A Survey of Multi-Agent Offline Policy Reinforcement Learning: Decentralized Learning and Cooperative Policy Optimization

Abstract

Multi-agent offline policy reinforcement learning (MARL) represents a significant advancement in artificial intelligence, enabling agents to optimize decision-making processes in environments where real-time data acquisition is challenging. This survey explores the integration of decentralized learning and cooperative policy optimization within MARL, highlighting its transformative potential across various domains such as robotics, healthcare, and energy management. The paper examines the core principles and challenges of MARL, including policy coordination, scalability, sample inefficiency, and distributional shifts. It emphasizes offline policy optimization's role in leveraging pre-collected datasets to enhance exploration and efficiency in MARL systems. The survey also discusses decentralized learning frameworks, communication strategies, and the pivotal role of cooperative agents in optimizing policies. Applications in real-world domains, such as microgrid management and autonomous systems, demonstrate MARL's capacity to improve system performance and stability. Future research directions are identified, focusing on scalability, robustness, and the integration of advanced models and techniques. The survey concludes by underscoring the need for innovative methodologies to fully realize MARL's potential, paving the way for more adaptive and intelligent multi-agent systems. By addressing these challenges, MARL can significantly enhance artificial intelligence's capabilities in complex, dynamic environments.

1 Introduction

1.1 Overview of Multi-Agent Reinforcement Learning (MARL)

Multi-Agent Reinforcement Learning (MARL) is a vital area in artificial intelligence that involves multiple agents learning and interacting within shared environments, characterized by complex cooperative and competitive dynamics [1]. A primary challenge in MARL is coordinating policies among agents, often necessitating advanced strategies to manage interactions without direct communication, as traditional methods typically depend on explicit inter-agent communication [2].

Scalability in MARL systems poses significant concerns; the complexity of agent interactions in dynamic environments can lead to considerable computational bottlenecks during training [3]. This complexity is intensified by the non-stationarity of the environment, requiring agents to adapt continuously to the evolving policies of others, thus necessitating robust strategies for stability and convergence.

Sample inefficiency is another challenge, particularly in model-free approaches that demand extensive data to learn effective policies. This issue is exacerbated in partially observable environments, where joint action spaces increase the curse of dimensionality, complicating scalability. Therefore, decentralized learning and cooperative strategies are essential for managing communication and coordination intricacies among agents, enabling efficient optimization of joint policies [1].

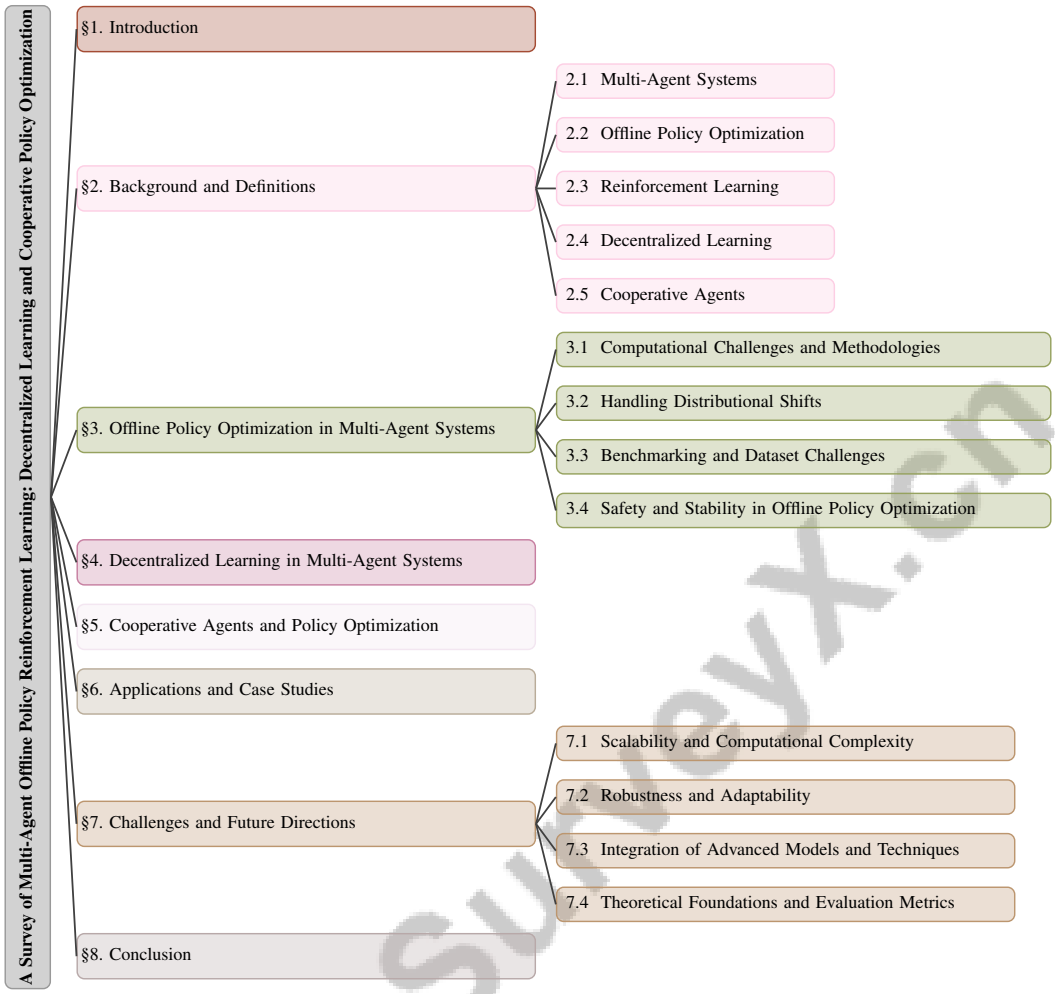


Figure 1: chapter structure

Practically, MARL has been effectively applied to optimize complex systems in various domains, including energy management and multi-robot systems, showcasing its potential to enhance real-world efficiency and performance. As MARL evolves, addressing emerging challenges through innovative strategies that improve communication, skill generalization, and knowledge transfer among agents is crucial. Recent research emphasizes developing frameworks for independent agents to share information effectively, as demonstrated by methods for fully independent communication and parallel knowledge transfer. Hybrid training approaches, such as HyGen, illustrate the potential for improved multi-task generalization by integrating online and offline learning, which is essential for realizing MARL's capabilities in intricate, dynamic environments [4, 5, 6, 7, 1].

1.2 Significance of Offline Policy Optimization

Offline policy optimization is crucial for advancing multi-agent reinforcement learning (MARL), particularly in settings where real-time data acquisition is impractical or hazardous. This method proves advantageous in domains like fleet management, where fluctuating demand and supply conditions require robust policy optimization without real-time data reliance [8]. By leveraging pre-collected datasets, offline policy optimization facilitates scalable solutions that enhance exploration through diverse strategies, thereby improving MARL systems' adaptability and efficiency [9].

A significant challenge in offline policy optimization is addressing the distributional shift between the behavior policy used for data collection and the target policy being optimized. These shifts can induce overgeneralization and scaling issues, as agents must navigate complex interactions and heterogeneous observation representations without real-time feedback. The intricacies of multiple

agents' interactions further complicate this process, necessitating efficient coordination strategies to manage the environment's non-stationarity and evolving policies [10].

Managing interactions in decentralized environments entails maximizing collective performance in collaborative tasks, often requiring the development of decentralized policies that effectively handle conflicting incentives among agents. The unpredictability of agent interactions and the stochastic nature of the environment highlight the need for advanced exploration methods to optimize policies effectively [11].

Existing methods often rely on expert datasets or uniform coverage datasets, which may not always be feasible, leading to struggles with distribution shifts during online fine-tuning and difficulties in retaining useful behaviors from offline training [12]. The inefficiency of single-agent reinforcement learning in multi-agent environments underscores the necessity for robust offline policy optimization strategies [13].

In energy management contexts, offline optimization significantly enhances efficiency without necessitating costly sensors or human expertise, showcasing its potential to improve system performance across various domains [13]. Effectively allocating tasks and controlling robots in scenarios with varying object and robot numbers complicates decision-making and adaptability [14]. Furthermore, the absence of theoretical guarantees in existing MARL methods can lead to unstable and suboptimal performance [15]. Addressing these challenges is essential for unlocking offline policy optimization's full potential and achieving significant advancements in artificial intelligence.

1.3 Structure of the Survey

This survey provides a comprehensive examination of Multi-Agent Offline Policy Reinforcement Learning, focusing on decentralized learning and cooperative policy optimization. The paper is organized into several key sections, each addressing specific aspects of the topic.

The introductory section presents an overview of Multi-Agent Reinforcement Learning (MARL), emphasizing core principles and challenges in policy coordination among agents, alongside the significance of offline policy optimization in contexts where real-time data is impractical or risky.

The second section delves into background and definitions, clarifying essential concepts such as multi-agent systems, offline policy optimization, reinforcement learning, decentralized learning, and cooperative agents. This foundational knowledge is crucial for understanding subsequent discussions.

The third section focuses on offline policy optimization in multi-agent systems, exploring challenges and methodologies for leveraging offline data to optimize policies without real-time interactions. It discusses computational challenges, strategies for managing distributional shifts, and the importance of safety and stability in offline policy optimization.

The fourth section examines decentralized learning in multi-agent systems, addressing the benefits and challenges of decentralized approaches, including scalability and communication among agents. It reviews frameworks and algorithms that support decentralized learning and explores strategies for agent communication and coordination.

The fifth section highlights cooperative agents and policy optimization, exploring various strategies for optimizing policies in cooperative settings and reviewing frameworks that facilitate cooperative learning while examining cooperation's impact on learning outcomes.

The sixth section illustrates practical applications and case studies of multi-agent offline policy reinforcement learning in domains such as robotics, healthcare, energy management, and transportation systems, demonstrating the potential of MARL techniques to enhance efficiency and performance across sectors.

The seventh section identifies current challenges in the field and discusses potential future research directions, emphasizing areas needing further investigation, such as scalability, robustness, and the integration of advanced models and techniques. This section underscores the importance of developing theoretical foundations and evaluation metrics to advance the field.

In conclusion, the paper synthesizes crucial insights, emphasizing multi-agent offline policy reinforcement learning (MARL) as a transformative approach in artificial intelligence. It highlights challenges in optimizing multi-agent policies from pre-collected datasets, such as the complexity of joint state-

action spaces and the risk of uncoordinated behaviors. Proposed methodologies, including In-Sample Sequential Policy Optimization (InSPO) and the Offline Value Function Memory with Sequential Exploration (OVMSE), address these challenges while demonstrating significant improvements in policy coordination and sample efficiency. The findings underscore the potential of offline MARL techniques to enhance collaborative agent systems' performance, paving the way for advancements in various AI applications [16, 17, 18, 19, 20].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Multi-Agent Systems

Multi-agent systems (MAS) comprise multiple agents interacting to achieve shared goals, each with distinct observations and information, posing challenges in convergence and efficiency. These systems underpin Multi-Agent Reinforcement Learning (MARL), where agents coordinate actions towards common objectives. The decentralized decision-making in MAS complicates optimization, as agents rely on local data, leading to complex dynamics [21]. Algorithm scalability is crucial in MARL due to the exponential growth of joint action spaces, necessitating efficient libraries for diverse tasks. MAS address decentralized decision-making challenges, especially when agents have limited global state visibility, making negotiation and communication essential in partially observable environments [22, 23, 24, 25, 26].

In heterogeneous environments, traditional message-passing may be inadequate, and as agent numbers increase, credit assignment for individual actions becomes complex, resulting in noisy policy gradients and learning inefficiencies. This complexity is exacerbated by varying dataset qualities, highlighting the importance of optimality inductive biases in offline reinforcement learning. Approaches like partial reward decoupling (PRD) simplify credit assignment by breaking tasks into subproblems, enhancing data efficiency and learning stability. Coordination as a learning objective can optimize agent policies, indicating its importance for MAS advancement [27, 28, 10]. High sample complexity in cooperative MARL further exacerbates learning inefficiencies.

MAS dynamics can be cooperative or competitive, requiring agents to adapt strategies based on other agents' behaviors. In cooperative scenarios, agents align individual goals with shared objectives, using methods like Altruistic Gradient Adjustment (AgA) to converge interests towards stable solutions. In competitive settings, agents balance personal ambitions with group goals, employing sophisticated tactics to navigate mixed-motive interactions, as seen in social dilemmas and control problems. Addressing these challenges is crucial for enhancing MAS capabilities in MARL, facilitating effective decentralized and cooperative policy optimization [29, 30, 31, 32].

2.2 Offline Policy Optimization

Offline policy optimization in MARL uses pre-collected datasets to develop robust policies, particularly where real-time data is impractical or risky, such as in autonomous driving. A key challenge is managing distributional shifts between behavior and target policies, leading to extrapolation errors [33]. Techniques like Reward Gap Minimization (RGM) correct imperfect rewards, enabling robust policy optimization [34], crucial in stochastic scenarios where agents lack direct gradient access [35]. Offline policy optimization fosters cooperation by allowing informed decision-making without real-time data, utilizing structured environmental representations.

It supports strategy development without real-time data, addressing dynamic challenges [8]. Techniques like policy finetuning bridge offline and online learning, enhancing sample efficiency [36], while -satisficing rules maintain adaptability in dynamic environments [37]. In non-cooperative systems, offline optimization is vital for accurate agent interaction modeling [38]. Addressing extrapolation errors and limited state-action space coverage is crucial for offline optimization's potential [39]. User-friendly software development facilitates advanced MARL research, overcoming existing frameworks' complexity [40]. Techniques like AMRS enhance learning in cooperative MARL environments [41], with accurate credit assignment using methods like PRD-AC advancing offline MARL [28].

Frameworks like TIHDP optimize task priority through hierarchical processing [14], while Heterogeneous-Agent Mirror Learning (HAML) ensures theoretical guarantees for improvement and

convergence [15]. These underscore offline policy optimization's role in robust, scalable multi-agent systems, advancing MARL by enabling efficient strategy refinement across complex applications [9].

2.3 Reinforcement Learning

Reinforcement Learning (RL) is a foundational AI paradigm where agents learn to make decisions to maximize cumulative rewards over time [42]. Agents take actions in states, receive rewards, and use feedback to enhance decision-making, learning optimal strategies for expected returns. The learning process is modeled as a Markov Decision Process (MDP), comprising states, actions, rewards, and transitions, defining environmental dynamics. The agent's objective is to identify an optimal policy mapping states to actions that maximize expected returns, often leveraging pre-existing teacher policies within a multi-objective framework, enhancing learning efficiency in continuous spaces. In competitive and cooperative settings, agents consider others' actions, employing game-theoretic principles for strategy formulation [37, 32, 42, 43, 44].

A major RL challenge, especially in MARL, is policy transferability across tasks and agents. Benchmarks evaluate policy effectiveness from offline datasets, assessing generalization to new tasks and configurations [17]. This is crucial for robust agents adapting to diverse environments without extensive retraining. The exploration-exploitation trade-off is fundamental, balancing exploration for better strategies against exploiting known high-reward actions. Effective management ensures RL algorithms converge to optimal policies. Techniques like Reward Gap Minimization address imperfect rewards, while uniform convergence methods provide optimal error bounds for policies. Policy finetuning enhances sample efficiency, leveraging reference policies for improved convergence in offline and online settings [45, 34, 36].

Reinforcement learning offers a robust framework for autonomous decision-making, effective across applications like robotics, games, and autonomous vehicles. Recent offline RL advancements, like expert-supervised frameworks, enhance applicability by extracting insights from offline data, enabling optimal policy formulation while quantifying uncertainty and accommodating risk levels [46, 47]. Its capacity to learn from interactions and adapt to dynamic environments makes it a powerful tool for intelligent systems in complex scenarios.

2.4 Decentralized Learning

Decentralized learning is vital in MARL, allowing agents to learn and decide based on local information without central authority, beneficial where centralized communication is unreliable. Agents solve local optimization problems, reducing communication overhead and enhancing scalability [21]. It facilitates diverse strategy exploration, constrained by parameter sharing leading to homogeneous behaviors [1]. However, quadratic communication complexity can hinder performance and scalability [2].

Despite challenges, decentralized learning supports cooperative exploration, enhancing collaboration and learning outcomes. The lack of global consensus requires reliance on local information, complicating joint action coordination. This can hinder credit assignment and slow learning, particularly in complex environments. Leveraging simplified dynamics through scalable approaches, like local reward structures in cooperative MARL, enhances coordination and performance. Frameworks enabling local policy learning without communication help reconstruct optimal solutions, addressing partial observability and coordination challenges [22, 48, 49]. As agent numbers grow, communication complexity increases, complicating action space exploration and timely feedback.

To mitigate challenges, decentralized learning incorporates asynchronous actions and control mechanisms, allowing independent task performance while achieving coordinated outcomes. This balances individual and joint cost minimization, crucial where agents lack access to each other's states and actions. Decentralized learning enhances robust, scalable solutions by enabling autonomous learning while effectively sharing essential information, beneficial in dynamic environments. Techniques like Centralized Training for Decentralized Execution (CTDE) and frameworks like LToS optimize cooperation, achieving global objectives with limited communication. Knowledge transfer and meta-learning architectures improve adaptability and performance in multi-agent systems [50, 51, 6, 48, 52].

2.5 Cooperative Agents

Cooperative agents in MARL optimize collective outcomes in environments requiring collaboration for shared objectives. They manage interdependencies and enhance learning, especially in complex environments with concurrent events [53]. The Parallel Attentional Transfer (PAT) method exemplifies cooperative agents, using student and self-learning modes to optimize learning, showcasing cooperative strategies' adaptability and efficiency [6].

In scenarios where individual rationality leads to suboptimal collective actions, like social dilemmas, cooperative agents enhance cooperation, mitigating irrational behaviors [29]. Optimal control laws for linear systems manipulated by multiple agents highlight cooperative agents' role in precise control and coordination [54]. Aligning individual objectives with collective goals is challenging, as discrepancies can lead to conflicts and inefficiencies [31]. The credit assignment problem complicates this, as agents must identify their contributions to a global reward, crucial for effective cooperation [55].

Training novice agents to cooperate with varying skill levels maximizes team rewards and facilitates learning [56]. Overfitting policies due to biased advantage values pose challenges, necessitating robust policy regularization for cooperation efficiency [57]. Cooperative agents must navigate unexpected failures, like hardware issues, disrupting coordination and causing system crashes [58]. Benchmarks comparing algorithms provide insights into cooperative strategies' effectiveness and efficiency, aiding robust cooperative MARL system development [44].

Cooperative agents optimize multi-agent systems' performance by aligning actions with shared objectives, improving learning through exploration and credit assignment, ensuring coordination in dynamic environments. This is crucial in scenarios like autonomous driving, where agents navigate conflicting goals, and in social dilemmas where individual incentives conflict with collective benefits. Recent advancements introduce innovative approaches like the CM3 framework, addressing individual and cooperative goal learning challenges through structured curricula and localized credit assignment, leading to faster learning and improved performance [30, 1].

In the context of multi-agent systems, the optimization of offline policies presents a myriad of computational challenges that necessitate a nuanced understanding of various methodologies. As illustrated in Figure 2, the hierarchical structure of offline policy optimization encompasses several primary categories, each detailing specific challenges and corresponding solutions. This figure not only highlights the intricacies involved in handling distributional shifts and benchmarking but also addresses dataset challenges, as well as essential safety and stability concerns. By providing a comprehensive overview of the field, the figure serves as a valuable reference for understanding the multifaceted nature of offline policy optimization in multi-agent systems.

3 Offline Policy Optimization in Multi-Agent Systems

3.1 Computational Challenges and Methodologies

Offline policy optimization in multi-agent systems involves significant computational challenges due to decentralized decision-making and dynamic agent interactions. The complexity of managing the expansive joint action space in Multi-Agent Reinforcement Learning (MARL) leads to inefficiencies with traditional centralized training methods [2]. The non-stationarity from concurrent learning by multiple agents adds to the difficulty, as agents face noisy, partial observations that impede optimal policy learning [59].

Identifying optimal decentralized policies requires navigating a double-exponential search space, resulting in inefficient learning and high sample demands. Balancing exploration and exploitation remains challenging, as existing methods often fail to achieve optimal policy performance during online training. Innovative approaches like the Multi-Agent Multi-Environment Mixed Q-Learning (MEMQ) algorithm address these challenges by allowing agents to minimize costs independently while coordinating through a leader agent [21].

Techniques such as Optimistic Search Lambda (OS()) and Advantage-Weighted Policy Optimization (AWPO) within frameworks like MAZero enhance search efficiency and learning in large action spaces [60]. The Federated Control with Reinforcement Learning (FCRL) framework combines hierarchical and multi-agent deep RL approaches, utilizing a meta-controller for agent communication

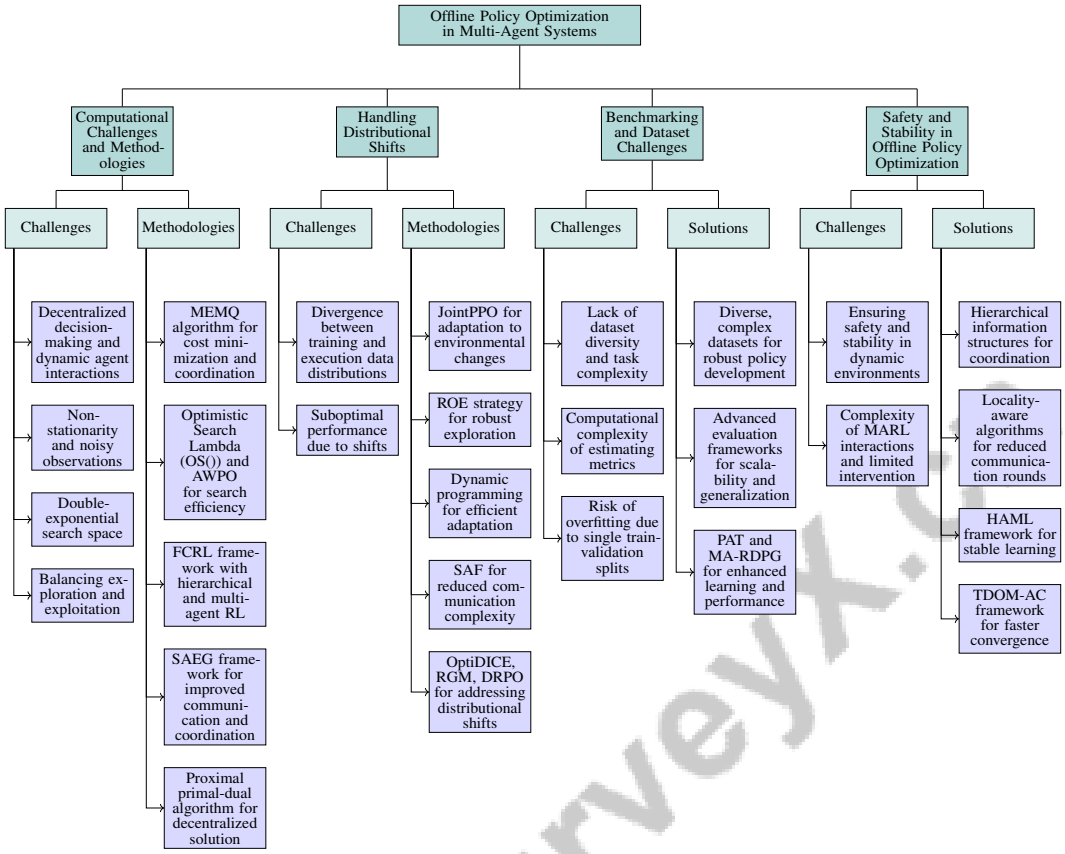


Figure 2: This figure illustrates the hierarchical structure of offline policy optimization in multi-agent systems, highlighting computational challenges and methodologies, handling distributional shifts, benchmarking and dataset challenges, and safety and stability concerns. Each primary category is further divided into specific challenges and corresponding solutions or methodologies, providing a comprehensive overview of the field.

[61]. The Shared Agent-Entity Graph (SAEG) framework improves communication and coordination by modeling agents and environmental entities as graph vertices [2].

Despite advancements, computational bottlenecks persist, particularly as agent numbers increase. Specific training phases consume excessive time, highlighting the need for scalable solutions [3]. The proximal primal-dual algorithm offers a decentralized solution by reformulating the problem into a saddle-point problem, enhancing computational efficiency [1]. Integrating sophisticated communication strategies, hierarchical frameworks, and adaptive exploration techniques is essential for overcoming decentralized decision-making challenges and achieving robust, scalable solutions in multi-agent systems [62, 18, 63, 64].

As illustrated in Figure 3, this figure categorizes the key computational challenges and emerging methodologies in MARL. It showcases the intricate processes and analytical techniques in optimizing agent performance in complex environments. The flowchart in the first example integrates a simulator and Shapley analysis, systematically evaluating agent performance. The second example compares Optimal Policy Evaluation (OPE) estimates and true rewards, providing insights into policy evaluation accuracy. The final example analyzes Q-values and normalized returns in AntMaze Large Play and Pen Sparse environments, emphasizing performance variance due to environmental conditions and learning parameters. These examples highlight the computational challenges and methodological advancements in offline policy optimization [23, 65, 64].

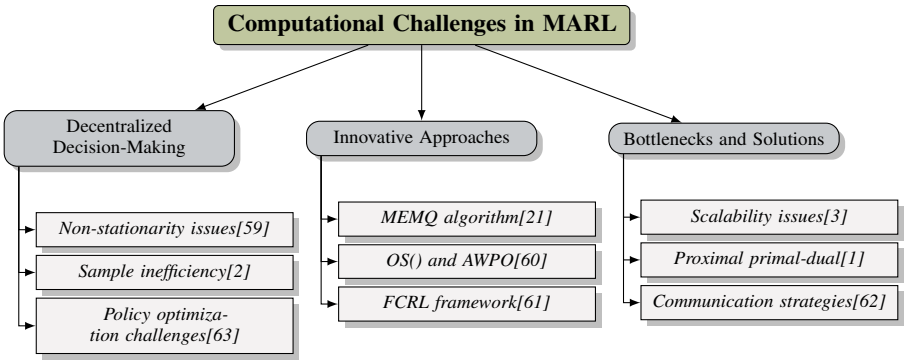


Figure 3: This figure illustrates the key computational challenges and emerging methodologies in Multi-Agent Reinforcement Learning (MARL). It categorizes the challenges into decentralized decision-making issues, innovative approaches addressing these challenges, and existing bottlenecks with potential solutions.

Method Name	Adaptation Strategies	Coordination Techniques	Optimization Methods
JPPO[39]	Sequence Generation Task	Centralized Critic Approximation	Proximal Policy Optimization
ROE[66]	Risk Scheduling	Aligned Risk Levels	Risk-based Approach
IA2C++[67]	Action Configurations	Action Anonymity Integration	Dynamic Programming Approach
SAF[2]	Faster Learning Adaptation	Intelligent Facilitator Mediation	Shared Pool Policies

Table 1: Comparison of various methods addressing distributional shifts in multi-agent reinforcement learning, highlighting their adaptation strategies, coordination techniques, and optimization methods. The table includes JointPPO, ROE, IA2C++, and SAF, each employing unique approaches to optimize policy performance amidst distributional changes.

3.2 Handling Distributional Shifts

Addressing distributional shifts in offline data is crucial for effective policy optimization in multi-agent reinforcement learning (MARL). These shifts occur when training data distributions diverge from those encountered during policy execution, risking suboptimal performance. JointPPO reformulates decision-making into a sequence generation task, facilitating adaptation to environmental changes [39]. The Risk-based Optimistic Exploration (ROE) strategy encourages optimistic exploration, maintaining robust strategies despite distributional shifts [66]. He et al.'s dynamic programming approach computes distributions over action configurations, enabling efficient adaptation [67].

The Structured Agent Framework (SAF) reduces communication complexity and enhances coordination, crucial for mitigating distributional shift impacts [2]. Employing methodologies like JointPPO, OptiDICE, RGM, and DRPO, combined with dynamic programming, effectively addresses distributional shifts. JointPPO optimizes joint policies, OptiDICE corrects overestimation in offline learning, RGM manages imperfect rewards, and DRPO enables federated learning without further interactions, navigating intrinsic distributional shifts [68, 63, 39, 44, 34]. Table 1 presents a comparative analysis of methodologies employed to handle distributional shifts in multi-agent reinforcement learning, detailing their respective adaptation strategies, coordination techniques, and optimization methods.

3.3 Benchmarking and Dataset Challenges

Benchmarking and dataset challenges are pivotal for developing robust, generalizable policies in offline multi-agent systems. Table 2 provides a detailed overview of the representative benchmarks used in offline multi-agent reinforcement learning, highlighting the diversity in dataset size, domain, task format, and evaluation metrics. Current benchmarks often lack sufficient dataset diversity and task complexity, hindering agent generalization evaluation [17]. Datasets, like those from StarCraft II, simulate varying task complexities but may not cover all real-world scenarios [69]. Estimating metrics such as the Fisher information matrix is computationally complex, limiting scalability [71]. Single train-validation splits risk overfitting, missing optimal algorithm-hyperparameter pairs due to sparse high-reward trajectories in small datasets [65].

Benchmark	Size	Domain	Task Format	Metric
MADT[17]	4,754,000	Multi-Agent Reinforcement Learning	Policy Learning	Sample Efficiency, Generalization
MADT[69]	4,178,846	Multi-Agent Reinforcement Learning	Policy Learning	Average Return
OB[70]	1,000,000	Multi-Agent Reinforcement Learning	Policy Optimization	Variance, Reward
MARL-Bench[3]	1,000,000	Multi-Agent Reinforcement Learning	Cooperative Navigation	Q loss, P loss

Table 2: This table presents a comparative analysis of various benchmarks used in multi-agent reinforcement learning, detailing their size, domain, task format, and evaluation metrics. The benchmarks include MADT, OB, and MARL-Bench, which are evaluated on metrics such as sample efficiency, generalization, average return, variance, reward, Q loss, and P loss. These benchmarks serve as critical tools for assessing the scalability, efficiency, and generalization capabilities of multi-agent systems.

Restrictive model architectures struggle with input-output variability, impeding knowledge transfer [7]. Experiments with 17 algorithms across 23 tasks from environments like SMAC, MPE, GRF, MAMuJoCo, and MAgent underscore the need for comprehensive benchmarking frameworks [60]. Continuous cooperative/competitive tasks, such as predator-prey scenarios, model complex agent interactions [10].

Addressing these challenges requires diverse, complex datasets and advanced evaluation frameworks. These should assess scalability, efficiency, and generalization capabilities, fostering robust policy development for diverse environments. Innovative approaches like Parallel Attentional Transfer (PAT) for knowledge sharing and Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG) for collaborative decision-making enhance team learning and performance. MARLlib streamlines multi-agent task implementation, while decentralized algorithms enable efficient policy evaluation and convergence in complex interactions [60, 23, 6, 24, 52].

3.4 Safety and Stability in Offline Policy Optimization

Safety and stability are critical in offline policy optimization for cooperative MARL, where agents operate in dynamic, uncertain environments. Ensuring these aspects requires strategies to mitigate risks of overestimation and unsafe actions. Hierarchical information structures enhance coordination and stabilize decentralized learning processes [59]. The complexity of MARL interactions and limited managerial intervention pose stability challenges as agent numbers grow, leading to expansive state and action spaces. The Heterogeneous-Agent Mirror Learning (HAML) framework manages heterogeneous agents, ensuring stable learning across configurations [15].

Locality-aware algorithms enhance safety by reducing communication rounds, minimizing instability risks from excessive information exchange [21]. The Multi-Agent Actor-Critic Time Dynamical (TDOM-AC) framework contributes to stable learning with faster convergence and improved opponent behavior prediction [61]. Addressing scalability bottlenecks is crucial for effective MARL application, impacting learning stability and safety [3].

A comprehensive strategy integrates hierarchical structures, locality-aware algorithms, and robust benchmarking techniques for safety and stability in offline policy optimization. This includes safe evaluation frameworks like approximate high-confidence off-policy evaluation (HCOPE) for performance estimation before deployment, federated reinforcement learning methods like DRPO for distributional shift management, and offline-to-online reinforcement learning techniques for balanced exploitation of pre-trained and online policies [72, 63, 18]. These strategies are essential for reliable performance in dynamic, uncertain multi-agent environments, advancing MARL.

4 Decentralized Learning in Multi-Agent Systems

Decentralized learning in multi-agent systems is underpinned by frameworks and algorithms that enable agents to function autonomously while optimizing collaboration and decision-making. The following subsection explores key frameworks and algorithms that enhance performance and scalability in these systems.

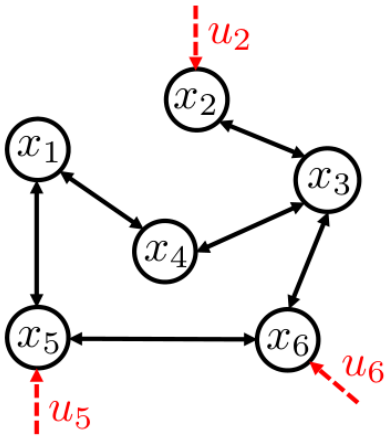
4.1 Decentralized Learning Frameworks and Algorithms

Decentralized learning frameworks are crucial for the scalability and robustness of multi-agent systems, allowing agents to operate without centralized control. The Shared Agent-Entity Graph (SAEG) framework facilitates agent communication via a learned facilitator, optimizing collective performance in cooperative tasks [2]. The Multi-Agent Multi-Environment Mixed Q-Learning (MEMQ) algorithm exemplifies decentralized learning by enabling agents to independently optimize policies while sharing limited information, balancing independent and collaborative approaches to enhance performance in offline reinforcement learning [27, 24, 18, 73].

Frameworks such as Federated Control with Reinforcement Learning (FCRL) integrate hierarchical and multi-agent deep reinforcement learning to address coordination challenges through meta-controllers [2]. Incorporating causality into MARL allows agents to leverage causal relationships for informed decision-making, addressing coordination and exploration challenges [74, 75, 76, 1, 77].

The Multi-Agent Actor-Attention-Critic (MAAC) model supports decentralized learning by enhancing agents' learning from limited data through attention mechanisms [22, 78, 79, 24, 52]. The diffusion GTD algorithm empowers agents to collaboratively estimate value functions through localized interactions, improving exploration efficiency and policy evaluation [22, 78, 79, 24, 52].

These frameworks and algorithms enhance multi-agent systems by integrating localized policies, innovative exploration strategies, and advanced architectures. The LOMAQ algorithm utilizes local rewards within a Centralized Training Decentralized Execution paradigm to address credit assignment challenges, while the PAT framework facilitates parallel knowledge transfer, boosting team learning rates and performance [22, 6].



(a) A Directed Graph with Node Labels and Arrows[49]

$$\sum_{t=1}^{T}(\mu_{1,1}-\mu_{a_t,b_t})$$
$$\le \sum_{t=1}^{T}\left(\frac{1}{n_t(1)^+}\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau=1]\mu_{1,b_\tau}+\sqrt{\frac{\kappa B\log(T/\delta)}{n_t(1)^+}}-\mu_{a_t,b_t}\right) \quad \text{(by Assumption 1 with } a=1\text{)}$$
$$\le \sum_{t=1}^{T}\left(\hat{\mu}_t(1)+c\sqrt{\frac{\log(T/\delta)}{n_t(1)^+}}+\sqrt{\frac{\kappa B\log(T/\delta)}{n_t(1)^+}}-\mu_{a_t,b_t}\right) \quad \text{(by Lemma 10)}$$
$$\le \sum_{t=1}^{T}\left(\hat{\mu}_t(a_t)+c\sqrt{\frac{\log(T/\delta)}{n_t(a_t)^+}}+\sqrt{\frac{\kappa B\log(T/\delta)}{n_t(a_t)^+}}-\mu_{a_t,b_t}\right) \quad \text{(by the selection rule of } a_t\text{)}$$
$$=\sum_{t=1}^{T}(\hat{\mu}_t(a_t)-\mu_{a_t,b_t})+\mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right)$$
$$=\underbrace{\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t=a]\left(\hat{\mu}_t(a)-\mu_{a,1}\right)}_{\textbf{term}_1}+\underbrace{\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t=a]\left(\mu_{a,1}-\mu_{a_t,b_t}\right)}_{\textbf{term}_2}+\mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right)$$

Notice that $\textbf{term}_2=\mathcal{O}\left(\sum_{a\in[A]}\sqrt{n_{T+1}(a)\log(T/\delta)}\right)=\mathcal{O}\left(\sqrt{ABT\log(T/\delta)}\right)$ due to Assumption 1. Besides, by Azuma's inequality, with probability at least $1-\mathcal{O}(\delta)$, for all t and a,

$$\hat{\mu}_t(a)=\frac{\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau=a]r_\tau}{n_t(a)^+}\le\frac{\sum_{\tau=1}^{t-1}\mathbb{I}[a_\tau=a]\mu_{a,b_\tau}}{n_t(a)^+}+\mathcal{O}\left(\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right)\le\mu_{a,1}+\mathcal{O}\left(\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right).$$

Therefore, $\textbf{term}_1\le\mathcal{O}\left(\sum_{a\in[A]}\sum_{t=1}^{T}\mathbb{I}[a_t=a]\sqrt{\frac{\log(T/\delta)}{n_t(a)^+}}\right)=\mathcal{O}\left(\sqrt{AT\log(T/\delta)}\right)$. Combining everything finishes the proof. ■

For MAB, there are also algorithms with refined gap-dependent regret bounds with only $\mathcal{O}(\log(T))$ dependence on T (e.g., the UCB1 algorithm of Auer et al. (2002a)). Below we show that if U2 executes such algorithms, the overall regret can also be of order $\mathcal{O}(\log(T))$. Such algorithms satisfy the following assumption:

Assumption 2 *U2 guarantees the following for some universal constant $\kappa\ge1$ with probability at least $1-\delta$:*

$$\forall t,a,\quad \sum_{\tau=1}^{t}\mathbb{I}[a_\tau=a](\mu_{a,1}-\mu_{a,b_\tau})\le\min\left\{\kappa\sum_{b\in\mathcal{B}_0}\frac{\log(T/\delta)}{\mu_{a,1}-\mu_{a,b}},\ \sqrt{\kappa B\sum_{\tau=1}^{t}\mathbb{I}[a_\tau=a]\log(T/\delta)}\right\},$$

(b) The image shows a mathematical proof involving summation and inequality calculations.[59]

Figure 4: Examples of Decentralized Learning Frameworks and Algorithms

Figure 4 illustrates the structural and theoretical components of decentralized learning in multi-agent systems, showcasing both practical and analytical dimensions [49, 59].

4.2 Communication and Coordination Strategies

In decentralized MARL systems, effective communication and coordination are essential. NeurComm enhances information sharing among agents, improving learning efficiency in non-stationary environments by dynamically adjusting communication strategies [80]. Shared agent-entity graphs enable agents to learn optimal communication patterns, optimizing interactions and coordination through shared insights [2, 51, 81, 24].

Attention mechanisms further refine coordination by focusing on relevant information, reducing noise, and enhancing decision-making capabilities in complex scenarios [27, 6, 26, 82]. These strategies promote effective information sharing and strategic coordination, enabling agents to tackle complex tasks collaboratively [24, 83].

4.3 Challenges in Decentralized Learning

Decentralized learning in multi-agent systems faces challenges such as maintaining policy stability while allowing exploration. The absence of a centralized coordinator complicates synchronization, leading to suboptimal decision-making. Sophisticated communication protocols and coordination strategies are needed to address communication delays and information asymmetry [62, 2, 24, 26, 84].

Scalability is another challenge, as increasing agent numbers lead to computational bottlenecks. Strategies like parallel knowledge transfer and locality-based reward decomposition enhance scalability and convergence speed [22, 52, 6].

Innovative solutions to these challenges improve adaptability, scalability, and robustness, enabling agents to optimize policies in dynamic environments. Approaches like Parallel Attentional Transfer (PAT) and Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG) enhance collaboration and performance [24, 52, 6].

4.4 Safety and Robustness in Decentralized Learning

Safety and robustness are critical in decentralized MARL systems. Ensuring optimal policy learning without compromising safety is crucial in dynamic environments. Methods that prevent unsafe actions while allowing effective learning are vital for system integrity [85]. Robustness is enhanced through adaptability to system crashes, with coach-assisted learning improving coordination and performance stability [58].

Safety-oriented methods compatible with various MARL algorithms ensure improved learning outcomes [12]. Addressing safety and robustness requires a comprehensive approach combining preventive measures, adaptability, and algorithmic compatibility.

4.5 Decentralization and Privacy Concerns

Decentralized learning raises privacy concerns as agents operate independently. While decentralization limits data exposure, it challenges information sufficiency for decision-making [86, 48, 87, 49]. Information leakage during interactions is a risk, requiring cryptographic techniques and privacy-preserving protocols to secure communication channels.

The absence of centralized authority complicates privacy policy enforcement and data usage monitoring. Robust mechanisms are needed to autonomously enforce privacy constraints and ensure compliance with regulations [48, 63, 49, 88]. Techniques like differential privacy and secure multi-party computation enhance privacy protection while maintaining learning efficiency.

Addressing privacy concerns requires balancing information sharing with data protection. Advanced frameworks and algorithms enable collaborative learning while safeguarding privacy [49, 63, 86, 48, 87]. Implementing privacy-preserving techniques ensures effective collaboration and decision-making among agents.

5 Cooperative Agents and Policy Optimization

In multi-agent reinforcement learning (MARL), policy optimization is deeply intertwined with agent cooperation, enhancing learning outcomes and aligning individual objectives with shared goals. This section delves into the pivotal role of cooperative agents in policy optimization, emphasizing their interactions and strategies that facilitate decision-making in complex environments.

5.1 Role of Cooperative Agents in Policy Optimization

Cooperative agents are crucial for optimizing policies in MARL settings, employing both shared and independent strategies to elevate system performance. Their collaboration enables the exchange of

learned features, enhancing model accuracy and control [13]. This synergy is vital for leveraging collective knowledge to improve decision-making.

In the MA-RDPG model, cooperation maximizes performance, underscoring collaboration's importance in policy optimization [24]. By aligning objectives, agents optimize policies in coordinated environments. Hierarchical decision-making structures further empower agents by reducing direct coordination needs, allowing independent decisions that contribute to collective goals [59]. This enhances policy optimization robustness and autonomy within a shared framework.

The MA-DAAC model's shared attention mechanisms facilitate policy optimization by focusing on relevant information, thus enhancing action coordination [89]. This is essential in data-intensive contexts where timely decisions are crucial. Cooperative agents also collaboratively evaluate target policies while following distinct behavior policies, highlighting their role in policy optimization [79]. This collaborative learning improves prediction accuracy and policy development.

Frameworks like MARLlib demonstrate cooperative agents' benefits by offering scalable solutions for diverse tasks [60]. Their user-friendly interfaces support cooperative strategies, enhancing policy optimization.

Cooperative agents are integral to policy optimization in MARL, fostering enhanced learning through strategic collaboration, communication, and optimization techniques. Their ability to leverage shared information promotes efficient and robust policy development, significantly advancing multi-agent systems [1].

5.2 Strategies for Policy Optimization in Cooperative Settings

Optimizing policies in cooperative MARL environments requires strategic agent collaboration to achieve collective objectives. Distributed gradient descent methods enable collaborative learning by sharing gradient information and updating policies in parallel, accelerating convergence to optimal policies [32].

Decentralized policy gradient methods allow agents to optimize independently using local data, reducing environmental interaction reliance. This fosters collaboration among agents with distinct datasets while aligning with shared goals through dual regularization, balancing local and global policy considerations. These methods help agents navigate distributional shifts and achieve policy improvements in a federated context, enhancing decision-making [49, 63, 48, 70, 52]. Local updates reduce communication overhead and improve scalability for large-scale systems.

Hierarchical reinforcement learning frameworks decompose complex tasks into sub-tasks, simplifying learning and enhancing performance. By promoting collaboration at hierarchical levels, these frameworks improve policy optimization efficiency through inter-agent communication techniques [10, 2, 28, 24].

Attention mechanisms significantly enhance coordination by allowing agents to focus on relevant information, improving communication clarity and efficiency. This focus enables agents to prioritize essential data, facilitating well-informed decisions crucial for policy optimization in dynamic environments [27, 23]. Such approaches enhance performance across varying dataset qualities, enabling nuanced understanding within cooperative systems.

Strategies for policy optimization in cooperative settings emphasize collaboration, decentralized learning benefits, and hierarchical task decomposition effectiveness. Techniques like gradient-based distributed policy search are advantageous in partially observable environments, while methods addressing credit assignment and federated offline policy optimization ensure robust policy improvements [28, 24, 32, 63]. Leveraging these strategies, agents can effectively optimize policies, improving performance and robustness in multi-agent systems.

5.3 Frameworks and Architectures Supporting Cooperative Learning

Frameworks and architectures promoting cooperative learning in MARL enhance agent interaction scalability and efficiency. Hierarchical architectures decompose complex tasks into sub-tasks, simplifying learning and improving coordination [59]. This enables agents to focus on specific task components, optimizing policies efficiently.

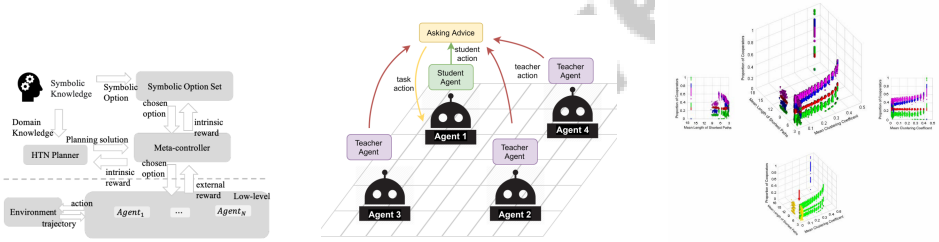
The Shared Agent-Entity Graph (SAEG) framework exemplifies cooperative learning architecture by enhancing communication through graph representation of agents and environmental entities [2]. This fosters cooperative behavior learning through effective information sharing.

Frameworks like MARLlib provide scalable solutions for implementing cooperative strategies across tasks [60]. Their user-friendly interfaces support cooperative learning integration, enhancing policy optimization in complex environments.

The Multi-Agent Actor-Attention-Critic (MAAC) model incorporates attention mechanisms to improve coordination, allowing agents to focus on relevant information, enhancing decision-making and collaboration [89].

Integrating causality into MARL frameworks boosts cooperation and performance, enabling informed decisions based on causal relationships among actions and rewards. This addresses challenges like lazy agent pathology and promotes intelligent behavior, improving coordination through causal influence [74, 4, 75, 76, 77]. This integration is crucial for developing effective cooperative learning outcomes and robust policy optimization.

Research into cooperative learning frameworks and architectures advances multi-agent systems, including curriculum learning and memory-based meta-learning architectures. Effective training strategies must consider agent collaboration and team reward optimization while adapting to varying skill levels. Findings suggest curricula with decreasing skill levels enhance performance, while shared communication policies foster rapid adaptation to new environments [51, 56]. Hierarchical structures, attention mechanisms, and graph-based representations enhance scalability, efficiency, and robustness, paving the way for sophisticated multi-agent systems.



(a) Symbolic Option-Based Meta-Controller for Multi-Agent Systems[5]

(b) Teacher-Agent-Based Learning Environment[6]

(c) Cooperator Proportion in a Network[83]

Figure 5: Examples of Frameworks and Architectures Supporting Cooperative Learning

As illustrated in Figure 5, developing frameworks and architectures supporting cooperative learning is essential for advancing multi-agent systems. The examples demonstrate various approaches to facilitating agent cooperation and learning. The "Symbolic Option-Based Meta-Controller for Multi-Agent Systems" showcases a framework where symbolic knowledge guides decision-making, emphasizing symbolic reasoning integration for enhanced interactions. The "Teacher-Agent-Based Learning Environment" presents a structured setting for dynamic information exchange, highlighting mentorship's importance in cooperative learning. Lastly, the "Cooperator Proportion in a Network" visually explores how network characteristics influence cooperative behavior prevalence. These examples underscore diverse methodologies employed to cultivate cooperation and optimize policies in multi-agent systems, paving the way for more sophisticated collaborative frameworks [5, 6, 83].

6 Applications and Case Studies

The transformative potential of multi-agent offline policy reinforcement learning (MARL) spans various domains, enhancing decision-making and operational efficiency. This section examines case studies that underscore MARL's impact in robotics, healthcare, education, energy management, traffic systems, and gaming, highlighting its ability to optimize complex environments.

6.1 Robotics and Autonomous Systems

MARL significantly advances decision-making in robotics and autonomous systems, especially where real-time data acquisition is limited. Techniques like In-Sample Sequential Policy Optimization (InSPO) and OMIGA enhance agent coordination by addressing out-of-distribution joint actions and premature convergence [16, 90]. Applications in autonomous driving and robotic control benefit from offline policy optimization, fostering robust strategies and mitigating real-time experimentation risks. PowerNet exemplifies MARL's role in stabilizing microgrid output voltages, optimizing control strategies using offline data for improved stability and efficiency [91]. Real-robot demonstrations validate MARL's effectiveness, emphasizing task prioritization and hierarchical decision-making for enhanced operations [14]. Datasets from environments like StarCraft II further assist in training robust systems capable of generalization across contexts [17]. By integrating offline and online learning, MARL systems adapt efficiently to new tasks, facilitating quicker deployment and enhanced real-world application generalization [18, 47, 64, 92].

6.2 Healthcare and Education

In healthcare, MARL enhances decision-making by developing personalized treatment plans from pre-collected patient data, reducing real-time experimentation risks and improving patient outcomes [27, 24, 65]. MARL optimizes healthcare resource allocation, such as staff scheduling and patient flow management, by simulating professional interactions to determine optimal workflow strategies, thus enhancing care quality. Advancements in hierarchical planning and knowledge transfer further adapt these systems to dynamic healthcare needs [3, 4, 5, 6, 1]. In education, MARL tailors learning experiences by integrating online and offline strategies, optimizing training processes for task adaptability. Reward shaping addresses sparse feedback challenges, leading to personalized strategies [7, 82]. MARL supports intelligent tutoring systems, providing real-time feedback for refined teaching strategies based on extensive datasets [72, 82]. MARL's application in healthcare and education improves decision-making efficiency and personalization, ensuring safe policy evaluations and enhancing outcomes [72, 65, 64].

6.3 Energy and Microgrid Management

MARL optimizes control and coordination in energy and microgrid management. The MADDPG-GCPN algorithm exemplifies effective cooperative behavior in managing microgrid energy storage systems [93]. Decentralized control allows agents to learn optimal strategies from local states and collaborative communication, enhancing renewable energy management and system resilience [13, 91]. MARL frameworks optimize energy resource scheduling and dispatch, reducing operational costs and enhancing sustainability through simulations that align economic and environmental objectives [27, 24, 43, 13]. MARL's integration in energy management highlights its transformative potential, facilitating cooperative control among distributed generators to optimize performance and adaptability in dynamic demand-supply conditions [13, 91, 8].

6.4 Traffic and Transportation Systems

MARL optimizes traffic flow and transportation network efficiency. Agents representing vehicles, traffic signals, and road users use offline data to formulate effective policies for improved system coordination. Combining pre-trained offline policies with adaptive online strategies enhances data efficiency and real-time performance [62, 18]. Traffic signal control optimization is a key advantage, where agents learn optimal timing sequences from historical data, reducing congestion and improving network performance. MARL systems optimize route planning and vehicle dispatching, minimizing travel time and fuel consumption, especially in congested urban environments [24, 8]. MARL in intelligent transportation systems (ITS) enhances responsiveness to traffic conditions, facilitating dynamic management solutions that improve decision-making amid fluctuating demands [94, 8]. This application supports sustainable urban mobility solutions through enhanced traffic management efficiency [18, 65, 8, 64].

6.5 Networked and Distributed Systems

MARL optimizes coordination and control in networked and distributed systems, benefiting from decentralized decision-making, enhancing communication efficiency, and reducing sample complexity [95]. In power networks, MARL coordinates distributed energy resources for grid stability, crucial with intermittent renewable sources, devising strategies to minimize losses and enhance sustainability [13, 91]. In autonomous driving networks, MARL enhances traffic flow and safety by coordinating vehicle actions based on pre-collected data, optimizing lane changing and speed adjustments [8, 82]. MARL's integration in networked systems optimizes real-time decision-making while utilizing pre-trained offline policies, enhancing efficiency, reliability, and scalability for advanced infrastructures in IoT and mobile edge computing scenarios [18, 96].

6.6 Game and Simulation Environments

MARL in game and simulation environments provides a platform for evaluating agent cooperation and strategy development. These settings simulate complex interactions, refining multi-agent strategies in various scenarios. The Simple Team Sports Simulator (STS2) captures team sports dynamics, evaluating cooperative agent behaviors [97]. Agents optimize decision-making and coordination using Imitation Learning and MARL, improving performance in complex scenarios. Embedding contextual information into reward functions addresses sparse reward challenges, facilitating efficient learning and complementary skill emergence [98, 97, 82]. Game environments test novel reinforcement learning algorithms, exploring cooperative and competitive strategies for enhanced multi-agent learning frameworks. Techniques like multi-scenario ranking and knowledge transfer improve team performance and learning efficiency, while advanced reward shaping addresses sparse reward challenges [82, 6, 24, 99, 83]. MARL's implementation in these environments advances artificial intelligence, developing optimal multi-agent policies from pre-collected datasets. Frameworks like the multi-agent decision transformer (MADT) and algorithms such as InSPO enhance learning efficiency and coordination, improving task generalization and sample efficiency in offline and online settings [16, 17]. These advancements contribute to effective and adaptable learning models, enhancing multi-agent systems' capabilities in real-world applications.

7 Challenges and Future Directions

7.1 Scalability and Computational Complexity

Scalability and computational complexity are significant hurdles in multi-agent reinforcement learning (MARL), particularly as the agent count and environmental complexity increase. The exponential growth of joint action-observation spaces results in inefficiencies and bottlenecks in large-scale systems [1]. Sparse rewards and the substantial computational demands of extensive state-action spaces further exacerbate these challenges, requiring considerable resources [1]. MARL frameworks like MARLlib demonstrate high memory usage, highlighting the need for efficient algorithms that function within computational limits [60]. The complexity of high-dimensional action spaces necessitates innovative methodologies to address these challenges without sacrificing performance [1].

Scalability is also hindered by reliance on predefined communication structures, which may not adapt to dynamic environments with shifting agent configurations [2]. While the SAF architecture shows promise, its scalability in complex scenarios needs further exploration [2]. Flexible communication protocols are essential to accommodate varying bandwidth constraints and agent dynamics. Centralized training methods in benchmark frameworks reveal limitations in fully decentralized settings, often failing to address scenarios demanding decentralized approaches [3]. Future research should enhance MARL methodologies' adaptability and explore sophisticated coordination mechanisms to improve scalability.

Developing methodologies that enhance credit assignment accuracy and agent cooperation is crucial for multi-agent task performance. Overcoming scalability and computational complexity challenges will enable MARL systems to be effectively deployed in intricate environments. This advancement paves the way for sophisticated applications, such as enhancing cooperative strategies in sparse reward environments through frameworks like SOMARL, which integrate hierarchical task planning and symbolic knowledge [60, 3, 5].

7.2 Robustness and Adaptability

Robustness and adaptability are crucial in MARL, as agents must effectively navigate dynamic and unpredictable environments. Future research should refine exploration strategies and enhance robustness across diverse multi-agent contexts [19]. This involves developing sophisticated mechanisms that enable agents to maintain high performance in varied scenarios, even under adversarial conditions or rapidly changing dynamics. Robustness does not always equate to optimal performance, particularly in dynamic environments, necessitating continuous refinement and adaptation [100]. Balancing robustness with adaptability requires ongoing innovation and advanced learning techniques.

Inaccurate cost estimations can hinder decision-making and adaptability in dynamic settings [73]. Reliable estimation techniques and adaptive frameworks are needed to adjust to varying conditions without compromising performance. Moreover, agents' ability to adapt to rapidly changing opponent strategies is crucial for maintaining robustness in competitive environments. Methods that struggle with evolving strategies may find it challenging to sustain robust performance [101]. Addressing these challenges involves enhancing algorithm flexibility and incorporating mechanisms for real-time adaptation.

A multifaceted approach integrating advanced exploration strategies, reliable estimation techniques, and adaptive learning frameworks is necessary for robustness and adaptability in MARL. By addressing scalability bottlenecks, enhancing multi-task generalization, and improving communication strategies, MARL systems can increase resilience and effectiveness, adapting efficiently to diverse environments [5, 3, 7, 4].

7.3 Integration of Advanced Models and Techniques

Integrating advanced models and techniques in MARL is promising for enhancing performance and scalability in complex environments. Refining algorithms to address optimality gaps and extending them to continuous state and action spaces is essential for operation in dynamic settings [102]. Incorporating models like Large Language Models into MARL frameworks, especially in preference-based learning scenarios, can enhance decision-making by providing richer contextual understanding [103]. Future development of frameworks like PowerNet should focus on integrating advanced models to optimize performance in complex dynamics and disturbances [91].

Refining algorithms for scalability and exploring applications across domains are pivotal for advancing credit assignment in complex environments [55]. This exploration will contribute to developing robust and efficient MARL systems for diverse tasks. Reducing computational overhead and enhancing scalability are critical research areas addressing current limitations in MARL frameworks [94]. Novel algorithms adaptable in resource-constrained environments are needed. Extending techniques like Independent Causal Learning (ICL) to centralized learning scenarios and real-world problems requiring online learning will enhance MARL systems' applicability [76]. Dynamic adjustments of key parameters through a parameter controller can significantly enhance MARL methods' adaptability [18].

Integrating advanced models and techniques in MARL, such as the Offline Pre-trained Multi-Agent Decision Transformer (MADT) and hybrid training frameworks like HyGen, is vital for addressing challenges. These innovations enhance sample efficiency and generalization across tasks, enabling agents to learn optimal policies from offline datasets and adapt to complex applications. Approaches facilitating independent communication among agents and hierarchical task planning contribute to more effective collaboration and performance, unlocking MARL systems' full potential [4, 17, 5, 69, 7].

7.4 Theoretical Foundations and Evaluation Metrics

Advancing MARL necessitates robust theoretical foundations and comprehensive evaluation metrics to facilitate scalable solutions. Enhancing model adaptability and robustness, particularly in multi-agent offline policy reinforcement learning, is critical for optimizing performance [59]. Future research should expand support for hyperparameter tuning and enhance off-policy capabilities, crucial for improving MARL frameworks' scalability and adaptability [60]. Exploring the balance between diversity and sharing in multi-agent systems is vital for improving theoretical foundations and

evaluation metrics, leading to more effective strategies for optimizing collaborative interactions among agents [13].

Developing efficient approximation methods for gradient computation and sampling techniques is necessary to optimize training processes in multi-agent settings [89]. Integrating function approximation techniques and extending methods to infinite state spaces can enhance MARL frameworks' scalability and applicability [79]. Insights from advanced learning environments can provide valuable understanding of cooperative task dynamics and zero-incentive scenarios, contributing to MARL's theoretical advancement [61].

Automated learning of efficient factorization schemes is another focus area aimed at enhancing scalability and applicability of factored shielding approaches in MARL [15]. Exploring Proximal Policy Optimization (PPO) in competitive multi-agent environments and heterogeneous agents can validate its effectiveness and contribute to robust evaluation metrics [3]. Integrating large language models into MARL frameworks, particularly in hybrid training environments, offers significant potential for enhancing efficiency and adaptability in diverse applications [60]. Future work should apply scalable approaches to mixed cooperative-competitive domains, validating their effectiveness for broader use cases [21].

The progress of MARL depends on the ongoing enhancement of its theoretical frameworks and evaluation metrics. These advancements are crucial for addressing scalability challenges and ensuring effective deployment of multi-agent systems across diverse applications, such as optimizing communication strategies among agents, reducing sample and communication complexities during policy evaluation, and facilitating cooperative learning in complex environments. By refining these foundational elements, researchers can better tackle scalability bottlenecks and improve MARL algorithms' overall performance in practical scenarios [3, 4, 5, 6, 95].

8 Conclusion

The exploration of multi-agent offline policy reinforcement learning (MARL) has underscored its transformative potential in advancing artificial intelligence by optimizing agent interactions in environments where real-time data acquisition is challenging or risky. By leveraging offline datasets, MARL enhances learning efficiency and system performance, enabling agents to achieve superior outcomes in new tasks [94]. Innovative methods such as Parallel Attentional Transfer (PAT) and the integration of teacher policies significantly improve team-wide learning performance and stability, consistently outperforming existing approaches in various multi-agent contexts.

Frameworks like Uni-O4 exemplify the dual capability of MARL to excel in both offline and online settings, broadening applicability to real-world scenarios [92]. Reward shaping techniques are vital in environments with sparse rewards, enhancing learning efficiency and facilitating effective agent cooperation [82]. Furthermore, containerized and distributed MARL algorithms, such as CMARL, demonstrate superior performance over traditional non-distributed methods, achieving better results in benchmarks like GRF and SMAC [104].

Advanced models and techniques, including the option-critic framework for Dec-POMDPs, have expanded MARL capabilities, allowing for competitive performance and improved convergence properties in complex environments [105]. The integration of contract theory with reinforcement learning, as proposed by Ivanov et al., holds promise for maximizing social welfare in sequential social dilemmas, highlighting the interdisciplinary potential of MARL [106].

The MRPG algorithm's success in achieving linear convergence to Nash equilibria in GS-MFTG settings marks significant improvements over existing methods, emphasizing MARL's potential to advance complex strategic interactions [107]. Additionally, the application of causal methods enhances the robustness, interpretability, and counterfactual reasoning capabilities of MARL systems, contributing to more sophisticated decision-making processes [74].

The effectiveness of Risk-based Optimistic Exploration (ROE) in managing exploration-exploitation trade-offs under uncertainty signifies its potential impact on the future of MARL [66]. Experiments indicate that explicitly driving learning with coordination leads to improved agent performance, further underscoring the importance of strategic collaboration in multi-agent systems [10].

As the field evolves, integrating innovative methodologies and frameworks will be crucial for unlocking the full potential of MARL, paving the way for more sophisticated and adaptive intelligent systems. The advancements in multi-agent offline policy reinforcement learning demonstrate a significant impact on artificial intelligence, offering robust solutions for optimizing agent interactions and enhancing system performance across various applications.

References

[1] Afshin OroojlooyJadid and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

[2] Dianbo Liu, Vedant Shah, Oussama Boussif, Cristian Meo, Anirudh Goyal, Tianmin Shu, Michael Mozer, Nicolas Heess, and Yoshua Bengio. Coordinating policies among multiple agents via an intelligent communication channel, 2022.

[3] Kailash Gogineni, Peng Wei, Tian Lan, and Guru Venkataramani. Scalability bottlenecks in multi-agent reinforcement learning systems, 2023.

[4] Rafael Pina, Varuna De Silva, Corentin Artaud, and Xiaolan Liu. Fully independent communication in multi-agent reinforcement learning, 2024.

[5] Xuechen Mu, Hankz Hankui Zhuo, Chen Chen, Kai Zhang, Chao Yu, and Jianye Hao. Hierarchical task network planning for facilitating cooperative multi-agent reinforcement learning, 2023.

[6] Yongyuan Liang and Bangwei Li. Parallel knowledge transfer in multi-agent reinforcement learning, 2020.

[7] Mingliang Zhang, Sichang Su, Chengyang He, and Guillaume Sartoretti. Hybrid training for enhanced multi-task generalization in multi-agent reinforcement learning, 2024.

[8] Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. Efficient collaborative multi-agent deep reinforcement learning for large-scale fleet management, 2019.

[9] Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:3991–4002, 2021.

[10] Sean L. Barton, Nicholas R. Waytowich, and Derrik E. Asher. Coordination-driven learning in multi-agent problem spaces, 2018.

[11] Qihan Liu, Jianing Ye, Xiaoteng Ma, Jun Yang, Bin Liang, and Chongjie Zhang. Efficient multi-agent reinforcement learning by planning, 2024.

[12] Ingy Elsayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding, 2021.

[13] Hussain Kazmi, Johan Suykens, Attila Balint, and Johan Driesen. Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads. *Applied energy*, 238:1022–1035, 2019.

[14] Yusei Naito, Tomohiko Jimbo, Tadashi Odashima, and Takamitsu Matsubara. Task-priority intermediated hierarchical distributed policies: Reinforcement learning of adaptive multi-robot cooperative transport, 2024.

[15] Jakub Grudzien Kuba, Xidong Feng, Shiyao Ding, Hao Dong, Jun Wang, and Yaodong Yang. Heterogeneous-agent mirror learning: A continuum of solutions to cooperative marl, 2022.

[16] Zongkai Liu, Qian Lin, Chao Yu, Xiawei Wu, Yile Liang, Donghui Li, and Xuetao Ding. Offline multi-agent reinforcement learning via in-sample sequential policy optimization, 2024.

[17] Linghui Meng, Muning Wen, Chenyang Le, Xiyun Li, Dengpeng Xing, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, Yaodong Yang, et al. Offline pre-trained multi-agent decision transformer. *Machine Intelligence Research*, 20(2):233–248, 2023.

[18] JaeYoon Kim, Junyu Xuan, Christy Liang, and Farookh Hussain. A non-monolithic policy approach of offline-to-online reinforcement learning, 2024.

[19] Hai Zhong, Xun Wang, Zhuoran Li, and Longbo Huang. Offline-to-online multi-agent reinforcement learning with offline value function memory and sequential exploration, 2024.

[20] Wesley Suttle, Zhuoran Yang, Kaiqing Zhang, Zhaoran Wang, Tamer Basar, and Ji Liu. A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning, 2019.

[21] Robin Brown, Federico Rossi, Kiril Solovey, Michael T. Wolf, and Marco Pavone. On local computation for optimization in multi-agent systems, 2020.

[22] Roy Zohar, Shie Mannor, and Guy Tennenholtz. Locality matters: A scalable value decomposition approach for cooperative multi-agent reinforcement learning, 2021.

[23] Towards a more efficient computation of individual attribute and policy contribution for post-hoc explanation of cooperative multi-agent systems using myerson values.

[24] Jun Feng, Heng Li, Minlie Huang, Shichen Liu, Wenwu Ou, Zhirong Wang, and Xiaoyan Zhu. Learning to collaborate: Multi-scenario ranking via multi-agent reinforcement learning. In *Proceedings of the 2018 World Wide Web Conference*, pages 1939–1948, 2018.

[25] Xiangyu Liu and Kaiqing Zhang. Partially observable multi-agent reinforcement learning with information sharing, 2024.

[26] Zeng Da. Research on multi-agent communication and collaborative decision-making based on deep reinforcement learning, 2023.

[27] Lionel Blondé, Alexandros Kalousis, and Stéphane Marchand-Maillet. Optimality inductive biases and agnostic guidelines for offline reinforcement learning, 2022.

[28] Benjamin Freed, Aditya Kapoor, Ian Abraham, Jeff Schneider, and Howie Choset. Learning cooperative multi-agent policies with partial reward decoupling, 2021.

[29] Zhenbo Cheng, Xingguang Liu, Leilei Zhang, Hangcheng Meng, Qin Li, and Xiao Gang. Improved cooperation by balancing exploration and exploitation in intertemporal social dilemma tasks, 2021.

[30] Jiachen Yang, Alireza Nakhaei, David Isele, Kikuo Fujimura, and Hongyuan Zha. Cm3: Cooperative multi-goal multi-stage multi-agent reinforcement learning, 2020.

[31] Yang Li, Wenhao Zhang, Jianhong Wang, Shao Zhang, Yali Du, Ying Wen, and Wei Pan. Aligning individual and collective objectives in multi-agent cooperation, 2024.

[32] Leonid Peshkin, Kee-Eung Kim, Nicolas Meuleau, and Leslie Pack Kaelbling. Learning to cooperate via policy search, 2014.

[33] Jian Hu, Siyang Jiang, Seth Austin Harding, Haibin Wu, and Shih-wei Liao. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2102.03479*, 2021.

[34] Jianxiong Li, Xiao Hu, Haoran Xu, Jingjing Liu, Xianyuan Zhan, Qing-Shan Jia, and Ya-Qin Zhang. Mind the gap: Offline policy optimization for imperfect rewards, 2023.

[35] Tatiana Tatarenko. Stochastic learning in potential games: Communication and payoff-based approaches, 2018.

[36] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning, 2022.

[37] Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Satisficing paths and independent multi-agent reinforcement learning in stochastic games, 2023.

[38] Jalal Etesami and Christoph-Nikolas Straehle. Non-cooperative multi-agent systems with exploring agents, 2020.

[39] Chenxing Liu and Guizhong Liu. Jointppo: Diving deeper into the effectiveness of ppo in multi-agent reinforcement learning, 2024.

[40] Ruan de Kock, Omayma Mahjoub, Sasha Abramowitz, Wiem Khlifi, Callum Rhys Tilbury, Claude Formanek, Andries Smit, and Arnu Pretorius. Mava: a research library for distributed multi-agent reinforcement learning in jax, 2023.

[41] Hangyu Mao, Zhibo Gong, and Zhen Xiao. Reward design in cooperative multi-agent reinforcement learning for packet routing, 2020.

[42] Shruti Mishra, Ankit Anand, Jordan Hoffmann, Nicolas Heess, Martin Riedmiller, Abbas Abdolmaleki, and Doina Precup. Policy composition in reinforcement learning via multi-objective policy optimization, 2023.

[43] Kevin Waugh, Brian D. Ziebart, and J. Andrew Bagnell. Computational rationalization: The inverse equilibrium problem, 2011.

[44] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

[45] Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning, 2020.

[46] Aaron Sonabend-W, Junwei Lu, Leo A. Celi, Tianxi Cai, and Peter Szolovits. Expert-supervised reinforcement learning for offline policy learning and evaluation, 2020.

[47] Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, and Ofir Nachum. Opal: Offline primitive discovery for accelerating offline reinforcement learning. *arXiv preprint arXiv:2010.13611*, 2020.

[48] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. *arXiv preprint arXiv:2102.04402*, 2021.

[49] Roel Dobbe, David Fridovich-Keil, and Claire Tomlin. Fully decentralized policies for multi-agent systems: An information theoretic approach, 2017.

[50] Yuxuan Yi, Ge Li, Yaowei Wang, and Zongqing Lu. Learning to share in multi-agent reinforcement learning, 2022.

[51] Marek Rosa, Olga Afanasjeva, Simon Andersson, Joseph Davidson, Nicholas Guttenberg, Petr Hlubuček, Martin Poliak, Jaroslav Vítku, and Jan Feyereisl. Badger: Learning to (learn [learning algorithms] through multi-agent communication), 2019.

[52] Lucas Cassano, Kun Yuan, and Ali H. Sayed. Multi-agent fully decentralized value function learning with linear convergence rates, 2019.

[53] Xuejing Zheng and Chao Yu. Multi-agent reinforcement learning with a hierarchy of reward machines, 2024.

[54] Ge Guo, Wing Shing Wong, and Zhongchang Liu. Cooperative target realization in multi-agent systems allowing choice-based actions, 2012.

[55] Jianhong Wang. Shapley value based multi-agent reinforcement learning: Theory, method and its application to energy network, 2024.

[56] Rupali Bhati, Sai Krishna Gottipati, Clodéric Mars, and Matthew E. Taylor. Curriculum learning for cooperation in multi-agent reinforcement learning, 2023.

[57] Jian Hu, Siyue Hu, and Shih wei Liao. Policy regularization via noisy advantage values for cooperative multi-agent actor-critic methods, 2023.

[58] Jian Zhao, Youpeng Zhao, Weixun Wang, Mingyu Yang, Xunhan Hu, Wengang Zhou, Jianye Hao, and Houqiang Li. Coach-assisted multi-agent reinforcement learning framework for unexpected crashed agents, 2022.

[59] Hsu Kao, Chen-Yu Wei, and Vijay Subramanian. Decentralized cooperative reinforcement learning with hierarchical information structure, 2021.

[60] Siyi Hu, Yifan Zhong, Minquan Gao, Weixun Wang, Hao Dong, Xiaodan Liang, Zhihui Li, Xiaojun Chang, and Yaodong Yang. Marllib: A scalable and efficient multi-agent reinforcement learning library, 2023.

[61] Yuan Tian, Klaus-Rudolf Kladny, Qin Wang, Zhiwu Huang, and Olga Fink. Multi-agent actor-critic with time dynamical opponent model, 2022.

[62] Haichao Zhang, We Xu, and Haonan Yu. Policy expansion for bridging offline-to-online reinforcement learning, 2023.

[63] Sheng Yue, Zerui Qin, Xingyuan Hua, Yongheng Deng, and Ju Ren. Federated offline policy optimization with dual regularization, 2024.

[64] Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.

[65] Allen Nie, Yannis Flet-Berliac, Deon R. Jordan, William Steenbergen, and Emma Brunskill. Data-efficient pipeline for offline reinforcement learning with limited data, 2023.

[66] Jihwan Oh, Joonkee Kim, Minchan Jeong, and Se-Young Yun. Toward risk-based optimistic exploration for cooperative multi-agent reinforcement learning, 2023.

[67] Keyang He, Prashant Doshi, and Bikramjit Banerjee. Many agent reinforcement learning under partial observability, 2021.

[68] Jongmin Lee, Wonseok Jeon, Byung-Jun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation, 2021.

[69] Linghui Meng, Muning Wen, Yaodong Yang, Chenyang Le, Xiyun Li, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, and Bo Xu. Offline pre-trained multi-agent decision transformer: One big sequence model tackles all smac tasks, 2022.

[70] Jakub Grudzien Kuba, Muning Wen, Yaodong Yang, Linghui Meng, Shangding Gu, Haifeng Zhang, David Henry Mguni, and Jun Wang. Settling the variance of multi-agent policy gradients, 2022.

[71] Prashant Trivedi and Nandyala Hemachandra. Multi-agent natural actor-critic reinforcement learning algorithms, 2022.

[72] Hager Radi, Josiah P. Hanna, Peter Stone, and Matthew E. Taylor. Safe evaluation for offline learning: Are we ready to deploy?, 2022.

[73] Tao Ma, Xuzhi Yang, and Zoltan Szabo. To switch or not to switch? balanced policy switching in offline reinforcement learning, 2024.

[74] St John Grimbly, Jonathan Shock, and Arnu Pretorius. Causal multi-agent reinforcement learning: Review and open problems, 2021.

[75] Xiao Du, Yutong Ye, Pengyu Zhang, Yaning Yang, Mingsong Chen, and Ting Wang. Situation-dependent causal influence-based cooperative multi-agent reinforcement learning, 2023.

[76] Rafael Pina, Varuna De Silva, and Corentin Artaud. Learning independently from causality in multi-agent environments, 2023.

[77] Ziyan Wang, Yali Du, Yudi Zhang, Meng Fang, and Biwei Huang. Macca: Offline multi-agent reinforcement learning with causal credit assignment, 2023.

[78] Jueming Hu, Zhe Xu, Weichang Wang, Guannan Qu, Yutian Pang, and Yongming Liu. Decentralized graph-based multi-agent reinforcement learning using reward machines, 2021.

[79] Sergio Valcarcel Macua, Jianshu Chen, Santiago Zazo, and Ali H. Sayed. Distributed policy evaluation under multiple behavior strategies, 2014.

[80] Tianshu Chu, Sandeep Chinchali, and Sachin Katti. Multi-agent reinforcement learning for networked system control. *arXiv preprint arXiv:2004.01339*, 2020.

[81] Qiliang Chen and Babak Heydari. Adaptive network intervention for complex systems: A hierarchical graph reinforcement learning approach, 2024.

[82] Chaoyi Gu, Varuna De Silva, Corentin Artaud, and Rafael Pina. Embedding contextual information through reward shaping in multi-agent learning: A case study from google football, 2023.

[83] Shijun Wang, Mate S. Szalay, Changshui Zhang, and Peter Csermely. Learning and innovative elements of strategy adoption rules expand cooperative network topologies, 2008.

[84] Rundong Wang, Xu He, Runsheng Yu, Wei Qiu, Bo An, and Zinovi Rabinovich. Learning efficient multi-agent communication: An information bottleneck approach, 2020.

[85] Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. *arXiv preprint arXiv:2101.11196*, 2021.

[86] Tanvi Verma, Pradeep Varakantham, and Hoong Chuin Lau. Entropy based independent learning in anonymous multi-agent settings, 2020.

[87] Roel Dobbe, David Fridovich-Keil, and Claire Tomlin. Fully decentralized policies for multi-agent systems: An information theoretic approach. *Advances in neural information processing systems*, 30, 2017.

[88] Andrea Coletta, Svitlana Vyetrenko, and Tucker Balch. K-shap: Policy clustering algorithm for anonymous multi-agent state-action pairs, 2023.

[89] Wonseok Jeon, Paul Barde, Derek Nowrouzezahrai, and Joelle Pineau. Scalable multi-agent inverse reinforcement learning via actor-attention-critic, 2020.

[90] Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. Offline multi-agent reinforcement learning with implicit global-to-local value regularization, 2023.

[91] Dong Chen, Kaian Chen, Zhaojian Li, Tianshu Chu, Rui Yao, Feng Qiu, and Kaixiang Lin. Powernet: Multi-agent deep reinforcement learning for scalable powergrid control. *IEEE Transactions on Power Systems*, 37(2):1007–1017, 2021.

[92] Kun Lei, Zhengmao He, Chenhao Lu, Kaizhe Hu, Yang Gao, and Huazhe Xu. Uni-o4: Unifying online and offline deep reinforcement learning with multi-step on-policy optimization, 2024.

[93] Heechang Ryu, Hayong Shin, and Jinkyoo Park. Multi-agent actor-critic with generative cooperative policy network, 2018.

[94] Tianxu Li, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Qihui Wu, Yang Zhang, and Bing Chen. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 24(2):1240–1279, 2022.

[95] Fnu Hairi, Zifan Zhang, and Jia Liu. Sample and communication efficient fully decentralized marl policy evaluation via a new approach: Local td update, 2024.

[96] Sangwon Hwang, Hoon Lee, Juseong Park, and Inkyu Lee. Decentralized computation offloading with cooperative uavs: Multi-agent deep reinforcement learning perspective, 2022.

[97] Jiachen Yang, Igor Borovikov, and Hongyuan Zha. Hierarchical cooperative multi-agent reinforcement learning with skill discovery, 2020.

[98] Yunqi Zhao, Igor Borovikov, Jason Rupert, Caedmon Somers, and Ahmad Beirami. On multi-agent learning in team sports games, 2019.

[99] Robert Loftin, Aadirupa Saha, Sam Devlin, and Katja Hofmann. Strategically efficient exploration in competitive multi-agent reinforcement learning, 2021.

[100] Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, Shaofeng Zou, and Fei Miao. What is the solution for state-adversarial multi-agent reinforcement learning?, 2024.

[101] Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach, 2018.

[102] Qiaosheng Zhang, Chenjia Bai, Shuyue Hu, Zhen Wang, and Xuelong Li. Provably efficient information-directed sampling algorithms for multi-agent reinforcement learning, 2024.

[103] Natalia Zhang, Xinqi Wang, Qiwen Cui, Runlong Zhou, Sham M. Kakade, and Simon S. Du. Preference-based multi-agent reinforcement learning: Data coverage and algorithmic techniques, 2025.

[104] Siyang Wu, Tonghan Wang, Chenghao Li, Yang Hu, and Chongjie Zhang. Containerized distributed value-based multi-agent reinforcement learning, 2021.

[105] Jhelum Chakravorty, Nadeem Ward, Julien Roy, Maxime Chevalier-Boisvert, Sumana Basu, Andrei Lupu, and Doina Precup. Option-critic in cooperative multi-agent systems, 2020.

[106] Dima Ivanov, Paul Dütting, Inbal Talgam-Cohen, Tonghan Wang, and David C. Parkes. Principal-agent reinforcement learning: Orchestrating ai agents with contracts, 2024.

[107] Muhammad Aneeq uz Zaman, Alec Koppel, Mathieu Laurière, and Tamer Başar. Independent rl for cooperative-competitive agents: A mean-field perspective, 2025.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.