}\left[\|\nabla \mathcal{J}_i^{\mathrm{ego}}(\theta; \zeta_i) - \nabla \mathcal{J}_i^{\mathrm{ego}}(\theta)\|^2\right] \leq \iota^2.$$

Assumption 3 (Bounded diversity)**.** *The variance of local gradients relative to the average gradients with the same parameter θ can be bounded as:*

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla \mathcal{J}_i^{\mathrm{ego}}(\theta) - \frac{1}{N}\sum_{j=1}^{N}\nabla \mathcal{J}_j^{\mathrm{ego}}(\theta)\right\|^2 \leq \sigma^2.$$

We can prove that the federated model, as well as the discrepancy between decentralized parameters $\tilde{\theta}_{i,t}$ and federated parameters ω_t, obtains an $\mathcal{O}(1/T)$ convergence rate.

Theorem 1. *Under Assumptions 1, 2, 3 and Lemmas 1, 2, 3 in Appendix 2 with* $\lambda_u > 2\sqrt{2}(\lambda_e+L)$, $\kappa := \alpha\eta\left(\lambda_u+\lambda_e+L\right)$ *and* $\kappa\left(\frac{N/N_p-1}{N-1}\cdot\frac{8(\lambda_e+L)^2}{\lambda_u^2-8(\lambda_e+L)^2}+1\right) < 1/3$, *then we have*

$$(a)\ \mathbb{E}\left[\left\|\nabla F\left(\omega_t\right)\right\|^2\right]$$
$$:= \mathcal{O}\left(\frac{\Delta_F}{\frac{\alpha\eta}{2}\left(1-3\kappa\left(\frac{N/N_p-1}{N-1}\cdot\frac{8(\lambda_e+L)^2}{\lambda_u^2-8(\lambda_e+L)^2}+1\right)\right)}\cdot\frac{1}{T}\right)$$
$$+\mathcal{O}\left(\frac{1+3\kappa\left(\frac{N/N_p-1}{N-1}\cdot\frac{2}{\lambda_u^2-8(\lambda_e+L)^2}\cdot\frac{\sigma^2}{\delta^2}+1\right)}{1-3\kappa\left(\frac{N/N_p-1}{N-1}\cdot\frac{8(\lambda_e+L)^2}{\lambda_u^2-8(\lambda_e+L)^2}+1\right)}\lambda_u^2\delta^2\right),$$

$$(b)\ \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\tilde{\theta}_{i,t}-\omega_t\right\|^2\right] \le \mathcal{O}\left(\mathbb{E}\left[\left\|\nabla F\left(\omega_t\right)\right\|^2\right]\right)$$
$$+\mathcal{O}\left(2\delta^2+\frac{4\sigma^2}{\lambda_u^2-8(\lambda_e+L)^2}\right),$$

where N_p *is the number of agents participating in the federation,* α *is the learning rate, and* η *is the update rate of the federated model in Algorithm 1. Then* $t \in \{0,1,...,T-1\}$, T *is the max time steps,* $\Delta_F := F(\omega_0)-F(\omega_T)$ *and* δ^2 *is defined in Lemma 1 of Appendix 2.*

The proof is provided in Appendix 3. Theorem 1 (a) demonstrates the convergence of FPF, where the first term, related to the initial error $\Delta_F := F(\omega_0)-F(\omega_T) \le F(\omega_0)$, can be bounded by a constant, and the second term consists entirely of constants. Furthermore, the coefficient is dependent on the number of agents (N_p) participating in the federation. As more agents engage, the upper bound becomes tighter. Consequently, the squared magnitude of the federated model's gradient achieves an $\mathcal{O}(1/T)$ convergence rate towards a ball with a constant radius.

Theorem 1 (b) establishes the convergence of the discrepancy between decentralized egoistic policies and the federated utilitarian policy. The last term in (b) indicates that the constrained coefficients λ_u and λ_e influence this discrepancy. As a result, the decentralized parameters $\tilde{\theta}_{i,t}$ on average converge to a ball centered around the federated parameters ω_t with a constant radius.

Local Multi-Federation

This section introduces a more efficient Local Multi-Federation (LMF) mechanism that operates without a global federation while preserving utilitarian and egalitarian policies. Each agent can participate in multiple local federations to exchange local information with others. We propose two communication methods within the same local federation. The first method involves the exchange of local policy parameters and can be theoretically proven to reach consensus. In the second method, agents exchange policy gradients instead of parameters, providing greater privacy while achieving comparable empirical performance in experiments.

Specifically, the LMF framework can be modeled as a time-varying directed graph $\mathcal{G}_t = (\mathcal{N}_p, \mathcal{E}_t)$, where $\mathcal{N}_p$ is the node set representing the participating agents, and the edge set $\mathcal{E}_t \subseteq \mathcal{N}_p \times \mathcal{N}_p$ represents the communication links within federations at time step t. The adjacency matrix of the graph

Algorithm 2: Local Multi-Federation (LMF)

Input: The finite set $\mathcal{N}$ of N agents, learning rate α_t, replay buffer $\mathcal{D} = \{\mathcal{D}_i\}_{i\in\mathcal{N}}$.
Initialize $\{\theta_i\} = \mathbf{0}$, the adjacent matrix C_t.
foreach *agent* $i \in \mathcal{N}$ **do**
  if *decentralized training* **then**
    Sample the mini-batch $\tilde{\mathcal{D}}_i \subseteq \mathcal{D}_i$.
    if *participation in federation* **then**
      Determine the adjacent matrix C_t.
      Update θ_i in Eq.(13) or Eq.(14).
    end
  end
end

$\mathcal{G}_t$ is denoted by $C_t = [c_t(i,j)]_{N_p\times N_p}$, where $c_t(i,j)$ is the weight of the message transmitted from agent j to agent i at time t and $c_t(i,j) = 0$ if and only if $(i,j) \notin \mathcal{E}_t$.

The detailed training procedure of FPF is outlined in Algorithm 2. In the LMF framework, each agent i can receive information from other agents within the same federation. The update rule for each agent i is defined as follows:

$$\theta_{i,t+1} = \sum_{j=1}^{N_p} c_t(i,j)\theta_{j,t} + \alpha_t\mathbb{E}[\nabla_{\theta_{i,t}} J_i^{\text{ego}}(\theta_{i,t})], \quad (13)$$

$$\theta_{i,t+1} = \theta_{i,t} + \alpha_t\sum_{j=1}^{N_p} c_t(i,j)\mathbb{E}[\nabla_{\theta_{j,t}} J_i^{\text{ego}}(\theta_{j,t})]. \quad (14)$$

Here, Eq.(13) represents the first communication method, which involves transmitting policy parameters, while Eq.(14) conveys only policy gradients to enhance privacy among agents. When $c_t(i,i) = 1$ and $c_t(i,j) = 0$ for all $j \neq i$, these two methods are equivalent. Note that LMF considering only the egoistic policy parameters θ rather than the utilitarian and egalitarian policies, can offer greater privacy.

Theoretical Analysis

In this section, we demonstrate the theoretical convergence of LMF under the policy parameter transmission method described in Eq. (13). First, we introduce several assumptions regarding the connectivity properties of the directed graph $\mathcal{G}_t$, the weights rule of the adjacency matrix C_t, and the bounds on local gradients in Assumptions 4, 5 and 6, respectively.

Assumption 4 (Bounded Interval)**.** *There exists an integer* $B \ge 1$ *such that for every edge* (i,j) *must be linked at least every* B *consecutive time steps. In other words, the graph* $(\mathcal{N}_p, \mathcal{E}_t \cup \cdots \cup \mathcal{E}_{t+B-1})$ *is strongly connected for all* $t \ge 0$.

Assumption 5 (Weights Rule)**.** *The weights of the adjacent matrix* C_t *satisfy that: (a)* $\mathbf{1}^\top C_t = \mathbf{1}^\top$ *and* $C_t\mathbf{1} = \mathbf{1}$, *where* $\mathbf{1}$ *is the column vector with all ones entries. (b) There exists a scalar* $\epsilon \in (0,1)$ *such that* $c_t(i,i) \ge \epsilon$ *for all* $i \in \mathcal{N}_p$ *and* $c_t(i,j) \ge \epsilon$ *if* $(i,j) \in \mathcal{E}_t$.

Assumption 6 (Bounded Gradients)**.** *There exists a scalar* $C_L > 0$ *such that* $\left\|\nabla_{\theta_{i,t}} J_i^{\text{ego}}(\theta_{i,t})\right\| \le C_L$, *for all* $i \in \mathcal{N}$ *and* $t \ge 0$.

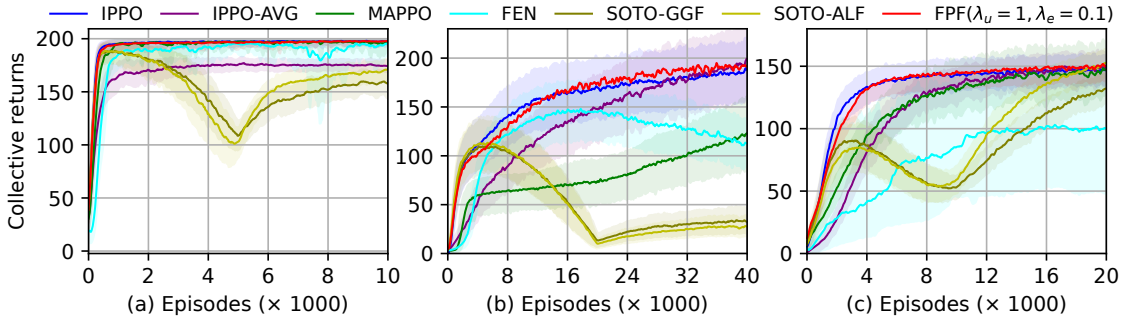

Figure 1: Collective returns of FPF and other baselines per 200 steps in 3 classical discrete scenarios. (a) Job Scheduling. (b) Matthew Effect. (c) Manufacturing Plant.

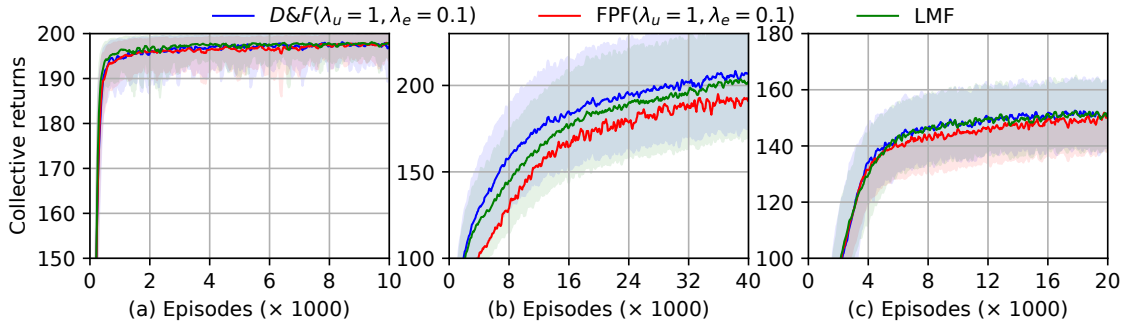

Figure 2: Collective returns of FPF, LMF and the mandatory participation D&F framework per 200 steps in 3 classical discrete scenarios. (a) Job Scheduling. (b) Matthew Effect. (c) Manufacturing Plant.

Theorem 2. *Under Assumptions 4, 5 and 6, if the learning rate* $\alpha_t = \alpha_0/t^\beta$, *where* $\alpha_0 \in (0, 1/N_p(2C_L + 1))$, $\beta \in (1/2, 1)$. *Then for all participation agents* $i \in \mathcal{N}_p$, *we have*

$$\left\| \frac{1}{N_p} \sum_{j=1}^{N_p} \theta_{j,T} - \theta_{i,T} \right\|^2 = \mathcal{O}\left(\frac{1}{T^\beta} + \bar{\epsilon}^T \right). \quad (15)$$

Here, $\bar{\epsilon} := \left(1 - \epsilon^{(N_p-1)B}\right)^{1/(N_p-1)B} \in (0, 1)$ *where* B, ϵ *and* C_L *are defined in Assumptions 4, 5 and 6.*

The proof is provided in Appendix 5. Theorem 2 shows that agents in the LMF framework can reach consensus on the policy parameters with a gap of $\mathcal{O}(1/T^\beta + \bar{\epsilon}^T)$, thereby balancing egoism, utilitarianism and egalitarianism.

Experiments

In this section, we demonstrate the superiority of FPF and LMF across various environments by addressing the following questions: (1) Can FPF and LMF outperform multiple baselines in balancing egoism, utilitarianism and egalitarianism? [in Figures 1, 4, 5, 6]. (2) Can FPF and LMF achieve empirical performance comparable to the mandatory participation D&F framework? [in Figure 2]. (3) Will agents who opt out of the federation experience a reduction in egoism? [in Figure 3]. (4) Does LMF exhibit robust performance in fully cooperative environments? [in Figure 7].

Baselines

We reference the following baselines, which respectively focus on **egoism** (Independent Learning methods such as IDQN (Mnih et al. 2015), IPPO (de Witt et al. 2020)), **utilitarianism** (AVG, QMIX (Rashid et al. 2020), MAPPO (Yu et al. 2022), and MADDPG (Lowe et al. 2017)), and **balancing egoism, utilitarianism and egalitarianism** (MIN, RMF (Zhang and Shah 2014), Inequality Aversion (IA) (Hughes et al. 2018), Social Influence (SOCIAL) (Jaques et al. 2019), FEN (Jiang and Lu 2019), SOTO-GGF (Zimmer et al. 2021), SOTO-ALF (Zimmer et al. 2021), and D&F (Dong et al. 2024b)). A detailed analysis of these baselines is in Appendix 6.1.

Classic Discrete Scenarios

Classic Discrete Scenarios (Jiang and Lu 2019), including Job Scheduling, Matthew Effect, and Manufacturing Plant, are defined by resource constraints that drive agents to engage in dynamic competition or cooperation within mixed-motive games. Detailed descriptions of these scenarios are provided in Appendix 6.2.1.

Utilitarianism. We first assess the utilitarian performance of FPF in comparison to other baselines. As depicted in Figure 1, our results indicate that IPPO and FPF achieve the highest social welfare, while IPPO-AVG and MAPPO underperform. FEN, SOTO-GGF, and SOTO-ALF overemphasize egoism and egalitarianism, leading to suboptimal utilitarian

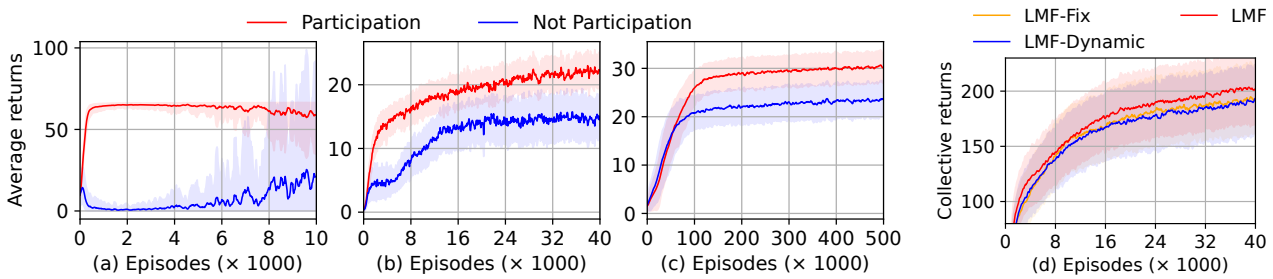

Figure 3: (a∼c) Average returns of the participating agents and not participating agents in FPF in Job Scheduling, Matthew Effect, and Manufacturing Plant scenarios, respectively. (d) Collective returns of LMF, LMF-Fix, and LMF-Dynamic mechanisms in the Matthew Effect scenario.

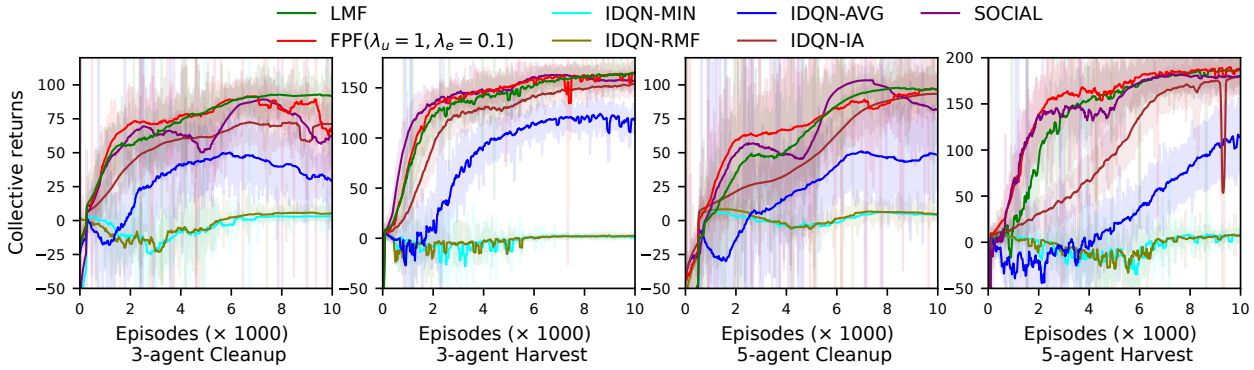

Figure 4: Collective returns of different algorithms per 100 steps in Cleanup and Harvest environments with 3 and 5 agents.

performance. Additionally, we compare the performance of different federation frameworks, including the mandatory participation D&F, our proposed FPF, and LMF. We initialize the experiment with half of the agents participating in FPF and all agents participating in LMF. The results in Figure 2 show that D&F achieves the highest social welfare, while LMF with full participation performs comparably and outperforms FPF with partial participation.

Egoism Analysis and Federation Mechanism. To further investigate the federation mechanisms of FPF and LMF, we compare the average returns of agents participating in the federation versus those opting out in FPF. As illustrated in Figure 3(a,b,c), agents who opt out of federation participation experience a decrease in rewards, demonstrating that FPF effectively balance egoism. Additionally, we analyze LMF formation as LMF-Fix and LMF-Dynamic. LMF-Fix refers to the division of all agents into two disjoint federations at the outset, while LMF-Dynamic allows for the formation of multiple joint federations. Figure 3(d) demonstrates that LMF-Fix achieves better utilitarian performance than LMF-Dynamic, as the formation of multiple joint federations complicates reaching consensus among all agents.

Egalitarianism. We evaluate the egalitarianism of different algorithms using the ratio of standard variance to mean value in agents' rewards. As depicted in Figure 5, our proposed FPF and LMF achieve a similarly low ratio, closely

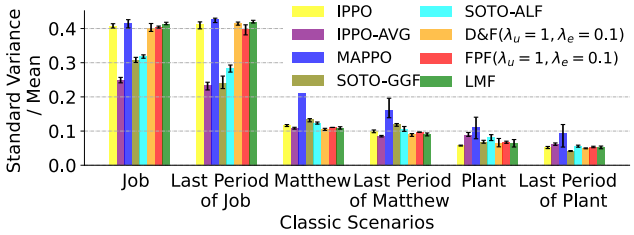

Figure 5: The ratio of standard variance to mean value in rewards in Classic Discrete Scenarios. Each scenario is compared in the average and the last period of the training.

matching the performance of the mandatory participation D&F in the scenarios of Matthew Effect and Manufacturing Plant. However, in the Job Scheduling scenario, IPPO-AVG and SOTO outperform FPF and LMF in terms of egalitarianism. We hypothesize that this is due to the presence of only one resource in the environment, which makes it challenging for FPF and LMF to penalize greedy agents effectively without explicitly incorporating reward shaping.

Sequential Social Dilemmas (SSD)

Subsequently, we conduct experiments in Sequential Social Dilemmas (SSD), a mixed-motive game, as introduced in

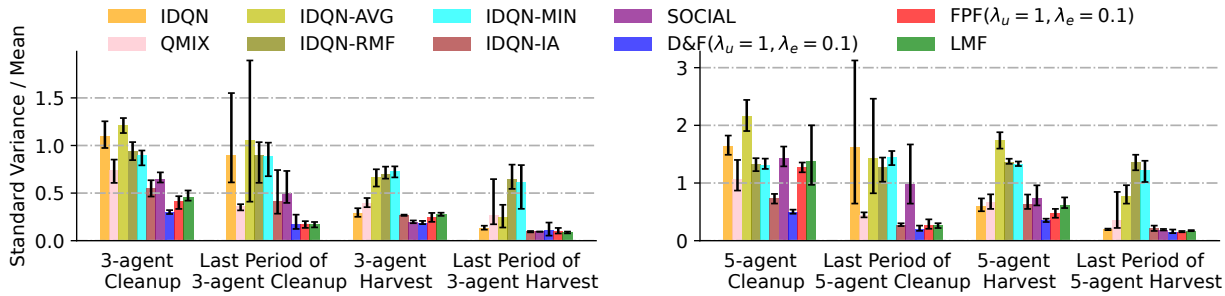

Figure 6: The ratio of standard variance to mean value in agents' rewards with different algorithms in Cleanup and Harvest environments with 3 agents and 5 agents. Each scenario is compared both in the average and the last period of the training.

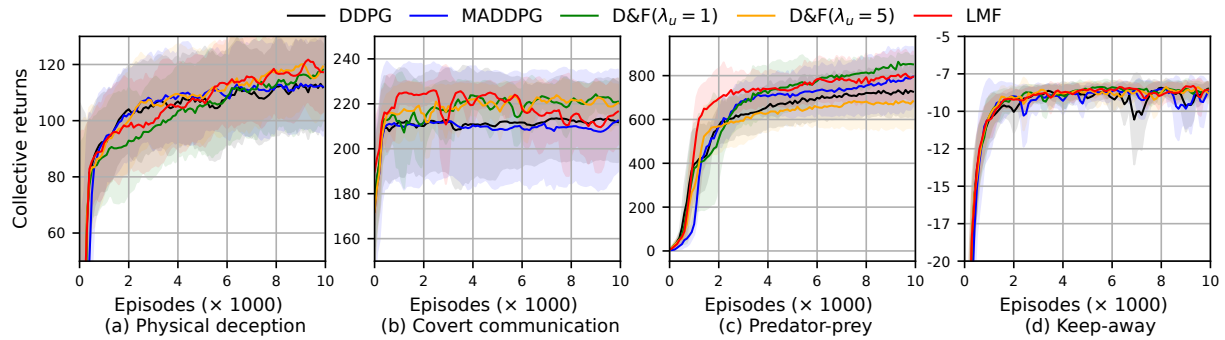

Figure 7: Collective returns per 1000 steps in MPE with random policy of adversary agent. (a) Physical deception. (b) Covert communication. (c) Predator-prey. (d) Keep-away.

(Leibo et al. 2017; Hughes et al. 2018; Jaques et al. 2019). The SSD comprises two social dilemma environments: *public goods dilemmas* (Cleanup) and *commons dilemmas* (Harvest). Each agent strives to maximize its individual returns through latent cooperative or competitive interactions with others.

Utilitarianism. We evaluate the utilitarian performance of FPF, LMF and other baselines in 3-agent and 5-agent Cleanup/Harvest environments. As shown in Figure 4, our proposed FPF and LMF outperform other baselines in achieving the best social welfare. Especially, IDQN-MIN and IDQN-RMF methods fail to learn an efficient policy that balances egoism and utilitarianism in the Cleanup and Harvest environments. Additional comparisons of utilitarian performance with the mandatory participation D&F, IDQN, and QMIX are available in Appendix 6.3.3.

Egalitarianism. In Figure 6, our proposed FPF and LMF achieve a similarly low ratio of standard variance to mean value in agents' rewards, comparable to that of the mandatory participation D&F framework. In contrast, IDQN, which focuses solely on egoism, and IDQN-AVG, which considers only utilitarianism, exhibit significant fluctuations.

Multi-Agent Particle Environments (MPE)

Finally, we evaluate the robust performance of LMF in fully cooperative MPE, where cooperative agents share the same reward. Specifically, we conduct experiments in the following four basic testing scenarios: Physical deception, Covert communication, Predator-prey and Keep-away. Detailed descriptions of these scenarios are available in Appendix 6.4.1.

Utilitarianism. First, we conduct the experiment with the adversary agent's policy fixed as a random policy. Since these scenarios only involve only three to four agents, we only compare the performance of full-participation LMF. As shown in Figure 7, we can find that LMF and D&F achieve nearly the best utilitarian performance. Notably, in the Predator-Prey scenario, LMF outperforms the mandatory participation D&F. Additionally, we pre-train the policy of the adversary agent as a DDPG policy, the final performance of different algorithms is available in Appendix 6.4.3.

Conclusion

In this paper, we present the Flexible-Participation Federation (FPF) and Local Multi-Federation (LMF) frameworks to balance egoism, utilitarianism, and egalitarianism in mixed-motive games. We provide a theoretical analysis of the convergence properties of FPF and the consensus properties among agents in LMF. Extensive experimental results demonstrate that FPF and LMF outperform multiple baselines in terms of both utilitarianism and egalitarianism. Agents who opt out of federation participation experience a reduction in egoism.

In future work, we will explore a free entry and exit mechanism within the federation during the training process. Additionally, a robust metric should be developed to evaluate the exploitation risk of participating agents in the federation.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No.62476136, No.62192783, No.62206133), the Collaborative Innovation Center of Novel Software Technology and Industrialization and the Qing Lan Project of Jiangsu Province, China. What's more, the authors greatly thank all anonymous reviewers for their valuable comments to this work.

References

Anastassacos, N.; García, J.; Hailes, S.; and Musolesi, M. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 115–123.

Anastassacos, N.; Hailes, S.; and Musolesi, M. 2020. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 7047–7054.

Arunarani, A.; Manjula, D.; and Sugumaran, V. 2019. Task scheduling techniques in cloud computing: A literature survey. *Future Generation Computer Systems*, 91: 407–415.

Baker, B. 2020. Emergent reciprocity and team formation from randomized uncertain social preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 15786–15799.

Chelarescu, P. 2021. *Mitigating exploitation caused by incentivization in multi-agent reinforcement learning*. Master's thesis, University of Edinburgh.

Chen, Z.-S.; Zhu, Z.; Wang, X.-J.; Chiclana, F.; Herrera-Viedma, E.; and Skibniewski, M. J. 2023. Multiobjective Optimization-Based Collective Opinion Generation With Fairness Concern. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(9): 5729–5741.

de Witt, C. S.; Gupta, T.; Makoviichuk, D.; Makoviychuk, V.; Torr, P. H. S.; Sun, M.; and Whiteson, S. 2020. Is independent learning all you need in the Starcraft multi-agent challenge? *arXiv:2011.09533*.

Dong, S.; Li, C.; Yang, G.; Ge, Z.; Cao, H.; Chen, W.; Yang, S.; Chen, X.; Li, W.; and Gao, Y. 2024a. Survey on Solutions and Applications for Mixed-motive Games. *Journal of Software*, 1–47.

Dong, S.; Li, C.; Yang, S.; An, B.; Li, W.; and Gao, Y. 2024b. Egoism, utilitarianism and egalitarianism in multi-agent reinforcement learning. *Neural Networks*, 178: 106544.

Heidari, H.; Ferrari, C.; Gummadi, K.; and Krause, A. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1273–1283.

Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; Dueñez-Guzman, E.; García Castañeda, A.; Dunning, I.; Zhu, T.; McKee, K.; Koster, R.; et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3330–3340.

Ivanov, D.; Egorov, V.; and Shpilman, A. 2021. Balancing Rational and Other-Regarding Preferences in Cooperative-Competitive Environments. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1536–1538.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 3040–3049.

Jiang, J.; and Lu, Z. 2019. Learning fairness in multi-agent systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 13854–13865.

Köster, R.; ; et al. 2020. Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences. *arXiv:2010.09054*, 1–24.

Köster, R.; Hadfield-Menell, D.; Everett, R.; Weidinger, L.; Hadfield, G. K.; and Leibo, J. Z. 2022. Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3): e2106028118.

Leibo, J. Z.; Zambaldi, V. F.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 464–473.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 6379–6390.

McKee, K. R.; Gemp, I.; McWilliams, B.; Duèñez-Guzmán, E. A.; Hughes, E.; and Leibo, J. Z. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 869–877.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M. A.; Fidjeland, A.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Ostrom, E. 2008. The challenge of common-pool resources. *Environment: Science and Policy for Sustainable Development*, 50(4): 8–21.

Perolat, J.; Leibo, J. Z.; Zambaldi, V.; Beattie, C.; Tuyls, K.; and Graepel, T. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3643–3652.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(1): 7234–7284.

Sen, S.; and Airiau, S. 2007. Emergence of Norms through Social Learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1507–1512.

Speicher, T.; Heidari, H.; Grgic-Hlaca, N.; Gummadi, K. P.; Singla, A.; Weller, A.; and Zafar, M. B. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2239–2248.

Vinitsky, E.; Köster, R.; Agapiou, J. P.; Duéñez-Guzmán, E. A.; Vezhnevets, A. S.; and Leibo, J. Z. 2023. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2): 26339137231162025.

Vinitsky, E.; Lichtlé, N.; Yang, X.; Amos, B.; and Foerster, J. 2022. Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3962–3974.

Yang, J.; Li, A.; Farajtabar, M.; Sunehag, P.; Hughes, E.; and Zha, H. 2020. Learning to incentivize other learning agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 15208–15219.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 24611–24624.

Zhang, C.; and Shah, J. A. 2014. Fairness in multi-agent sequential decision-making. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2636–2644.

Zhou, M.; Luo, J.; Villella, J.; Yang, Y.; Rusu, D.; Miao, J.; Zhang, W.; Alban, M.; Fadakar, I.; Chen, Z.; et al. 2021. SMARTS: An open-source scalable multi-agent rl training school for autonomous driving. In *Conference on Robot Learning (CoRL)*, 264–285.

Zimmer, M.; Glanois, C.; Siddique, U.; and Weng, P. 2021. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 12967–12978.