Achieving Equilibrium Under Utility Heterogeneity: An Agent-Attention Framework for Multi-Agent Multi-Objective Reinforcement Learning

Zhuhui Li[1], Chunbo Luo[1], Liming Huang[2], Luyu Qi[3], and Geyong Min[1]
[1]Department of Computer Science, University of Exeter, EX4 4QF, UK
[2]School of Mechanical and Electrical Engineering, Central South University, Hunan, China
[3]Faculty of Science and Engineering, University of Bristol, BS8 1UB Bristol, U.K
{zl462, c.luo}@exeter.ac.uk, huanglm.me@gmail.com, luyu.qi@bristol.ac.uk, g.min@exeter.ac.uk

Abstract

Multi-agent multi-objective systems (MAMOS) have emerged as powerful frameworks for modelling complex decision-making problems across various real-world domains, such as robotic exploration, autonomous traffic management, and sensor network optimisation. MAMOS enhances scalability and robustness through decentralised control and more accurately captures inherent trade-offs between conflicting objectives. In MAMOS, each agent uses utility functions that map return vectors to scalar values. Existing MAMOS optimisation methods face significant challenges in handling heterogeneous objective and utility function settings, where training non-stationarity is intensified due to private utility functions and the associated policies. In this paper, we first theoretically prove that direct access to, or structured modeling of, global utility functions is necessary to achieve the Bayesian Nash Equilibrium under decentralised execution constraints. To access the global utility functions while preserving the decentralised execution, we propose an Agent-Attention Multi-Agent Multi-Objective Reinforcement Learning (AA-MAMORL) framework. Our approach implicitly learns a joint belief over other agents' utility functions and their associated policies during centralised training, effectively mapping global states and utilities to each agent's policy. During execution, each agent independently selects actions based on local observations and its private utility function to approximate a BNE, without relying on inter-agent communication. We evaluate our framework through extensive experiments in a custom-designed MAMO Particle environment and the standard MOMALand benchmark. The results demonstrate that accessibility to global preferences and our proposed AA-MAMORL significantly improves performance and consistently outperforms state-of-the-art methods.

Introduction

Multi-Agent Multi-Objective Systems (MAMOSs) have been spotlighted in real-world applications, such as balancing exploration and exploitation in networked robotic systems (Paine and Benjamin 2024), managing the trade-off between efficiency and energy consumption in autonomous traffic control (Shi et al. 2021), and optimising the trade-off between resolution and coverage in mobile sensor monitoring tasks (Hayat et al. 2020). In contrast to single-agent systems where the decision burden and failure risk are centralised, the MAMOS distributes both computation and control across agents. This decentralisation enhances system scalability and enables resilience and robustness to partial agent failures (He et al. 2021). Meanwhile, the multi-objective formulation in MAMOSs also better reflects the inherent trade-offs in real-world applications. To be specific, most real-world systems require trade-offing between multiple, often conflicting, performance metrics. Representative MO settings include energy efficiency (Niu et al. 2023), energy performance index (Chang, Iqbal, and Chen 2023), and water use efficiency (Mallareddy et al. 2023), often in conjunction with advanced integrated technologies such as Simultaneous Wireless Information and Power Transfer (Wei et al. 2021), Integrated Sensing and Communication (Qi et al. 2022), and piezoelectric roads (Jiang et al. 2023).

Although steady progress has been made in the development of MAMOSs, several critical challenges remain, including the joint interdependencies among objectives and agents, the dynamic nature of real-world systems, and the non-differentiability of many environments (Wong et al. 2023). The heterogeneity and diversity of reward (corresponding to objective) and utility function settings in MAMOS further pose significant challenges to the MAMO optimisation. As discussed in (Rădulescu et al. 2020), the utility function is defined as a mapping from the vectorised rewards to a scalar utility. The utility function in this paper is referred to as the preference, where the weighted sum based on rewards and preferences forms the most basic utility function for all agents. The decision-making problems in MAMOS can be categorised into five settings, based on the combinations of reward and utility function types. On the reward side, agents receive either a team reward, where all agents share the same reward vector reflecting collective performance, or an individual reward, where each agent obtains a personalised reward vector. On the utility side, agents optimise a shared team utility, pursue a social choice utility that aggregates all agents' rewards into a global social welfare function, or optimise their own individual utility, where each agent maintains a private function. Examples of these combinations in real-world companies are illustrated in Fig. 1.

In the simplest setting: team utility with team reward, the problem can be simplified into a single-agent formulation,

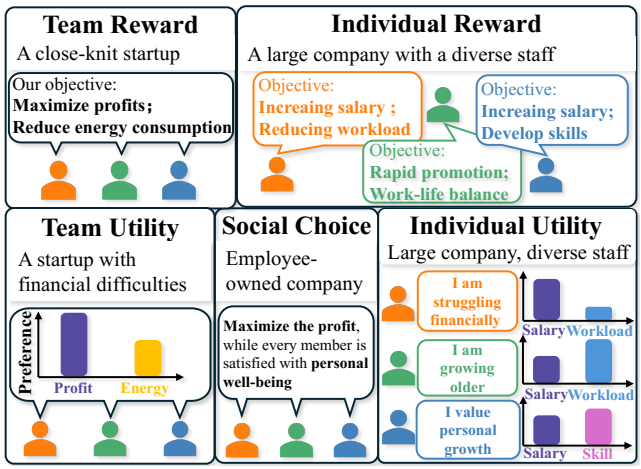


Figure 1: Illustrative examples of the five types of reward and utility function setting, using different companies as analogies.

where one entity optimises its joint policy by acting over the entire joint action space (Rădulescu et al. 2020). In the more challenging team utility with individual reward setting, Hu et al. (Hu et al. 2023) have proposed MO-MIX to solve Partially Observable MO Markov Decision Process (POMOMDP)s within the Centralised Training with Decentralised Execution (CTDE) framework. Their approach incorporates a Multi-Objective Conditioned Agent Network for evaluating action-values under the team utility, a parallel mixing network for estimating joint action-values, and a preference-based exploration strategy to promote diverse and well-distributed Pareto-optimal policies.

However, the optimisation in the most general setting: individual utility functions, remains unsolved. But this setting is critical as it is able to reflect most real-world scenarios where agents act based on personal intentions, constraints, or their distinct roles in the environment. Designing a general optimisation framework for this setting is challenging, with the reason that each agent must optimise its policy according to its own utility function, while the reward of that policy depends on the joint policy with other agents under their individual utility functions. Since these utility functions are heterogeneous and potentially conflicting, each agent's policy cannot be updated and performed in isolation (Assos, Dagan, and Daskalakis 2024).

This naturally leads to the optimisation in MAMOSs as a Bayesian game, where each agent possesses private information, such as utility functions, objectives, or local observations, and must form beliefs about the policies of others to make its own optimal decision. The canonical solution concept in such settings is the Bayesian Nash Equilibrium (BNE) (Saglam 2025), a joint policy in which no agent can unilaterally improve its expected utility, given its beliefs about the types and policies of others. Achieving BNEs in MAMOS settings is non-trivial due to the intensified non-stationarity introduced by other agents' private utility functions, multiple rewards settings, and the associated policies (Assos, Dagan, and Daskalakis 2024).

While CTDE has become a dominant paradigm in MARL to optimise the individual policy (actor) by forming beliefs about the policies of others in the centralised critic network (Li, Wang, and Xu 2025), it does not guarantee convergence to BNE when agents' utility functions are partially or entirely unknown to each other in MAMOSs, since the CTDE framework will still fail to capture or model utility-based policies with the unknown utility function. As a result, the learning process is still highly non-stationary from each agent's perspective, and the BNE remains inapplicable. Thus, BNE in MAMOs requires that every agent possesses a model of others' utility functions and the corresponding policies those utilities induce. This fundamental requirement constitutes the main problem to be addressed in this paper: **How to map the global state, utility functions, and the associated joint policies to each agent's individual policy, so that each agent can learn its own decentralised optimal policy that maximise the global utility, even under heterogeneous objective and utility function settings.**

We find that belief modeling on the joint policy becomes tractable, and BNE becomes theoretically attainable when each agent's utility function is a deterministic function of the state or observation (e.g., $\boldsymbol{w}_i = g(o_i)$). This case also better aligns with many real-world applications. For instance, when purchasing train tickets, a traveler facing tight time constraints may prioritize on-time arrival over cost, whereas one planning in advance during a pre-sale period is more likely to prefer the lowest possible price. Building on this finding, we first prove that when each agent's utility function is either directly observable or can be modeled from local observations, convergence to a BNE becomes theoretically attainable. Then, we propose an agent-attention MAMORL framework. This framework implicitly learns a joint belief over other agents' utility functions and their corresponding policies through centralized agent-attention-based critic training. Thus, each agent learns a mapping from the global state and utility functions to its own policy under such belief. The learnt distributed policy of each agent can be executed using only its local observations and private utility function. And no agent has an incentive to unilaterally deviate from its decision given the system-wide context. Thereby the BNE of MAMOSs can be approximated in a fully decentralized and communication-free setting. Our main contributions are summarized as follows:

1. We formalise the MAMOS optimisation as a general POMOMDP, capable of abstracting various applications with heterogeneous reward and utility function settings.

2. We bridge between the POMOMDP and Bayesian games, and rigorously prove that even under the CTDE paradigm, the utility function is essential for achieving BNE in decentralised decision-making.

3. We propose an agent-attention MAMORL framework for scenarios where utility functions are deterministic functions of agents' local observations. This framework enables the optimal decentralised policy conditioned solely on its local observation and private utility function for each agent.

4. We conduct comprehensive experiments in the MAMO Particle environment and the MOMALand benchmark. The

results demonstrate that the global utility functions and our proposed AA-MAMORL framework consistently improve multiple MO metrics.

Preliminaries

Partially Observable MO Markov Decision Process

POMOMDP is defined by the tuple: $\Omega = (S, \boldsymbol{\mathcal{A}}, \boldsymbol{R}, \boldsymbol{W}, \boldsymbol{P}_o, \boldsymbol{P}_{wt}, P_t, P_0, \gamma)$. Within this process, S is the state space describing the possible states of all agents and the environment, $\mathcal{A}_1, \ldots, \mathcal{A}_N \in \boldsymbol{\mathcal{A}}$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N \in \boldsymbol{W}$ are the action and preference spaces for all agents. At each time slot, agent i first uses its observation function $P_o^i \in \boldsymbol{P}_o : S \to O_i$ to obtain its own observation o_i based on the state $s \sim S$. Each agent i uses its MO policy $\pi_i : O_i \times \boldsymbol{w}_i \to \mathcal{A}_i$ to select its action $a_i \sim \mathcal{A}_i$ since observations coupled with the utility function affect agents' decisions jointly. The transition function $P_t : S \times \mathcal{A}_1 \times \ldots \times \mathcal{A}_N \to S$ transits the current state s_t to s_{t+1}. The preference transition function $P_{wt}^t \in \mathbf{P}_{wt}$ transits the preference $\boldsymbol{w}_i[t]$ to $\boldsymbol{w}_i[t+1]$. Its setting is divided in two cases in the following sections. P_0 is the distribution function of the initial state of the environment. Finally, each agent i obtains vectorised rewards as a function of the state and joint action $\boldsymbol{r}_i \in \boldsymbol{R} : S \times \mathcal{A}_1 \times \ldots \times \mathcal{A}_N \to \boldsymbol{R} : \mathbb{R}^m$. The objective of MAMO optimisation is the optimal joint policy which maximises the weighted sum of all agents' rewards and corresponding preferences: $R = \sum_{i=0}^{N} \sum_{t=0}^{T} \gamma^t \boldsymbol{w}_i^\top \boldsymbol{r}_i^t$, where $\gamma \in [0, 1]$ denotes the discount factor.

The Necessity of Global Preferences in POMOMDP Decision-Making

In this section, we theoretically demonstrate that the attainability of BNE in POMOMDP depends on the observability or structural modeling of global preference.

Case I: Preferences as Unstructured Random Variables

Assume $\boldsymbol{w}_i \sim \text{Unif}(\Delta^k)|_{\Delta^k = \boldsymbol{w} \in \mathbb{R}^k | \sum_j w_j = 1, w_j \geq 0}$. Each agent's preference is independently and uniformly distributed.

Theorem 1 (BNE Inapplicability with Unobservable, Uniform Preferences)**.** *Suppose that for any $i \neq j$, agent i knows only that other agent's preference $\boldsymbol{w}_j \sim \text{Unif}(\Delta^k)$. Then the classical BNE concept is inapplicable*

Theorem 2 (BNE Attainability with Observable Uniform Preferences)**.** *Let each agent j's preference weight $\boldsymbol{w}_j \sim \text{Unif}(\Delta^k)$ be drawn independently, and assume that for every pair $i \neq j$, agent i* observes *$\boldsymbol{w}_j$ prior to choosing its action. Then BNE in behavioral strategies exists.*

Case II: Preferences as State-dependent Functions

Theorem 3 (BNE Existence when $\boldsymbol{w}_i = g(o_i)$)**.** *Suppose each agent i 's preference weight $\boldsymbol{w}_i$ is a* deterministic *function of its private observation $\boldsymbol{w}_i = g(o_i)$ where $g : O_i \to \Delta^k$ is continuous. Then, under the usual compactness and continuity assumptions on observations and actions, a mixed-strategy BNE exists.*

Proof. The proofs of above theorems are provided in (Li et al. 2025). □

General Multi-Agent Multi-objective Reinforcement Learning

In a game with N agents, each agent has M objectives, and the corresponding preference vector $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N\}$ indicates the importance of each objective for the corresponding agent, where $\boldsymbol{w}_i = \{w_i^1, \ldots, w_i^M | \sum_{j=1}^{M} w_i^j = 1\}$. Based on the above theorem, we further develop distinct MAMORL frameworks designed for whether preferences are modeled as unstructured random variables or observation-dependent functions. These frameworks are designed to generalise across diverse real-world settings of states, actions, preferences, and rewards, thereby enabling robust optimisation in MAMOS scenarios.

Global-preference-based MAMORL for Case I

The policy set $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_N\}$ is assigned to all agents and is parametrised by $\boldsymbol{\theta^\pi} = \{\theta^{\pi_1}, \ldots, \theta^{\pi_N}\}$. According to Theorem 2, the global preference $\boldsymbol{W}$ is necessary for all agents to reach BNE. Thus, the input is the observation o_i achieved from the state s and the global preference $\boldsymbol{W}$. The probability of each action $\pi_i(a_i|o_i, \boldsymbol{W})$ is the output.

$\boldsymbol{v}_i^{\pi_i} : \mathbb{R}^m$ is the vectorised MO state-value of the policy π_i, which approximates the expected rewards under the state $s[t]$ and the given global preference $\boldsymbol{W}$, denoted as:

$$\begin{aligned} v_i^{\pi_i}(s[t], \boldsymbol{W}) =& \mathbb{E}_{\boldsymbol{A}[t] \sim \boldsymbol{\pi}(\cdot|s[t], \boldsymbol{W}), s[t+1] \sim P(\cdot|s[t], \boldsymbol{A}[t])} \\ & \left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_i \left(s[t], \boldsymbol{A}[t] \right) \right], \\ \boldsymbol{A}[t] =& \{a_1[t], .., a_N[t]\}, \\ \boldsymbol{W} =& \{\boldsymbol{w}_1, .., \boldsymbol{w}_N \| \boldsymbol{w}_i \sim \text{Unif}(\Delta^k)\}. \end{aligned} \tag{1}$$

This MO state-value vector can be linearly combined with the preference $\boldsymbol{w}_i$: $v_i^{\pi_i} = \boldsymbol{w}_i^\top \boldsymbol{v}_i^{\pi_i}$. The objective of each agent is to find the policy π_i which maximises the excepted $v_i^{\pi_i}$ under the initial state distribution P_0 and given preference $\boldsymbol{w}$: $\mathbb{E}_{s[0] \sim p[0]} v_i^{\pi_i}(s[0], \boldsymbol{W})$.

The vectorised MO action-value function for the policy π_i based on the state-action-preference tuple $(s, \boldsymbol{A}, \boldsymbol{W}) | \boldsymbol{A} = \{a_i, .., a_N\}$ is utilised to approximate the expected rewards under the policy π_i, which is defined as Eq. 2.

$$\begin{aligned} & \boldsymbol{q}_i^{\pi_i}(s, \boldsymbol{A}, \boldsymbol{W}) = \mathbb{E}_{\pi_i}[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_i(s[t], \boldsymbol{A}[t])], \\ & = \mathbb{E}_{s[t+1] \sim P(\cdot|s[t], \boldsymbol{A}[t])} \left[\boldsymbol{r}_i(s[t], \boldsymbol{A}[t]) + \gamma \boldsymbol{v}_i^{\pi_i}(s[t+1], \boldsymbol{W}) \right], \end{aligned} \tag{2}$$

where $\boldsymbol{q}_i^{\pi_i}(s, \boldsymbol{A}, \boldsymbol{W})$ is an m-dimensional vector representing the expected rewards of m objectives for the agent i. It extends the MO state-value vector by explicitly incorporating the current action, which can be directly optimised in policy learning.

A centralised trained MO action-value function $\boldsymbol{Q}_i^{\pi_i}(s, a_1, \ldots, a_N, \boldsymbol{W} | \theta^{\boldsymbol{Q}_i})$ parameterised by $\theta^{\boldsymbol{Q}_i}$ is deployed to represent $\boldsymbol{q}_i^{\pi_i}(s, \boldsymbol{A}, \boldsymbol{W})$. The inputs are the actions of all agents $a_1, \ldots, a_N$, the preference settings of all agents $\boldsymbol{W}$, and the state information s. It outputs represent the approximate expected rewards $\boldsymbol{v}_i^{\pi_i}$.

The policy of each agent π_i is updated by the gradient of the expected return $J(\theta^{\pi_i}) = \mathbb{E}_{s[t],a[t]\sim\pi}\left[\boldsymbol{w}_i^\top \boldsymbol{r}_i\left(s[t],a[t]\right)\right]$ aimed at maximising the weighted sum of the approximate expected rewards $\boldsymbol{Q}_i^{\pi_i}$ and the current preference $\boldsymbol{w}_i$, which is represented as:

$$\begin{aligned}\nabla_{\theta^{\pi_i}} J(\theta^{\pi_i};\boldsymbol{W}) =&\mathbb{E}_{s\sim\rho^\pi,a_i\sim\pi_i}[\nabla_{\theta^{\pi_i}}\log\pi_i(a_i\mid o_i,\boldsymbol{W})\\ &\boldsymbol{w}_i^\top \boldsymbol{Q}_i^{\pi_i}(s,a_1,\ldots,a_i,\ldots,a_N,\boldsymbol{W}\mid\theta^{\boldsymbol{Q}_i})],\end{aligned} \tag{3}$$

where ρ^π is the state distribution induced by the policy π. This framework can also be extended to deterministic policies. The policy is reformulated as the continuous action version: $\boldsymbol{\mu} = \{\mu_1(o_1,\boldsymbol{W}\mid\theta^{\mu_1}),\ldots,\mu_N(o_N,\boldsymbol{W}\mid\theta^{\mu_N})\}$. The corresponding MAMO Deep Deterministic Policy Gradient (MAMODDPG) is denoted as:

$$\begin{aligned}\nabla_{\theta^{\mu_i}} J(\theta^{\mu_i};\boldsymbol{W}) = \mathbb{E}_{s,a,\boldsymbol{W}\sim\mathcal{D}}[\nabla_{a_i}\boldsymbol{w}_i^\top \boldsymbol{Q}_i^{\mu_i}(s,a_1,\ldots,a_i,\ldots,\\ a_N,\boldsymbol{W}\mid\theta^{\boldsymbol{Q}_i})\mid_{a_i=\mu_i(o_i,\boldsymbol{W}\mid\theta^{\mu_i})}\nabla_{\theta^{\mu_i}}\mu_i(o_i,\boldsymbol{W}\mid\theta^{\mu_i})],\end{aligned} \tag{4}$$

where the experience replay buffer $\mathcal{D}$ contains tuples $(s,s',a_1,\ldots,a_N,\boldsymbol{W},\boldsymbol{r}_1,\ldots,\boldsymbol{r}_N)$. By sampling the experiences from $\mathcal{D}$, all agents' actor networks are updated by maximising the MAMODDPG, and all agents' critic networks are updated by minimising the MO temporal difference (MOTD) error for a more accurate approximation:

$$\begin{aligned}L(\theta^{\boldsymbol{Q}_i}) =&\mathbb{E}_{s,s',\boldsymbol{a},\boldsymbol{W},\boldsymbol{r}\sim\mathcal{D}}\\ &\left[\boldsymbol{Q}_i^{\mu_i}(s,a_1,\ldots,a_i,\ldots,a_N,\boldsymbol{W}\mid\theta^{\boldsymbol{Q}_i})-\boldsymbol{y}_i\right]^2.\end{aligned} \tag{5}$$

$$\begin{aligned}\boldsymbol{y}_i =&\boldsymbol{r}_i+\gamma\boldsymbol{Q}_i^{'\mu_i'}(s',a_1',\ldots,a_i^{GPI},\ldots,a_N',\boldsymbol{W}\mid\theta^{\boldsymbol{Q}_i'})\\ &\mid_{a_j'=\mu_j'(o_j',\boldsymbol{W}),a_i^{GPI}=\mu_i^{GPI}(o_i',\boldsymbol{W})},\end{aligned} \tag{6}$$

where $\boldsymbol{\mu}' = \{\mu_1',\ldots,\mu_N'\}$ and $\boldsymbol{Q}_i^{'\mu_i'}$ are the target actor networks and critic networks for all agents maintained during the centralised training. μ_i^{GPI} is the agent's policy integrated with Generalised Policy Improvement (Yang, Sun, and Narasimhan 2019), which assists in the rapid exploration of the entire preference space. In GPI, an alternative policy set for each agent Π_i is maintained. All policies in Π_i are used to generate multiple actions, and the one with the maximum expected return $\boldsymbol{Q}_{max}^{\pi_i^*}(s,a)$ is selected by the agent i. For the deterministic policy and MAMO context, it is reformulated as:

$$\begin{aligned}\mu_i^{GPI}(o_i,\boldsymbol{W}) =&\mu_i(o_i,\arg\max_{\boldsymbol{w}_i'\sim\boldsymbol{\Psi}_i}\boldsymbol{w}_i^\top \boldsymbol{Q}_i^{\mu_i}(o_i,a_1',\ldots,\\ &\mu_i(o_i,\boldsymbol{W}'),\ldots,a_N',\boldsymbol{w}_1,\ldots,\boldsymbol{w}_i',\ldots,\boldsymbol{w}_N)).\end{aligned} \tag{7}$$

The policy set Π_i is replaced by policies generated with the random global preferences $\boldsymbol{W}' = \{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_i',\ldots,\boldsymbol{w}_N\}$, where other agents' preferences are fixed while the preference of itself is randomly sampled from the preference distribution $\boldsymbol{\Psi}_i$. The critic network $\boldsymbol{Q}^{\mu_i}$ generates the expected action-value vectors from different preferences. After the weighted sum with the given preference $\boldsymbol{w}_i$, the maximum summation is selected, and the associated action $\mu_i(o_i,\boldsymbol{W}^*)$ is selected as the optimal action.

For the optimisation for unstructured random preference, the global preference ensures each agent maintains a consistent belief over the global joint policy during decision-making. This mechanism facilitates consensus among agents toward maximising the global utility, mitigating the non-stationarity in the evolving joint MO policy, and enabling the BNE. The following experiments demonstrate the global preference leads to an improvement in overall utility. Such a design is appropriate in scenarios where global coordination is critical and communication is available, such as swarm robotics or UAV formations.

Agent-Attention MAMORL for Case II

The global preference above inevitably violates the principle of decentralised execution in the CTDE paradigm, introducing non-negligible communication overhead into distributed systems. In scenarios where inter-agent communication is entirely infeasible, such as disaster-response missions in disastrous environments, this approach becomes incompatible with the constraints of fully decentralised systems. Consequently, we observe that assigning preferences as completely random variables departs from several real-world applications, where agent preferences are often shaped by environmental states or agents' observations.

To address this gap, we then design an agent-attention MAMORL (AA-MAMORL) framework that better aligns with practical scenarios where preferences are observation-dependent in this section. This framework maintains the integrity of CTDE, while allowing all agents to converge toward a BNE and jointly optimise the global utility in a scalable and communication-efficient manner. Based on theorem 3, when preference $\boldsymbol{w}_i$ is a deterministic function of the agent's observation $\boldsymbol{w}_i = g_i(o_i)$, the global decision is no longer needed for the decision. Thus, the policy for each agent is defined as $\pi_i(a_i|o_i)$. It is updated by the gradient of the expected return $J(\theta^{\pi_i})$, represented as:

$$\begin{aligned}\nabla_{\theta^{\pi_i}} J(\theta^{\pi_i}) =&\mathbb{E}_{s\sim\rho^\pi,a_i\sim\pi_i,\boldsymbol{w}_i=g_i(o_i)}[\nabla_{\theta^{\pi_i}}\log\pi_i(a_i\mid o_i)\\ &\boldsymbol{w}_i^\top \boldsymbol{Q}_i^{\pi_i}(s,a_1,\ldots,a_i,\ldots,a_N,\boldsymbol{W}\mid\theta^{\boldsymbol{Q}_i})[i]].\end{aligned} \tag{8}$$

This framework is extended to deterministic policies. The policy is reformulated as the continuous action version: $\boldsymbol{\mu} = \{\mu_1(o_1\mid\theta^{\mu_1}),\ldots,\mu_N(o_N\mid\theta^{\mu_N})\}$, which is updated by the MAMODDPG:

$$\begin{aligned}\nabla_{\theta^{\mu_i}} J(\theta^{\mu_i}) = \mathbb{E}_{s,a\sim\mathcal{D},\boldsymbol{w}_i=g_i(o_i)}[\nabla_{a_i}\boldsymbol{w}_i^\top \boldsymbol{Q}_i^{\mu_i}(s,a_1,\ldots,\\ a_i,\ldots,a_N,\boldsymbol{W}\mid\theta^{\boldsymbol{Q}_i})\mid_{a_i=\mu_i(o_i\mid\theta^{\mu_i})}\nabla_{\theta^{\mu_i}}\mu_i(o_i\mid\theta^{\mu_i})].\end{aligned} \tag{9}$$

According to Theorem 2, the structural modeling of global preference and policy is necessary. Thus, for the critic network, let $h^S = \{h_1^S,\ldots,h_N^S\}$ denote the set of feature embeddings from N agents, where each feature embedding h_i^S combines the observation $o_i \in \mathbb{R}^{D_o}$, agent-specific actions a_i , and local preference $\boldsymbol{w_i}$: $h_i^S = [o_i;a_i;\boldsymbol{w}_i]$. Each embedding is mapped into a common embedding space of

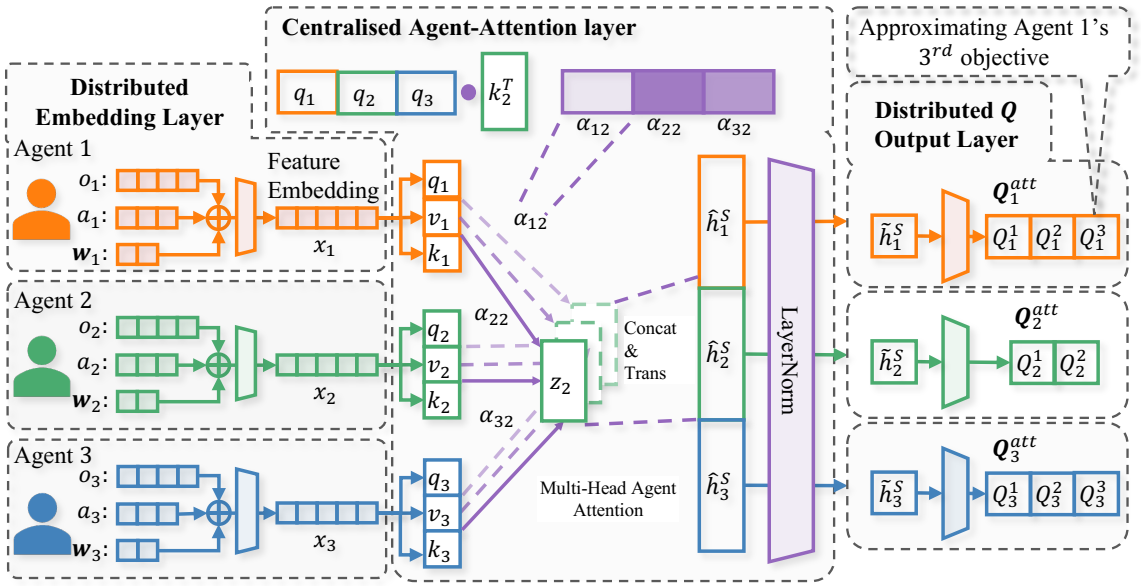


Figure 2: Agent Attention-based MAMORL Framework. Each agent uses its individual embedding layer to extract feature embeddings x_i based on o_i, a_i, and $\boldsymbol{w}_i$. All agents' embeddings are concatenated and fed into a centralised agent-attention layer. In this layer, each agent's preference and policy are accessible to others, enabling agents to model the influence of others' preferences and associated policies on their own policies and rewards. The output of this layer, $\hat{H}^S$, is sliced to obtain the feature corresponding to agent i: $\tilde{h}_i^S$ through the LayerNorm layer, which is then passed to agent i's Q output layer to produce a vectorised Q-value $\mathbf{Q}_i^{att}$ that approximates its multiple rewards.

dimension d through each agent's linear encoder layer, forming $\mathbf{x}_i \in \mathbb{R}^d$ corresponding to i-th agent's latent feature.

To effectively capture inter-agent influence under a given state, the global preference that is inferable, and the associated joint policy, we employ an agent-level attention mechanism. Unlike traditional attention applied to language or vision (Han et al. 2024), where tokens or patches are arranged with spatial or sequential priors, here each $\mathbf{x}_i$ represents an independent and potentially heterogeneous agent's state, action, and preference. This attention module is shared across all agents. It serves as an explicit relational reasoning mechanism, enabling each agent to dynamically incorporate inter-agent influences based on learnt preference-specific and policy-specific patterns, thereby modeling agent-aware relational dependencies in a fully differentiable manner.

To capture the relational dependencies between agent i and all agents, the model transforms the agent's embedding via projection matrices into query, key, and value vectors:

$$\mathbf{q}_i = \mathbf{x}_i \mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V, \tag{10}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_h}$ are shared across agents within a given attention head, $\boldsymbol{X}$ is the collective embedding of all N agents. Each agent thus uses its own query vector $\mathbf{q}_i$ to compute attention weights over all N key vectors:

$$\boldsymbol{\alpha}_i = \text{softmax}\left(\frac{\mathbf{q}_i \mathbf{K}^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{1 \times N}, \tag{11}$$

which reflects how much agent i's utility and policy attend to other agents when updating its representation. i's resulting embedding under a single attention head is computed as:

$$\mathbf{z}_i = \boldsymbol{\alpha}_i \mathbf{V} = \sum_{j=1}^{N} \alpha_{ij} \mathbf{v}_j. \tag{12}$$

The interaction weights α_{ij} adaptively quantify how much agent j's utility function and associated policy affect agent i's rewards, which is critical in maintaining the joint policy stationary in MAMOSs.

To enrich the model's expressiveness, multiple attention heads are used in parallel. Each head independently projects the inputs using distinct parameters, producing head-specific embeddings $\mathbf{z}_i^{(1)}, \ldots, \mathbf{z}_i^{(h)}$. These are concatenated and linearly transformed to obtain the final output embedding:

$$\hat{h}^S[i] = Concat\left(\mathbf{z}_i^{(1)}, \ldots, \mathbf{z}_i^{(h)}\right) \mathbf{W}^O, \quad \mathbf{W}^O \in \mathbb{R}^{hd_h \times d}. \tag{13}$$

All agent's final embedding of different heads will be concatenated: $\hat{H}^S = Concat\left(\hat{h}_1^S, \ldots, \hat{h}_N^S\right)$. Then a feed-forward network (FFN) with residual connections and layer normalisation further enhances representational capacity:

$$\tilde{h}^S = \text{LayerNorm}(\hat{h}^S + \text{FFN}(\hat{h}^S)). \tag{14}$$

Finally, the slice of each agent $\tilde{h}_i^S$ is processed by the individual output layer to estimate agent-specific Q-values, enabling accurate approximation of vectorised action values under complex, multi-agent dynamics for agent i:

$$\boldsymbol{Q}_i^{att} = Output(\tilde{h}_i^S). \tag{15}$$

And the attention-based MOTD error becomes:

$$L^{att}(\theta^{\boldsymbol{Q}_i}) = \mathbb{E}_{s,s',\boldsymbol{a},\boldsymbol{W},\boldsymbol{r}\sim\mathcal{D}}\left[\boldsymbol{Q}_i^{att}(s,A,\boldsymbol{W}) - \boldsymbol{y}_i^{att}\right]^2. \tag{16}$$

$$\boldsymbol{y}_i^{att} = \boldsymbol{r}_i + \gamma \boldsymbol{Q}_i^{'att}(s',A',W') \mid_{a'_j=\mu'_j(o'_j),a_i^{GPI}=\mu_i^{GPI}(o'_i)}. \tag{17}$$

By minimising the attention-based MOTD error, the parameters of each agent's embedding and output layers, as well as the shared agent-attention module, are jointly updated. This enables each agent to better learn how variations in its own policy affect the vectorised rewards under the global utility functions and associated joint policies. Such relational modeling facilitates utility-aware policy improvement, guiding agents toward both global utility maximisation and convergence to a BNE. The pseudocode of these two learning frameworks can be found in (Li et al. 2025).

Experiment Results and Analysis

Experiment Settings

Datasets

- **MOMA particle environments**: A series of environments extended from the grounded particle environment (Lowe et al. 2017) into an MO version. The first objective aligns with the original one, the second one is the energy consumption related to movement and communication. The benchmark includes the following scenarios: *Push*, *Adversary*, *Reference*, *Spread*, and *Tag*.
- **MOMALand** (Felten et al. 2024): A benchmark that builds on the PettingZoo API and supports MAMO learning by returning vector-valued rewards. It includes diverse scenarios such as *Mountain Walker*, *Escort*, *Catch*, and *Surround*.

  Detailed descriptions of these environments can be found in (Li et al. 2025).

Baselines

- **MOMIX** (Hu et al. 2023): Utilises preference-conditioned local action-value estimation and a parallel mixing network to compute joint value functions. A preference-based exploration mechanism is introduced to encourage well-distributed Pareto-optimal solutions.
- **GPI-PD** (Alegre et al. 2023): Combines GPI with a Dyna-style MORL approach to prioritise updates for improved sample efficiency. Modifications are made to support the multi-agent setting.
- **Individual Preference (IP)**: Each agent learns its policy based solely on its local observation and private preference vector. This can be viewed as a part of ablation tests.
- **MADDPG** (Zhang et al. 2024) : A standard single-objective MADDPG baseline with scalarised rewards computed from multiple objectives using current preferences. This can be viewed as a part of ablation tests.

Evaluation Metrics

- **Global Utility (GU)** (Alegre et al. 2023): Multiple objectives are weighted summed by the preference $\boldsymbol{w}_i$ to achieve the individual utility $v_i^{\pi_i}(\boldsymbol{w}_i)$ for each agent. All agent's individual utility will be averages to get the GU. We average over 128 initial states for diverse preference settings to approximate the preference space.
- **Hypervolume (HV)** (Zitzler, Brockhoff, and Thiele 2007): The volume of the area in the objective space enclosed by the reference points and the non-dominated solutions obtained by the algorithm.

Performance Comparison

In this experiment, each agent's preference is modeled as a linear function of the observation. The mapping function is agent-specific but kept consistent across all rounds and baselines to ensure fairness. Hyperparameter settings are presented in (Li et al. 2025).

Table 1 presents the performance across 10 training seeds, reported as the mean and standard deviation of GU and HV. Agent-Attention MAMORL (AA) consistently achieves the best and most stable performance across most environments. Although global-preference-based MAMORL (GP) performs comparably to AA in certain environments (*Walker*, *Push*, and *Reference*), its lack of consensus among agents undermines its robustness in settings with conflicting individual preferences (*Catch*). GPI-PD is designed for single-agent MO settings. As a result, it shows poor performance and occasionally fails to converge. This demonstrates the limitation of single-agent frameworks in modeling the multi-agent policy-preference space. MOMIX is designed for team preference settings and a discrete action space. It lacks generalisability for individual rewards and preferences. Consequently, MO-MIX shows its shortcomings in generalisation to complex reward and preference settings.

Learning curves in Fig. 3 show that in the most challenging environment, *Multi-Walker*, only AA and GP successfully acquire meaningful policies, whereas all other baselines fail to progress. Similar trends are observed in the MOMA particle environments, where AA and GP demonstrate the most stable learning dynamics, while others struggle with convergence and exhibit high variance.

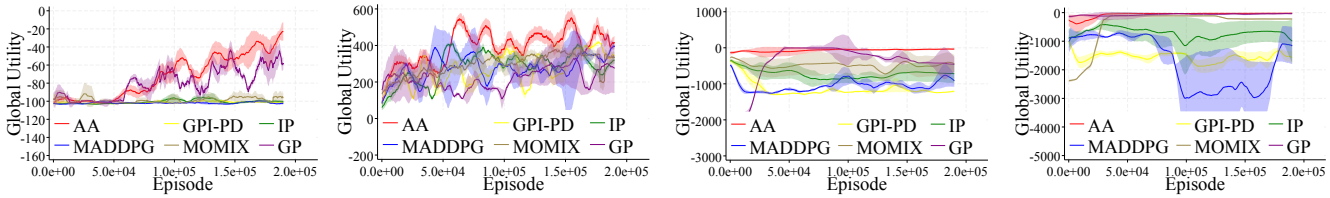
Ablation Study

The comparison among AA, GP, MADDPG, and IP serves as the ablation study aimed at evaluating the impact of vectorised action, GP, and AA mechanisms on MAMO learning.

The struggles of MADDPG highlight the importance of vectorised representations of rewards, action-values, and preferences in multi-objective settings. Scalarised rewards might be effective when preferences are fixed and aligned between training and execution, but they become impractical in real-world scenarios where preferences evolve dynamically. As shown in our dynamic preference setting, scalarised reward methods struggle to generalise and fail to extract meaningful policies. The poor performance of IP provides empirical support for Theorem 1. Without direct access or structured modeling to the global preference,

Env	Metric	GPI- PD	MOMIX	IP	MADDPG	AA	GP
Catch	GU	315.2 ± 3.9	82.0 ± 2.1	161.0 ± 2.4	397.8 ± 2.7	$\mathbf{528.1 \pm 26.7}$	255.9 ± 91.4
	HV	251.6 ± 43.0	2.8 ± 4.0	93.5 ± 3.9	124.1 ± 15.7	$\mathbf{215.8 \pm 22.9}$	78.4 ± 22.7
Escort	GU	269.0 ± 66.3	67.6 ± 2.9	290.0 ± 3.4	316.0 ± 7.4	$\mathbf{613.8 \pm 90.8}$	569.9 ± 0.6
	HV	385.1 ± 43.8	12.9 ± 3.4	137.6 ± 11.7	46.4 ± 7.7	$\mathbf{184.8 \pm 75.0}$	133.8 ± 37.2
Walker	GU	-103.3 ± 0.2	-99.3 ± 0.4	-102.4 ± 0.0	-100.9 ± 0.1	$\mathbf{-25.8 \pm 6.8}$	-31.4 ± 21.1
	HV	12.6 ± 6.6	22.0 ± 4.3	15.9 ± 3.0	19.9 ± 9.0	$\mathbf{3345.7 \pm 475.6}$	2115.7 ± 511.6
Sur	GU	405.2 ± 1.1	254.6 ± 19.6	225.7 ± 33.8	405.4 ± 0.3	$\mathbf{615.7 \pm 58.9}$	436.1 ± 1.2
	HV	70.9 ± 13.9	143.6 ± 24.7	88.1 ± 29.6	261.2 ± 13.9	$\mathbf{318.4 \pm 72.0}$	96.3 ± 85.5
Adv	GU	-155.1 ± 11.7	-61.2 ± 11.7	-160.8 ± 34.3	-150.0 ± 16.4	$\mathbf{-26.9 \pm 2.4}$	-64.5 ± 6.2
	HV	276.9 ± 29.2	622.0 ± 80.7	379.7 ± 88.4	201.0 ± 36.1	$\mathbf{963.0 \pm 43.4}$	707.5 ± 77.6
Push	GU	-83.4 ± 3.5	-44.3 ± 33.5	-216.6 ± 54.2	-202.1 ± 19.2	$\mathbf{-21.1 \pm 1.9}$	-40.6 ± 15.7
	HV	726.5 ± 101.7	658.3 ± 69.7	848.0 ± 57.6	741.1 ± 79.0	$\mathbf{2183.1 \pm 469.9}$	1382.6 ± 87.3
Ref	GU	-99.5 ± 22.8	-66.9 ± 22.8	-109.1 ± 15.1	-265.7 ± 431.6	$\mathbf{-51.1 \pm 2.2}$	-67.7 ± 9.7
	HV	564.2 ± 135.5	763.0 ± 43.9	542.6 ± 58.4	488.4 ± 78.2	749.8 ± 91.1	$\mathbf{784.2 \pm 114.8}$
Spread	GU	-73.8 ± 1.2	-168.6 ± 41.2	-130.7 ± 25.9	-198.0 ± 42.8	$\mathbf{-53.8 \pm 0.6}$	-128.3 ± 38.7
	HV	143.2 ± 195.5	278.3 ± 25.1	175.0 ± 12.4	673.1 ± 84.6	$\mathbf{846.0 \pm 27.6}$	384.2 ± 59.3
Tag	GU	-116.4 ± 4.0	-31.7 ± 7.0	-145.2 ± 33.7	-105.1 ± 17.2	$\mathbf{-15.0 \pm 1.5}$	-57.8 ± 15.4
	HV	176.7 ± 67.5	358.9 ± 24.2	289.4 ± 28.6	243.9 ± 537.4	$\mathbf{521.8 \pm 56.2}$	295.1 ± 28.6

Table 1: Performance comparison in 9 MAMO environments. Results are reported as mean ± standard deviation over 10 seeds.



(a) Mountain Walker (b) Surround (c) Tag (d) Push

Figure 3: Average and 95% confidence intervals of GU from AA, GP, and baselines on 4 MAMO environments.

agents cannot form accurate beliefs over others' policies, making BNE unattainable. Consequently, learning becomes unstable and ineffective. AA even outperforms GP in several environments. This result demonstrates that how heterogeneous and dynamic utilities influence inter-agent coordination. By explicitly modeling these relational factors, AA enables each agent to update its own policy, adapting to other policies more effectively, resulting in improved performance and convergence towards the BNE.

Conclusion

In this paper, we mathematically prove that direct access to or structured modeling of global preferences during decision-making is a necessary condition for achieving BNE in MAMOS. For the case where preferences are randomly generated, we incorporate global preferences into distributed decision-making and design a corresponding MAMORL framework. For the more realistic setting where preferences are generated by agents based on their own observations, we develop an AA-MAMORL framework, where a centralised attention-based critic network is employed to model inter-agent influences and preference-policy dependencies and is shared among all agents. To evaluate AA and global preference, we constructed experiments using 9 standard MAMO environments. The results demonstrate that the proposed AA-MAMORL consistently outperforms baselines across diverse environments by effectively modeling heterogeneous preferences and coordinating decentralised policies. Ablation results highlight the necessity of global preference modeling and vectorised objectives for stable and convergent learning in multi-agent multi-objective settings.

Acknowledgments

This work was supported in part by EU Horizon 2020 Grant No. 101008297, UKRI Grant No. EP/Y036786/1, and Horizon EU Grant No. 101129910.

References

Alegre, L. N.; Bazzan, A. L.; Roijers, D. M.; Nowé, A.; and da Silva, B. C. 2023. Sample-efficient multi-objective learning via generalized policy improvement prioritization. *arXiv preprint arXiv:2301.07784*.

Assos, A.; Dagan, Y.; and Daskalakis, C. 2024. Maximizing utility in multi-agent environments by anticipating the behavior of other learners. *Advances in Neural Information Processing Systems*, 37: 38769–38798.

Chang, L.; Iqbal, S.; and Chen, H. 2023. Does financial inclusion index and energy performance index co-move? *Energy Policy*, 174: 113422.

Felten, F.; Ucak, U.; Azmani, H.; Peng, G.; Röpke, W.; Baier, H.; Mannion, P.; Roijers, D. M.; Terry, J. K.; Talbi, E.-G.; et al. 2024. Momaland: A set of benchmarks for multi-objective multi-agent reinforcement learning. *arXiv preprint arXiv:2407.16312*.

Han, D.; Wang, Z.; Xia, Z.; Han, Y.; Pu, Y.; Ge, C.; Song, J.; Song, S.; Zheng, B.; and Huang, G. 2024. Demystify mamba in vision: A linear attention perspective. *Advances in neural information processing systems*, 37: 127181–127203.

Hayat, S.; Yanmaz, E.; Bettstetter, C.; and Brown, T. X. 2020. Multi-objective drone path planning for search and rescue with quality-of-service requirements. *Autonomous Robots*, 44(7): 1183–1198.

He, W.; Xu, W.; Ge, X.; Han, Q.-L.; Du, W.; and Qian, F. 2021. Secure control of multiagent systems against malicious attacks: A brief survey. *IEEE Transactions on Industrial Informatics*, 18(6): 3595–3608.

Hu, T.; Luo, B.; Yang, C.; and Huang, T. 2023. MO-MIX: Multi-objective multi-agent cooperative decision-making with deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12098–12112.

Jiang, W.; Li, P.; Sha, A.; Li, Y.; Yuan, D.; Xiao, J.; and Xing, C. 2023. Research on pavement traffic load state perception based on the piezoelectric effect. *IEEE Transactions on Intelligent Transportation Systems*, 24(8): 8264–8278.

Li, M.; Wang, Q.; and Xu, Y. 2025. Gtde: Grouped training with decentralized execution for multi-agent actor-critic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18368–18376.

Li, Z.; Luo, C.; Huang, L.; Qi, L.; and Min, G. 2025. Achieving Equilibrium under Utility Heterogeneity: An Agent-Attention Framework for Multi-Agent Multi-Objective Reinforcement Learning. arXiv:2511.08926.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.

Mallareddy, M.; Thirumalaikumar, R.; Balasubramanian, P.; Naseeruddin, R.; Nithya, N.; Mariadoss, A.; Eazhilkrishna, N.; Choudhary, A. K.; Deiveegan, M.; Subramanian, E.; et al. 2023. Maximizing water use efficiency in rice farming: A comprehensive review of innovative irrigation management technologies. *Water*, 15(10): 1802.

Niu, H.; Lin, Z.; An, K.; Wang, J.; Zheng, G.; Al-Dhahir, N.; and Wong, K.-K. 2023. Active RIS assisted rate-splitting multiple access network: Spectral and energy efficiency tradeoff. *IEEE Journal on Selected Areas in communications*, 41(5): 1452–1467.

Paine, T. M.; and Benjamin, M. R. 2024. A model for multi-agent autonomy that uses opinion dynamics and multi-objective behavior optimization. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 8305–8311. IEEE.

Qi, Q.; Chen, X.; Khalili, A.; Zhong, C.; Zhang, Z.; and Ng, D. W. K. 2022. Integrating sensing, computing, and communication in 6G wireless networks: Design and optimization. *IEEE Transactions on Communications*, 70(9): 6212–6227.

Rădulescu, R.; Mannion, P.; Roijers, D. M.; and Nowé, A. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems*, 34(1): 10.

Saglam, I. 2025. Bayesian Nash Equilibrium. In *Mastering Game Theory: A Comprehensive Introduction to Strategic Decision Making*, 101–112. Springer.

Shi, M.; He, H.; Li, J.; Han, M.; and Jia, C. 2021. Multi-objective tradeoff optimization of predictive adaptive cruising control for autonomous electric buses: A cyber-physical-energy system approach. *Applied Energy*, 300: 117385.

Wei, Z.; Yu, X.; Ng, D. W. K.; and Schober, R. 2021. Resource allocation for simultaneous wireless information and power transfer systems: A tutorial overview. *Proceedings of the IEEE*, 110(1): 127–149.

Wong, A.; Bäck, T.; Kononova, A. V.; and Plaat, A. 2023. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, 56(6): 5023–5056.

Yang, R.; Sun, X.; and Narasimhan, K. 2019. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32.

Zhang, C.; Sun, G.; Li, J.; Wu, Q.; Wang, J.; Niyato, D.; and Liu, Y. 2024. Multi-objective aerial collaborative secure communication optimization via generative diffusion model-enabled deep reinforcement learning. *IEEE Transactions on Mobile Computing*.

Zitzler, E.; Brockhoff, D.; and Thiele, L. 2007. The hypervolume indicator revisited: On the design of Pareto-compliant indicators via weighted integration. In *Evolutionary Multi-Criterion Optimization: 4th International Conference, EMO 2007, Matsushima, Japan, March 5-8, 2007. Proceedings 4*, 862–876. Springer.