Emergence and Resilience in Multi-Agent Reinforcement Learning

Thomy Phan

München 2023

Emergence and Resilience in Multi-Agent Reinforcement Learning

Thomy Phan

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

vorgelegt von
Thomy Phan

Tag der Einreichung: 17. Oktober 2022

1. Gutachter/in:	Prof. Dr. Claudia Linnhoff-Popien
2. Gutachter/in:	Prof. Dr. Sven Koenig
3. Gutachter/in:	Prof. Dr. Long Tran-Thanh
Tag der Einreichung:	17. Oktober 2022
Tag der Disputation:	19. Juni 2023

Eidesstattliche Versicherung

(siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Thomy Phan
München, 20. Juni 2023

Abstract

Our world represents an enormous *multi-agent system (MAS)*, consisting of a plethora of agents that make decisions under uncertainty to achieve certain goals. The interaction of agents constantly affects our world in various ways, leading to the emergence of interesting phenomena like life forms and civilizations that can last for many years while withstanding various kinds of disturbances. Building artificial MAS that are able to adapt and survive similarly to natural MAS is a major goal in artificial intelligence as a wide range of potential real-world applications like autonomous driving, multi-robot warehouses, and cyber-physical production systems can be straightforwardly modeled as MAS. *Multi-agent reinforcement learning (MARL)* is a promising approach to build such systems which has achieved remarkable progress in recent years. However, state-of-the-art MARL commonly assumes very idealized conditions to optimize performance in best-case scenarios while neglecting further aspects that are relevant to the real world.

In this thesis, we address emergence and resilience in MARL which are important aspects to build artificial MAS that adapt and survive as effectively as natural MAS do. We first focus on emergent cooperation from local interaction of self-interested agents and introduce a peer incentivization approach based on mutual acknowledgments. We then propose to exploit emergent phenomena to further improve coordination in large cooperative MAS via decentralized planning or hierarchical value function factorization. To maintain multi-agent coordination in the presence of partial changes similar to classic distributed systems, we present adversarial methods to improve and evaluate resilience in MARL. Finally, we briefly cover a selection of further topics that are relevant to advance MARL towards real-world applicability.

Zusammenfassung

Unsere Welt stellt ein riesiges *Multiagentensystem (MAS)* dar, welches aus einer Vielzahl von Agenten besteht, die unter Unsicherheit Entscheidungen treffen müssen, um bestimmte Ziele zu erreichen. Die Interaktion der Agenten beeinflusst unsere Welt stets auf unterschiedliche Art und Weise, wodurch interessante emergente Phänomene wie beispielsweise Lebensformen und Zivilisationen entstehen, die über viele Jahre Bestand haben und dabei unterschiedliche Arten von Störungen überwinden können. Die Entwicklung von künstlichen MAS, die ähnlich anpassungs- und überlebensfähig wie natürliche MAS sind, ist eines der Hauptziele in der künstlichen Intelligenz, da viele potentielle Anwendungen wie zum Beispiel das autonome Fahren, die multirobotergesteuerte Verwaltung von Lagerhallen oder der Betrieb von cyberphyischen Produktionssystemen, direkt als MAS formuliert werden können. *Multi-Agent Reinforcement Learning (MARL)* ist ein vielversprechender Ansatz, mit dem in den letzten Jahren bemerkenswerte Fortschritte erzielt wurden, um solche Systeme zu entwickeln. Allerdings geht der Stand der Forschung aktuell von sehr idealisierten Annahmen aus, um die Effektivität ausschließlich für Szenarien im besten Fall zu optimieren. Dabei werden weiterführende Aspekte, die für die echte Welt relevant sind, größtenteils außer Acht gelassen. In dieser Arbeit werden die Aspekte Emergenz und Resilienz in MARL betrachtet, welche wichtig für die Entwicklung von anpassungs- und überlebensfähigen künstlichen MAS sind. Es wird zunächst die Entstehung von emergenter Kooperation durch lokale Interaktion von selbstinteressierten Agenten untersucht. Dazu wird ein Ansatz zur Peer-Incentivierung vorgestellt, welcher auf gegenseitiger Anerkennung basiert. Anschließend werden Ansätze zur Nutzung emergenter Phänomene für die Koordinationsverbesserung in großen kooperativen MAS präsentiert, die dezentrale Planungsverfahren oder hierarchische Faktorisierung von Evaluationsfunktionen nutzen. Zur Aufrechterhaltung der Multiagentenkoordination bei partiellen Veränderungen, ähnlich wie in klassischen verteilten Systemen, werden Methoden des Adversarial Learning vorgestellt, um die Resilienz in MARL zu verbessern und zu evaluieren. Abschließend wird kurz eine Auswahl von weiteren Themen behandelt, die für die Einsatzfähigkeit von MARL in der echten Welt relevant sind.

Acknowledgments

This thesis is the emergent result of an exciting journey involving many wonderful people and comprising more than just the mere sum of encounters.
I would like to express my gratitude to my advisor Claudia Linnhoff-Popien for her trust, her guidance, and the freedom I was given to conduct my research. Her experience and enthusiasm as well as the opportunity to work at her chair was very valuable to me to grow scientifically and personally.
I thank Sven Koenig for his passionate support and inspiration even beyond my thesis work. I thank Long Tran-Thanh for his interest in my research and support of my work. It was my great honor meeting Sven and Long right before the global lockdown and finally having them both serving on my dissertation committee. I also thank Dieter Kranzlmüller for chairing the committee and Albrecht Schmidt for serving as substitute examiner.
During my journey, I was blessed with several amazing mentors. I am particularly grateful to Alf Zugenmaier, Joerg Bewersdorf, and Fang Huang for their initial guidance, patience, and wisdom, which greatly influenced my scientific path at an early stage. I am deeply indebted to Lenz Belzner and Thomas Gabor for continually sharing valuable knowledge, experience, and ideas with me – even late after hours.
I thank my colleagues Fabian Ritz, André Ebert, Sebastian Feld, Markus Friedrich, Carsten Hahn, Steffen Illium, Marie Kiermeier, Robert Müller, Christoph Roch, Kyrill Schmid, Andreas Sedlmeier, Philipp Altmann, Jonas Nüßlein, Michael Kölle, Maximilian Zorn, and all other members of the Mobile and Distributed Systems chair for the great time I had at LMU Munich along with countless inspiring and entertaining conversations, gatherings, and activities that pushed me forward.
Finally, I would like to thank my family and friends for their love, patience, and open ears. Having them around gave me the strength and the resilience to pursue my PhD in the best way possible.

Contents

1. Introduction

> Do cities learn? Not the individuals who populate cities, not the institutions they foster, but the cities themselves. I think the answer is yes.
>
> *Emergence*
> Steven Johnson

1.1. Motivation

Agents are everywhere. An *agent* is an autonomous entity which is able to make decisions under uncertainty to achieve a certain goal. Our world consists of a plethora of agents, naturally representing an enormous *multi-agent system (MAS)*. Agents may represent MAS themselves, consisting of several sub-agents. All agent decisions, no matter what level, constantly affect our world and our future in various unpredictable ways.
Genes represent natural MAS on a micro-level that jointly decide how a living organism is supposed to be built – without actually "knowing" the future behavior and lifestyle of the resulting organism [12]. Such organisms represent agents on a higher level that can make decisions on their own with uncertain future consequences to themselves, their surroundings – and their genes. Humans are highly intelligent specimen of organisms that can make complex decisions to pursue a variety of goals by collaborating or competing with each other regardless of kinship and race [3, 7].
Unlike genes, where failures are ultimately punished by nature with extinction, humans and many other organisms are able to *learn* from experience and therefore being able to adapt to changing circumstances, e.g., caused by natural disasters, conflicts, or economic crises [4]. Through learning, humans successfully established MAS in form of civilizations and nations that last for centuries while withstanding various kinds of disturbances [23].
We define *external disturbances* in a MAS as events that naturally occur in the environment regardless of any agent behavior like seismic activities or astronomical events [23]. *Internal disturbances* in a MAS, on the other hand, are directly caused by agent behavior, e.g., in case of defection, manipulation, or failures [79]. Human society exhibits impressively high degrees of resilience against both kinds of disturbances as humankind managed to adapt and survive for thousands of years [3, 12].

Building artificial MAS that are able to solve complex tasks while withstanding disturbances similarly to natural MAS is a major goal in *artificial intelligence (AI)* [74, 87]. Many potential real-world applications of AI like autonomous driving, multi-robot warehouses, and cyber-physical production systems can be straightforwardly modeled as MAS, consisting of several autonomous components with individual or common objectives [31, 46, 91]. *Reinforcement Learning (RL)* provides a general framework to enable agents to learn and adapt from experience [69, 76]. *Multi-agent RL (MARL)* extends RL to MAS, where all agents are additionally trained to interact with each other [3, 87]. Recent advances in complex board and video games have demonstrated the potential of MARL to build artificial MAS under idealized conditions [70, 84, 85]. However, there are several important aspects that need further consideration in order to advance MARL towards real world applicability.
In this thesis, we focus on *emergence* and *resilience* in MARL, where

Emergence is a phenomenon that arises from local interaction of multiple agents, representing an effect that comprises *more than just the mere sum of parts or interactions* [11]. Life in organisms is the emergent result of multiple interacting genes despite genes themselves being actually non-living molecules [12]. Living organisms exhibit interactive behavior on their own to jointly form swarms or civilizations which steadily evolve and adapt as a whole. In these examples, agents are usually only aware of their local interaction but not of potential global consequences. Emergence therefore represents some kind of self-organizing *macrobehavior* without any centralized controller [23].

Resilience is the ability of a MAS to withstand external and internal disturbances. Animal groups like swarms, flocks, or herds are resilient as the occasional loss or mutation of some individuals does not affect the existence of the whole group [23]. Global economy is typically resilient against crises, since most companies are able to adapt through emergency plans and innovation. Even if single companies die out, the global economy itself persists. Classic distributed systems are designed with resilience in mind to ensure constant availability of data and services without failing entirely. Resilience is therefore required to *maintain* certain emergent properties of the MAS [79].

The main contents presented in the thesis are based on the following central hypothesis:

Hypothesis 1.1

Emergence and resilience are important aspects that need to be considered in MARL in order to build artificial multi-agent systems that adapt and survive in the real world as effectively as natural multi-agent systems do.

1.2. Scope and Research Questions

Our goal is to provide methods that explicitly consider emergence and resilience to advance MARL towards building artificial MAS according to our central hypothesis 1.1. We focus on *multi-agent cooperation* in various settings to build cohesive agent systems that directly benefit our world[1] [79]. Unless stated otherwise, we simply use the term *agent* for *artificial agent*.
The research presented in this thesis is guided by the following questions:

(Q1) Can cooperation emerge in MARL from local interaction? As self-learning agents become more and more omnipresent in the real world, they will inevitably learn to interact with each other. Non-cooperative game theory and empirical studies have shown that naive RL approaches commonly fail to cooperate, possibly leading to undesirable emergent results. We present a peer incentivization protocol, where agents learn to cooperate from mutual acknowledgments. Our approach is only based on local information and communication therefore offering scalability which is useful in a steadily expanding world of agents.

(Q2) How to maintain emergent cooperation in MARL? Achieving emergent cooperation under specific circumstances is an important step towards real-world deployment of agents. However, the real world is messy with many unpredictable external factors that could destabilize cooperative situations. In our peer incentivization protocol mentioned in **Q1**, we introduce a penalization mechanism that enables agents to reciprocate in order to maintain cooperation even under social pressure, where many agents compete for scarce resources. We also demonstrate that the locality of information and communication naturally provides some degree of resilience against random protocol defections and communication failures.

(Q3) How to consider emergence to improve performance in MARL? Learning typically involves self-reflection about past or potential future behavior. Cooperative MARL often exploits global information like states and joint actions during training to produce coordinated strategies for decentralized decision making. However, emergent phenomena can dynamically occur in various forms and levels which are difficult to deduce from mere states and joint actions therefore limiting performance and scalability. We propose to explicitly consider emergence in form of globally tracked models for prediction and dynamic team structures for hierarchization to overcome these limitations.

[1] We point out that emergence and resilience are not necessarily limited to cooperation though. The insights drawn from the thesis can be applied to the opposite as well, where multi-agent defection is desirable, e.g., to uncover weaknesses or flaws in a MAS [23, 82].

(Q4) How to improve resilience in MARL against partial changes? As known from classic distributed systems, *partial changes* can occur as internal disturbances in MAS, where agents unexpectedly alter their behavior due to updates, manipulation, or flaws. Thus, even fully cooperative MAS should be prepared for partial changes to prevent catastrophic failure. While many works on MARL assume idealistic conditions to optimize performance for the best-case, we focus on worst-case scenarios in cooperative MAS and propose antagonist-based methods, where all agents can potentially change their behavior adversarially to expose the target system to less idealistic situations.

(Q5) How to evaluate resilience in MARL against partial changes? The performance of a MAS is often evaluated in best-case scenarios, where the MAS under test consists of the exact same agents as seen during training therefore not considering the possibility of partial change at all. In addition to our antagonist-based methods mentioned in **Q4**, we propose testing methods to evaluate resilience of any MAS using dedicated test sets of cooperative and adversarial agents.

(Q6) What are further relevant topics? Beside emergence and resilience, there are many other topics that are also relevant to building artificial MAS for the real world. We briefly present a selection of work regarding resource efficiency, state uncertainty, and non-cooperative emergence.

1.3. Thesis Structure

Chapter 2 provides background knowledge and important terms that are relevant for the main body of the thesis. In each main chapter as listed in Table 1.1, we give a brief introduction and present a selection of results but also refer to the corresponding literature and the attached publications for a better overall understanding.

Table 1.1.: Overview of the main contents of the thesis.

Thesis chapter	Questions	Publications
3. Emergent Cooperation in General-Sum Games	**Q1**, **Q2**	[55]
4. Emergence in Cooperative Multi-Agent RL	**Q3**	[49, 53, 54]
5. Resilience in Cooperative Multi-Agent RL	**Q4**, **Q5**	[47, 51]
6. Further Topics	**Q6**	[18, 48, 50, 52]

Chapter 7 concludes the thesis by connecting the findings to the research questions of Section 1.2 and provides potential directions for future research.
All publications that contributed to this thesis are listed in Table 1.1. Further information on the publications is provided in the appendix.

2. Background

> [...] if bacteria can play games, so can people and nations.
>
> *The Evolution of Cooperation*
> Robert Axelrod

2.1. Multi-Agent Settings

We formulate MAS as *partially observable stochastic game (POSG)* $M = \langle \mathcal{D}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{Z}, \Omega, b_0 \rangle$, where $\mathcal{D} = \{1, ..., N\}$ is a set of *agents* i, $\mathcal{S}$ is a set of *states* s_t at time step t, $\mathcal{A} = \langle \mathcal{A}_1, ..., \mathcal{A}_N \rangle = \langle \mathcal{A}_i \rangle_{i \in \mathcal{D}}$ is the set of *joint actions* $a_t = \langle a_{t,i} \rangle_{i \in \mathcal{D}}$, $\mathcal{P}(s_{t+1}|s_t, a_t)$ is the state transition probability, $\mathcal{R}(s_t, a_t) = \langle \mathcal{R}_i(s_t, a_t) \rangle_{i \in \mathcal{D}} = \langle r_{t,i} \rangle_{i \in \mathcal{D}} \in \mathbb{R}$ is the *joint reward*, $\mathcal{Z}$ is a set of *(local) observations* $z_{t,i}$ for each agent i, $\Omega(s_t) = z_t = \langle z_{t,i} \rangle_{i \in \mathcal{D}} \in \mathcal{Z}^N$ is the *joint observation* of state s_t, and b_0 is the probability distribution over initial states $s_0 \in \mathcal{S}$ [13, 19]. Each agent i maintains a *(local) history* $\tau_{t,i} \in (\mathcal{Z} \times \mathcal{A}_i)^t$ with $\tau_t = \langle \tau_{t,i} \rangle_{i \in \mathcal{D}}$ being the *joint history*. A *belief state* $b(s_t|\tau_t)$ is a sufficient statistic for joint history τ_t and defines a probability distribution over states s_t given τ_t which can be updated by Bayes' theorem [25, 41, 61].

A *joint policy* $\pi = \langle \pi_i \rangle_{i \in \mathcal{D}}$ with *(local) policies* π_i, defines the joint action probability $\pi(a_t|\tau_t) = \prod_{i \in \mathcal{D}} \pi_i(a_{t,i}|\tau_{t,i})$. We define π_{-i} as *opponent policies* such that $\pi = \langle \pi_i, \pi_{-i} \rangle$. For convenience, we use the term *opponent* of agent i for all other agents $j \neq i$ without strictly assuming a competitive game as suggested in [14]. The *return* of agent i is defined by $G_{t,i} = \sum_{c=0}^{T-1} \gamma^c r_{t+c,i}$, where T is the *horizon* and $\gamma \in [0, 1]$ is the *discount factor*. Local policies π_i can be evaluated with a *value function* defined by $V_i^{\pi}(\tau_t) = \mathbb{E}_{\pi, b_0, \mathcal{P}}[G_{t,i}|\tau_t]$.

Nash equilibria are a solution concept describing joint policies $\pi^+ = \langle \pi_i^+ \rangle_{i \in \mathcal{D}}$, where all π_i^+ maximize their values w.r.t. the opponent policies π_{-i}^+ such that $V_i^{\pi^+} = V_i^+ \geq V_i^{\langle \pi_i, \pi_{-i}^+ \rangle}$ for all π_i. Another concept are *optimal joint policies* $\pi^* = \langle \pi_i^* \rangle_{i \in \mathcal{D}}$, where *social welfare*, i.e., the *utilitarian metric* or *efficiency* $U = \sum_{i \in \mathcal{D}} G_{0,i}$, is maximized such that $\mathbb{E}_{\pi^*, b_0, \mathcal{P}}[U] \geq \mathbb{E}_{\pi, b_0, \mathcal{P}}[U]$ for all π.

We assume that all agents only perceive their own observations $z_{t+1,i}$ and rewards $r_{t,i}$ but have no knowledge about the actual state s_t or their opponents' actions, observations, or rewards. In addition, we assume all agents to be

rational, i.e., act to maximize their individual values[1] V_i^π [62, 66].
POSGs can model various multi-agent scenarios with different purposes and challenges. In this thesis, we focus on the following settings:

Cooperative Games are situations of pure coordination, where all agents $i, j \in \mathcal{D}$ share the same global reward $r_t = r_{t,i} = r_{t,j}$ thus need to collaborate to achieve a common goal [6, 41, 66]. Tasks that are solvable in a distributed manner like multi-robot warehouse commissioning or automated manufacturing in smart factories, naturally represent a cooperative game [31, 46]. A Nash equilibrium in a cooperative game is equivalent to an optimal joint policy such that $\pi^+ = \pi^*$ [6].

Zero-Sum Games are situations of pure competition with $N = 2$ agents having opposing rewards $r_{t,1} = -r_{t,2}$ [33, 66]. Zero-sum games are often used for learning and testing policy robustness in worst-case scenarios [16, 56, 82]. A Nash equilibrium in a zero-sum game consists of two *minimax optimal policies* $\pi^+ = \langle \pi_1^+, \pi_2^+ \rangle$ which maximize their respective *worst-case value* defined by $V_i^+ = max_{\pi_i} min_{\pi_j} \{V_i^{\langle \pi_i, \pi_j \rangle}\}$ [33].

General-Sum Games represent the middle ground between pure coordination and competition, where all agents $i \in \mathcal{D}$ have individual preferences expressed by their respective rewards $r_{t,i}$. Self-learning systems designed for different tasks form a general-sum game when sharing the same environment [14, 17]. Those systems or agents can learn to benefit from interaction with each other [87]. In general, there can be multiple Nash equilibria with different levels of social welfare, where the true system goal depends on domain-dependent aspects [66].

Due to our focus on multi-agent cooperation, we adopt an *optimization perspective* to find optimal joint policies[2] π^* in cooperative and general-sum games or minimax optimal policies π^+ in zero-sum games respectively [33, 41, 67].
The joint policy π could in principle be realized with a centralized controller which has access to the joint history τ_t. However, this would limit scalability, privacy, and resilience in practice [41]. We therefore focus on learning local policies π_i for *decentralized decision making*.

2.2. Dynamic Programming

Local policies π_i can be optimized with *dynamic programming (DP)* by assuming that the opponent policies π_{-i} are fixed and known [40]. Given a model of

[1] The values may change, e.g., due to opponent adaptation or peer incentivization.
[2] Searching for Nash equilibria π^+ is a sensible goal for cooperative and zero-sum games but not necessarily desirable for general-sum games because they can be globally inefficient [3, 12, 67]. Furthermore, prior work implies that knowing one's opponent is often more beneficial than merely striving for a Nash equilibrium [62, 67, 73].

the environment M, π_i can be evaluated by computing V_i^π for all τ_t:

$$V_i^\pi(\tau_t) = \sum_{a_t \in \mathcal{A}} \pi(a_t|\tau_t) \sum_{s_t \in \mathcal{S}} b(s_t|\tau_t) \Bigg(\mathcal{R}_i(s_t, a_t) + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{P}(s_{t+1}|s_t, a_t) V_i^\pi(\tau_{t+1}) \Bigg) \tag{2.1}$$

where $\tau_{t+1} = \langle \tau_t, a_t, z_{t+1} \rangle$ is the updated joint history as concatenation of τ_t, a_t, and $z_{t+1} = \Omega(s_{t+1})$.

π_i can be improved with V_i^π by selecting actions that maximize the value for all joint histories τ_t. The iterative process of alternating *policy evaluation* and *policy improvement* is known as *policy iteration* and provably converges to an optimal local policy given that all other policies π_{-i} are fixed [58, 76].

Value iteration is an alternative DP algorithm to compute the optimal value function V_i^* from a random guess V_i^0 by repeatedly applying the following update for all τ_t until convergence [19, 40]:

$$V_i^{k+1}(\tau_t) \leftarrow max_{a_t \in \mathcal{A}} \left\{ \sum_{s_t \in \mathcal{S}} b(s_t|\tau_t) \Bigg(\mathcal{R}_i(s_t, a_t) + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{P}(s_{t+1}|s_t, a_t) V_i^k(\tau_{t+1}) \Bigg) \right\} \tag{2.2}$$

An optimal policy for agent i is then determined by maximizing V_i^*.

DP can find optimal joint policies π^* in cooperative games via policy or value iteration, e.g., by optimizing all local policies alternatingly while keeping the opponent policies fixed or by backward policy construction [19, 40].

However, finding optimal joint policies in POSGs is NEXP-complete hence being intractable for long-horizon games with many states and agents because of the curse of dimensionality w.r.t. the space of belief states, which is $|\mathcal{S}|$-dimensional, as well as exponentially scaling joint action and history spaces [41]. Furthermore, DP requires the explicit distributions of b_0 and $\mathcal{P}$, which are infeasible to specify for large state spaces $\mathcal{S}$. Practical algorithms based on DP often use a black box simulator as *generative model* for Monte Carlo planning, function approximation techniques for RL, or a combination of both [27, 39, 69, 71].

2.3. Multi-Agent Reinforcement Learning

Multi-agent RL (MARL) can approximate optimal policies π_i^* from experience which is generated through explorative[3] interaction between agent i with a simulated or real environment along with all opponents. Many MARL algorithms approximate the *action-value function* Q_i^π instead of V_i^π [15, 60, 72]:

[3]The exploration-exploitation dilemma is a fundamental RL problem to balance between data acquisition and convergence speed, which is out of scope for this thesis.

$$Q_i^\pi(\tau_t, a_t) = \mathbb{E}_{\pi,b_0,\mathcal{P}}\left[\sum_{c=0}^{T-1} \gamma^c \mathcal{R}_i(s_{t+c}, a_{t+c}) \Big| \tau_t, a_t\right] \tag{2.3}$$

$$= \mathbb{E}_{b_0}\left[\mathcal{R}_i(s_t, a_t) + \gamma \sum_{s_{t+1}\in\mathcal{S}} \mathcal{P}(s_{t+1}|s_t, a_t) V_i^\pi(\tau_{t+1}) \Big| \tau_t, a_t\right] \tag{2.4}$$

$$= \sum_{s_t\in\mathcal{S}} b(s_t|\tau_t)\left(\mathcal{R}_i(s_t, a_t) + \gamma \sum_{s_{t+1}\in\mathcal{S}} \mathcal{P}(s_{t+1}|s_t, a_t) V_i^\pi(\tau_{t+1})\right) \tag{2.5}$$

The optimal action-value function Q_i^* is defined analogously to V_i^*. In fully observable games, π_i can be improved by maximizing Q_i^π without depending on $\mathcal{P}$ therefore enabling *model-free control* [6, 76]. Most real-world domains are too large for a tabular representation of Q_i^π, V_i^π, and π_i though. Thus, function approximation techniques like *deep learning* are used to train an approximator $\hat{Q}_{i,\omega} \approx Q_i^\pi$ or $\hat{V}_{i,\omega} \approx V_i^\pi$ with parameters ω [39, 68, 80]. π_i can be improved through maximization of $\hat{Q}_{i,\omega}$ or *actor-critic* methods by training an approximator $\hat{\pi}_{i,\theta}$ with parameters θ via gradient ascent on $\hat{Q}_{i,\omega}$ or $\hat{V}_{i,\omega}$ [77]. For simplicity, we always omit θ, ω and write $\hat{\pi}_i$, $\hat{Q}_i$, and $\hat{V}_i$ instead.
We focus on the following MARL paradigms that can be combined and applied to any multi-agent setting from Section 2.1:

Independent Learning applies single-agent RL techniques to each agent in the game [78]. Functions like Q_i^π and V_i^π that require *global information*, i.e., s_t, τ_t, or a_t, are approximated using *local information*, i.e., $\tau_{t,i}$ and $a_{t,i}$, per agent i such that $\hat{Q}_i(\tau_{t,i}, a_{t,i}) \approx Q_i^\pi(\tau_t, a_t)$ or $\hat{V}_i(\tau_{t,i}) \approx V_i^\pi(\tau_t)$ respectively. Independent learning offers optimal scalability in all settings due to the locality of information and parallelization of training but ignores concurrent adaptation of opponents which causes non-stationarity therefore lacking convergence guarantees [28, 72].

Centralized Training for Decentralized Execution (CTDE) assumes training to take place in a laboratory or a simulator hence having access to global information to learn $\hat{Q}_i(\tau_t, a_t) \approx Q_i^\pi(\tau_t, a_t)$ or $\hat{V}_i(\tau_t) \approx V_i^\pi(\tau_t)$. Local policy approximators $\hat{\pi}_i$ are trained with $\hat{Q}_i$ or $\hat{V}_i$, e.g., via actor-critic methods or *value function factorization* and can be executed independently afterwards without requiring $\hat{Q}_i$ or $\hat{V}_i$ anymore [15, 60].

Adversarial Learning assumes a zero-sum game to train two agents or agent teams with opposing objectives. Adversarial learning can be realized in different ways like self-play, alternating training, or co-evolution and is used to learn minimax optimal policies π^+ to achieve safe and resilient behavior, to validate RL systems, or to build MAS with potentially unknown agents [16, 33, 38, 82].

3. Emergent Cooperation in General-Sum Games

> Not being nice may look promising at first, but in the long run it can destroy the very environment it needs for its own success.
>
> *The Evolution of Cooperation*
> Robert Axelrod

3.1. Learning in Social Dilemmas

What will happen if two or more agents meet each other for the first time? Let us assume that all agents originate from different manufacturers and have individual goals therefore being *self-interested*. Let us further assume that all agents – regardless of their origins – are unaware of each others' goals and any global information while being able to adapt to individual experience using independent learning. In autonomous driving, e.g., several self-driving cars need to coordinate their decisions without knowing each others' destinations or the preferences of each others' passengers. In a smart home, multiple intelligent devices like cleaning robots and smart speakers share the same household while having obviously different purposes and probably never met each other before. Despite initial mutual unawareness, all agents will inevitably learn to interact with each other – either directly or through manipulation of the shared environment [3, 87]. As self-learning systems are becoming more and more omnipresent in the future, such scenarios are becoming the norm, not the exception [14, 17]. We can naturally formulate them as general-sum games.
Sharing an environment means competing over its resources which can cause self-interested agents to adopt defective policies that exploit each other and harm social welfare, e.g., self-driving cars causing traffic jams due to greedy navigation or the cleaning robot of a smart home removing all smart speakers to keep the household "quiet and clean". Such tension between individual and collective rationality is typically modeled as *social dilemma (SD)*, i.e., a general-sum game, where Nash equilibria π^+ are different from optimal joint policies π^* such that individual rationality would lead to worse outcomes than is possible [3, 12]. Although scenarios of mutual defection are certainly not desirable for real-world applications, non-cooperative game theory and empir-

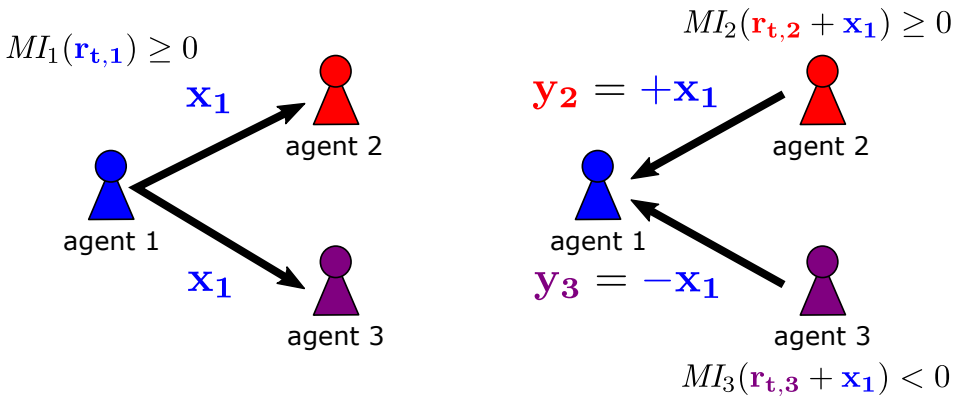
ical studies have shown that naive independent learning agents commonly fail to cooperate in SDs as studied extensively for the prisoner's dilemma [4, 59]. Fortunately, studies from evolutionary biology and human society have shown that cooperation can still emerge in various ways despite all individuals being (presumably) selfish at the core [4, 7, 12, 81]. Especially human society is seen as an "unusual and interesting special case" of emergent cooperation that is not exclusively based on kinship or race thus providing inspiration for building agents that are able to cooperate out of self-interest [3, 7, 12]. However, merely establishing emergent cooperation is not sufficient to sustain social welfare under social pressure, where many agents compete for scarce resources [30, 45]. Furthermore, cooperative situations could be easily destabilized through disturbances [3, 12]. Hence, *reciprocity* is important to achieve *and* maintain cooperation in general-sum games by adequately responding to both cooperative and defective opponent behavior [3].
To address the above challenges, we present a scalable peer incentivization approach based on a two-phase communication protocol in the next section.

3.2. Mutual Acknowledgment Exchange

Peer incentivization (PI) is an increasingly popular MARL approach, where agents learn to reward each other to achieve emergent cooperation in SDs [35, 64, 83, 90]. Flawless communication is often assumed to exchange rewards which are simply integrated into the learning process without further feedback. PI is motivated by evolutionary biology and human society, where individuals incentivize each other via supporting actions or side payments in the hope of future compensation [22, 81]. While conceptually interesting, most PI approaches rely on global information like joint actions [90], central market functions [64], or publicly available information [83] which limits scalability and applicability to real-world scenarios. Another common weakness is the lack of opponent penalization for reciprocity, which makes most PI approaches vulnerable to social pressure and disturbances.
Therefore, we propose *Mutual Acknowledgment Token Exchange (MATE)*, a PI approach defined by a two-phase communication protocol to mutually exchange acknowledgment tokens $x_{token} > 0$ to shape rewards for independent learning [55]. After every state transition, each agent i evaluates the *monotonic improvement* MI_i of its situation defined by:

$$MI_i(\hat{r}_{t,i}) = MI_{\langle \tau_{t,i}, a_{t,i}, r_{t,i}, z_{t+1,i} \rangle, \hat{V}_i}(\hat{r}_{t,i}) = \hat{r}_{t,i} + \gamma \hat{V}_i(\tau_{t+1,i}) - \hat{V}_i(\tau_{t,i}) \tag{3.1}$$

which is the *temporal difference (TD)* residual of $\hat{V}_i$ w.r.t. to some arbitrary reward $\hat{r}_{t,i}$ estimating the expected *long-term* improvement of agent i.
The two-phase communication protocol of exchanging rewards using acknowledgment tokens $x_i = x_{token} > 0$ is based on humans thanking and acknowledging each other as illustrated in Figure 3.1. After the acknowledgment token



(a) Request phase (b) Response phase

Figure 3.1.: MATE protocol example. (a) If agent 1 estimates a monotonic improvement $MI_1(r_{t,1}) \geq 0$ of its situation, it "thanks" its neighbor agents 2 and 3 by sending an *acknowledgment request* $x_1 > 0$ as reward. (b) Agent 2 and 3 check if the request x_1 monotonically improves their own situation along with their own respective reward. If so, a positive reward (e.g., $y_2 = +x_1$) is sent back as a response. If not, a negative reward (e.g., $y_3 = -x_1$) is sent back. Image taken from [55].

exchange, the shaped reward $\hat{r}_{t,i}^{MATE}$ for each agent i is computed as follows:

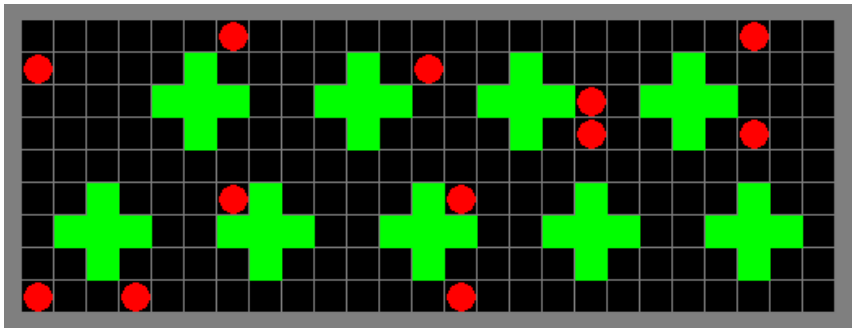
$$\hat{r}_{t,i}^{MATE} = r_{t,i} + \hat{r}_{req} + \hat{r}_{res} = r_{t,i} + max\{\langle x_j \rangle_{j \in \mathcal{N}_{t,i}}\} + min\{\langle y_j \rangle_{j \in \mathcal{N}_{t,i}}\} \quad (3.2)$$

where $\hat{r}_{req} \in \{0, x_{token}\}$ is the aggregation of all received requests x_j and $\hat{r}_{res} \in \{-x_{token}, 0, x_{token}\}$ is the aggregation of all received responses y_j. $\hat{r}_{t,i}^{MATE}$ replaces the original reward $r_{t,i}$ to stably learn cooperative local policies using any RL algorithm, e.g., actor-critic methods.

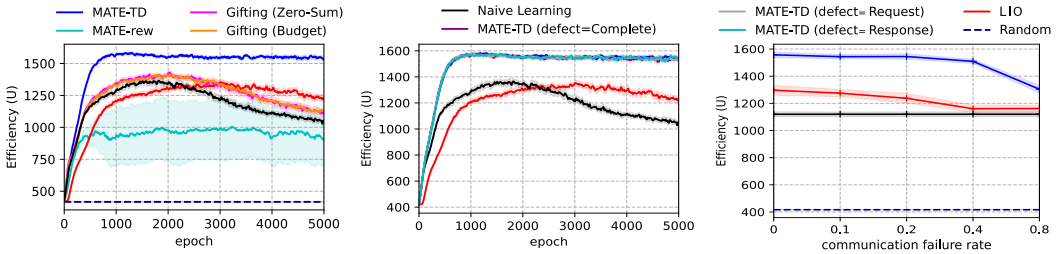
MATE advances existing PI approaches w.r.t. scalability, stability, and resilience against external disturbances. MATE is completely decentralized, only relying on local information, i.e., individual experience and rewards exchanged within a local neighborhood $\mathcal{N}_{t,i}$ thus being more scalable than most prior PI approaches. The term $\hat{r}_{res}$ in Equation 3.2 enables penalization which is necessary for reciprocity to maintain social welfare. Due to the locality of information and communication, MATE naturally exhibits some degree of resilience as local disturbances should not affect the whole MAS.

Empirical results from [55] show that MATE achieves and maintains significantly higher levels of cooperation than alternative MARL w.r.t. different metrics in sequential SD domains as shown in Figure 3.2 for Harvest with 12 agents. The modified TD residual of $\hat{V}_i$ in Equation 3.1 is shown to be a suitable monotonic improvement estimator to incentivize emergent cooperation. MATE is able to maintain social welfare in contrast to prior PI approaches,

which become unstable under social pressure and switch to more defective strategies [35, 90]. MATE is also able to maintain its superior cooperation in the presence of random protocol defections and communication failures as shown in Figure 3.2c and 3.2d.



(a) Harvest Domain ($N = 12$)



(b) Original Setting (c) Protocol Defectors (d) Communication Failures

Figure 3.2.: (a) Harvest domain inspired by [45]. All agents (red circles) need to collect apples (green squares) while avoiding to be tagged and exhaustion of all apples which would prevent regrowth of apples. (b-d) Performance of MATE (as MATE-TD) compared to alternative MARL approaches in different settings of Harvest. (b) Learning progress of MATE and alternative MARL in the standard setting. (c) MATE exhibits resilience w.r.t. protocol defectors as social welfare is not harmed. (d) MATE is able to maintain its superior cooperation even when communication fails with a chance of up to 80%. Image and results taken from [55]

4. Emergence in Cooperative Multi-Agent RL

> There is great power and creative energy in self-organization [...] but it needs to be channeled toward specific forms for it to blossom into something like intelligence.
>
> *Emergence*
> Steven Johnson

How do we know that our society is doing well? Or that it is improving after all? In the previous chapter, we studied how cooperation can emerge in general-sum games as a positive effect of local interaction and adaptation. The agents are not "aware" that their individual decisions lead to overall cooperation though [23]. They simply adapt out of self-interest.
Many distributed real-world applications like multi-robot warehouses or smart factories represent purely cooperative games, where all agents are part of a team, sharing a common goal [31, 46]. While still being able to make individual decisions, all agents now need to collaborate to not just achieve emergent cooperation but overall *coordination*, i.e., cooperation in the best way possible. In such cooperative settings, we assume all agents to be trained together in a controlled environment, e.g., in a laboratory or a simulator with access to global information, e.g., states and joint actions, to learn a centralized value function $\hat{Q}$ or $\hat{V}$ according to the CTDE paradigm explained in Section 2.3. The resulting policies are deployable for decentralized decision making.
Multi-agent credit assignment is a central challenge in cooperative games, since all agents only observe a single global reward. Hence, the deduction of individual contributions is difficult and could lead to uncoordinated policies or "lazy" agents [8, 75, 89]. Another challenge is the limited representational capacity of the centralized value function approximator which becomes a *performance bottleneck* [5, 49, 53]. Both challenges can lead to the emergence of poor local optima, especially in large MAS with many agents.
As state-of-the-art CTDE is often limited to a handful of agents despite exploiting states and joint actions during training, we hypothesize that explicit consideration of emergence is needed to effectively address the above challenges. From a theory of mind perspective, providing higher level knowledge beyond mere states and joint actions could enable agents and CTDE regimes to

effectively "read the (emergent) mind" of the whole MAS therefore potentially improving overall coordination in large cooperative games [23, 57].
In the following sections, we present two MARL approaches that explicitly consider emergence in form of globally tracked models for prediction or dynamic team structures for hierarchization to overcome the limitations of state-of-the-art CTDE.

4.1. Distributed Policy Iteration

According to the dual-system theory, human reasoning is guided by a *system 1*, which is fast and intuitive, and a *system 2*, which is slow and (self-)reflective [26]. *Hybrid (RL) approaches* like *AlphaZero* or *Expert Iteration* realize system 1 using statistical methods to learn a fast and generalizing policy $\hat{\pi}$ while system 2 is based on symbolic or model-based methods like planning, representing a stronger but computationally more expensive policy $\widetilde{\pi}$ [2, 69]. The fast policy approximator of $\hat{\pi}$ is trained via imitation of the symbolic policy $\widetilde{\pi}$ while $\widetilde{\pi}$ is improved by using the fast policy $\hat{\pi}$ as a prior for efficient reasoning. So far, state-of-the-art hybrid approaches mainly focused on single-agent systems or symmetric zero-sum games hence only requiring a single planning process to guide policy learning without any coordination [65].
To learn optimal policies in cooperative games, system 2 can be implemented on different levels. Centralized planning could realize high-level (but not emergent) reasoning for optimal coordination in the MAS. However, as noted in Section 2.2, centralized planning of optimal joint policies π^* is intractable due to exponential computation time. Centralized *Monte Carlo planning (MCP)* could address the curse of dimensionality regarding (belief) states via statistical sampling and a generative model but also does not scale to large MAS due to exponential branching factors caused by the joint action space $\mathcal{A}$ [1, 71]. Decentralized MCP, on the other hand, could realize low-level reasoning in a scalable way but possibly leads to the emergence of uncoordinated behavior. Beside the generative model, decentralized MCP requires knowledge of the opponent policies π_{-i} for coordinated planning [10].
To address this dilemma, we propose *Stable Emergent Policies (STEP)*, a distributed policy iteration scheme, combining centralized learning for policy evaluation and decentralized MCP for policy improvement [49, 54]. Centralized learning approximates the emergent joint policy $\hat{\pi}$ via agent-wise imitation of the decentralized planners $\widetilde{\pi} = \langle \widetilde{\pi}_i \rangle_{i \in \mathcal{D}}$ and a centralized value function $\hat{V}$ to evaluate $\widetilde{\pi}$. Decentralized MCP is implemented with agent-wise open-loop *Monte Carlo Tree Search (MCTS)* and reintegrates $\hat{\pi}$ and $\hat{V}$ to predict emergent effects in order to improve efficiency and coordination w.r.t. $\hat{V}$. After training, the approximated policies $\hat{\pi}_i$ can be executed in a decentralized way without depending on $\hat{V}$, opponent policies $\hat{\pi}_{-i}$, or the slower MCTS anymore. The whole STEP training scheme is illustrated in Figure 4.1.
Integrating $\hat{\pi}$ and $\hat{V}$ improves decentralized MCP in the following aspects:

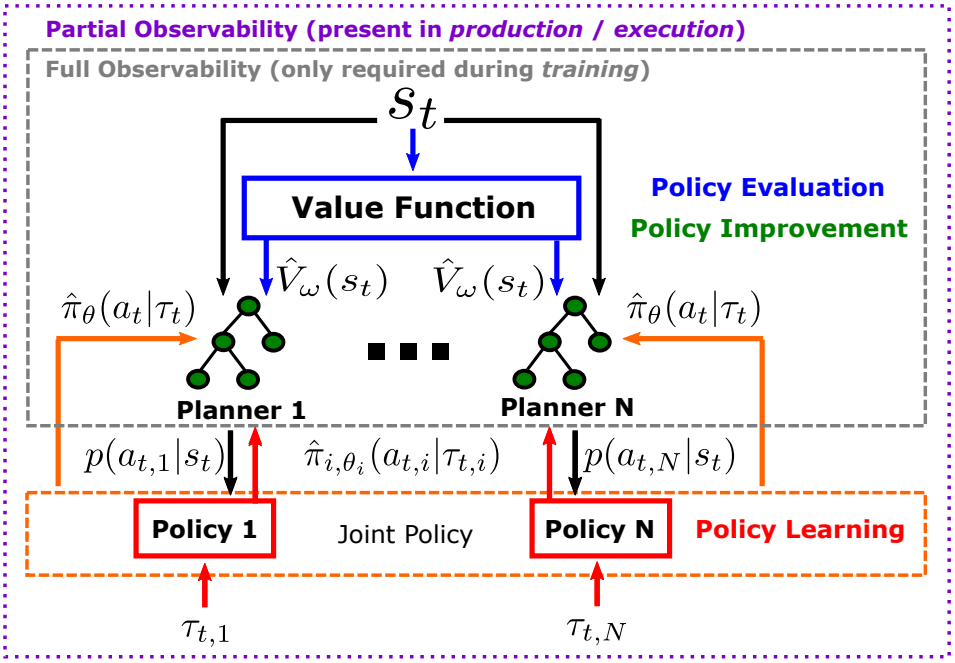


Figure 4.1.: Training architecture of STEP and information flow between the learned policies $\hat{\pi}_{i,\theta_i}$ (red), the learned value function $\hat{V}_\omega$ (blue), and the decentralized planners (green). The policy iteration components (blue and green) within the gray dashed rectangle use global information and are only required during centralized learning, while the learned policies (red) can act under partial observability. Image taken from [49].

Prediction of potential emergent effects is generally a non-trivial task due to non-stationarity and opponent uncertainty. Since $\hat{\pi}$ is trained via imitation of all decentralized planners $\widetilde{\pi}$, reintegrating $\hat{\pi}$ into the MCTS process enables explicit consideration of potential opponent behavior. This effectively addresses the credit assignment problem, allowing for decentralized improvement of individual decisions for the global benefit.

Infinite horizon planning is enabled by $\hat{V}$, which can extend the foresight of MCP to potentially infinite horizons [68, 70]. $\hat{V}$ is learned from global experience produced by all decentralized planners $\widetilde{\pi}$ thus being able to predict the expected return based on simulated states in each planner. Learning $\hat{V}$ instead of $\hat{Q}$ alleviates the performance bottleneck problem, since the input space of $\hat{V}$ only depends on $\mathcal{S}$ without $\mathcal{A}$.

Open-loop planning with closed-loop priors can significantly reduce the branching factor by merely focusing on individual action sequences. Despite lacking optimality guarantuees in stochastic domains, open-loop planning can still be effective in large domains with restricted computational resources, often outperforming closed-loop counterparts in practice

[29, 44, 86]. We use $\hat{\pi}_i$ as *closed-loop prior* to weight actions depending on simulated histories in the MCTS selection rule which maximizes:

$$UCB1^{\hat{\pi}_i}_{Nd_t}(\tau_{t,i}, a_{t,i}) = \overline{Q}(Nd_t, a_{t,i}) + \hat{\pi}_i(a_{t,i}|\tau_{t,i})c\sqrt{\frac{2log(n_{t,i})}{n_{a_{t,i}}}} \tag{4.1}$$

where Nd_t is a node in the open-loop tree, $\overline{Q}$ is the tracked open-loop action value, $c > 0$ is an exploration constant, $n_{t,i}$ is the visitation count of node Nd_t, and $n_{a_{t,i}}$ is the action count of $a_{t,i}$ in node Nd_t. $\hat{\pi}_i$ can potentially recommend different prior actions for the same action sequences due to explicitly considering observations which would be otherwise ignored therefore enabling efficient and effective search.

We evaluated STEP in a predator-prey and a smart factory domain. Our ablation results suggest that integrating $\hat{\pi}$ and $\hat{V}$ can significantly improve overall coordination in cooperative games therefore improving $\hat{\pi}_i$ [49, 54]. Combining decentralized open-loop MCTS with closed-loop priors can be superior to closed-loop MCTS when the computation budget is low or when the branching factor is intractably large [54]. The policy approximations of STEP outperform model-free CTDE approaches like COMA and QMIX in settings with up to 200 agents. The results indicate that our hybrid approach offers a more effective solution to the multi-agent credit assignment and performance bottleneck problem through explicit consideration of emergent effects instead of merely regarding states and joint actions [49].

4.2. Variable Agent-Sub Teams

Hybrid approaches like STEP combine statistical and symbolic methods to produce fast and effective policy approximations. However, MCP requires sufficient simulations per decision hence adding computational costs and being prohibitive if simulations are expensive. Thus, we focus on model-free MARL in the following.

The state-of-the-art in model-free MARL is based on *value function factorization (VFF)*, where the centralized value function $\hat{Q} = Q_{tot}$ is decomposed into local value functions $\langle Q_i \rangle_{i \in \mathcal{D}}$ using a learned *VFF operator* Ψ. Therefore, VFF directly addresses the multi-agent credit assignment problem via end-to-end learning [60, 72, 75]. However, most VFF approaches are limited to a handful of agents in most domains due to the flat factorization scheme, where Ψ becomes a performance bottleneck with increasing number of agents N.

To this end, we propose VFF with *variable agent sub-teams (VAST)* to address that performance bottleneck problem [53]. Instead of directly factorizing $\hat{Q}$ for each agent, VAST approximates a factorization using a VFF operator Ψ for $K \leq N$ disjoint *agent sub-teams* $D_{t,k} \subseteq \mathcal{D}$ which can be defined in an arbitrary way and vary over time, e.g., to adapt to different situations. The sub-team

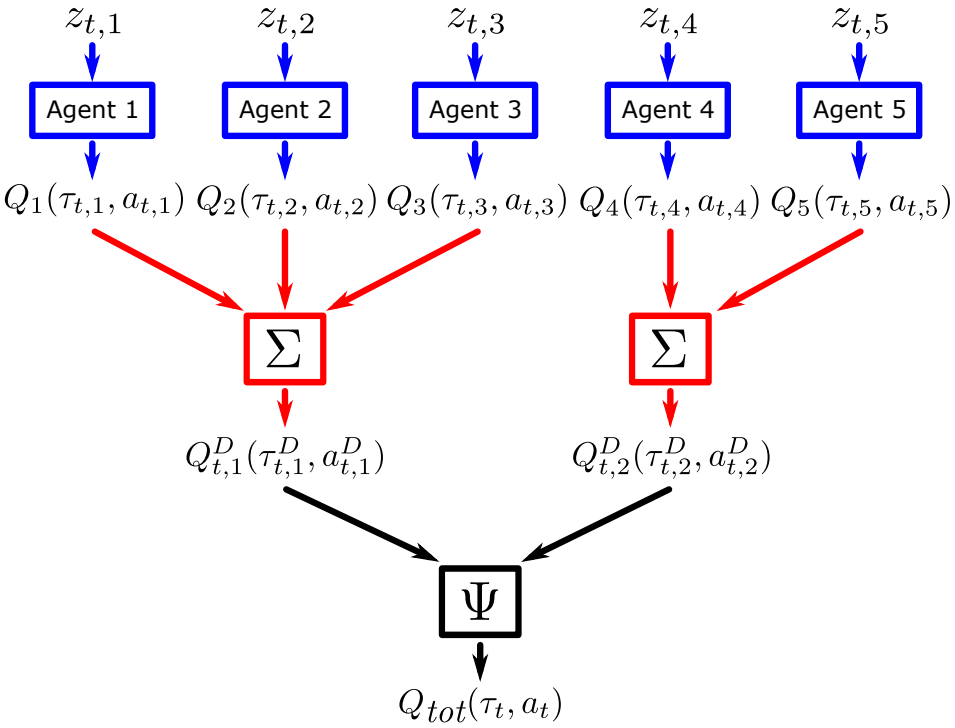
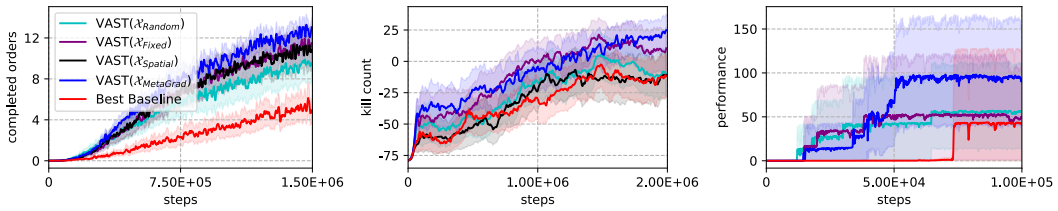


Figure 4.2.: Our hierarchical factorization scheme based on $K = 2$ sub-team values $Q^D_{t,k}$, which are linearly decomposed into local values Q_j per sub-team member $j \in D_{t,k} \subseteq \mathcal{D}$. Image reproduced from [53].

values $Q^D_{t,k}$ are linearly decomposed for all sub-team members $j \in D_{t,k}$ via *value decomposition networks (VDN)* [75]. VDN offers a simple and flexible way to learn sub-team member values Q_j of variable sized sub-teams without adding further complexity through additional parameters. The hierarchization enables Ψ to learn on a more focused and compact input representation. Agents can be assigned to sub-teams by an arbitrary *assignment operator* $\mathcal{X}$ without loosing any guarantuees w.r.t. decentralizability. The hierarchical VFF scheme is illustrated in Figure 4.2.

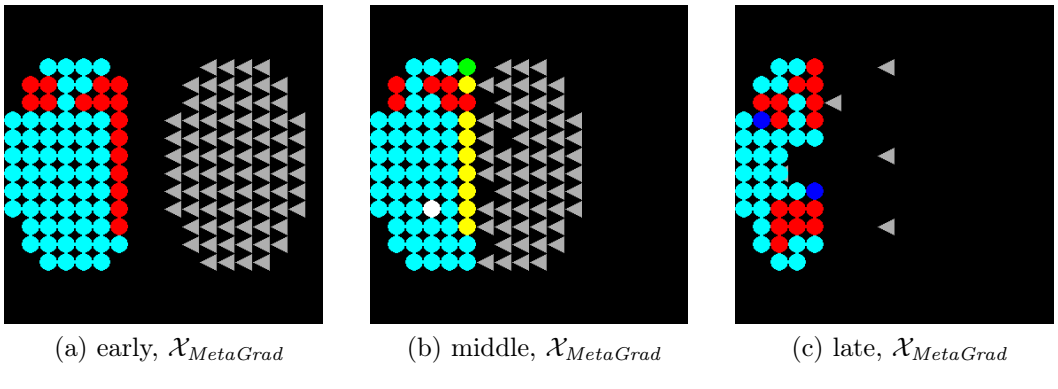
Our empirical results in Figure 4.3 show that a random sub-team assignment $\mathcal{X}_{Random}$ can already be sufficient to outperform flat VFF approaches like QMIX and QTRAN in large-scale domains with up to 800 agents [53]. However, we found that meta optimization of agent sub-teams w.r.t. the return or utilitarian metric U is most effective due to successfully exploiting emergence in form of dynamic team structures.

Figure 4.4 shows an example for emergent sub-teams in the Battle domain determined by a *meta-gradient optimizing* assignment operator $\mathcal{X}_{MetaGrad}$. The sub-teams vary depending on the situation, enabling the learning process to adapt accordingly in order to further improve performance compared to alternative assignments like spatial clustering ($\mathcal{X}_{Spatial}$), which would simply determine sub-teams based on spatial distances regardless of the actual situation.



(a) *Warehouse[16]* (b) *Battle[80]* (c) *GaussianSqueeze[800]*

Figure 4.3.: Average training progress of VAST with sub-team assignment operator $\mathcal{X} \in \{\mathcal{X}_{Random}, \mathcal{X}_{Fixed}, \mathcal{X}_{Spatial}, \mathcal{X}_{MetaGrad}\}$, Ψ_{QTRAN}, and $\eta = \frac{1}{4}$ as well as the respective best VFF baselines reported in [53]. Results taken from [53].



(a) early, $\mathcal{X}_{MetaGrad}$ (b) middle, $\mathcal{X}_{MetaGrad}$ (c) late, $\mathcal{X}_{MetaGrad}$

(d) early, $\mathcal{X}_{Spatial}$ (e) middle, $\mathcal{X}_{Spatial}$ (f) late, $\mathcal{X}_{Spatial}$

Figure 4.4.: Visualizations of the determined sub-teams of $\mathcal{X}_{MetaGrad}$ with $\eta = \frac{1}{4}$ and $\mathcal{X}_{Spatial}$ with k-means clustering using 10 centroids at different stages (early, middle, late) in *Battle[80]* after training. All learning agents (round circles) of the same sub-team have the same color. Image taken from [53].

5. Resilience in Cooperative Multi-Agent RL

> Do we assume, in real life [...], that we are playing a zero sum game when we are not?
>
> *The Selfish Gene*
> Richard Dawkins

What happens if something goes wrong or not the way we expected? In human society, this question often arises when planning activities under uncertainty to prepare for undesirable or unexpected events. In Chapter 3 and 4, we focused on emergent cooperation and coordination via MARL, where we assumed all agents to act exactly according to their learned policies even in the presence of external disturbances. However in the real world, agents themselves can change unexpectedly by altering their policies due to updates, manipulation, or flaws in hardware and software [79]. Such internal disturbances or *partial changes* can cause catastrophic failures, i.e., mutual defection in general-sum games or degrading performance in cooperative games with potentially severe consequences in industrial and safety-critical domains [82].

Resilience has been considered a main motivation for artificial MAS and is commonly found in natural MAS like animal groups or human society, and in classic distributed systems, which are able to survive various kinds of partial change. Resilient artificial MAS are supposed to withstand partial changes by ensuring "graceful degradation" of performance without failing entirely [41]. However, many works on cooperative MARL commonly assume idealized conditions, where all agents only optimize the best-case performance while being evaluated against the exact same agents as seen during training [15, 60, 85]. The possibility of partial change is therefore not considered at all.

We argue that even when being trained for a fully cooperative task, a MAS should *always* be prepared for partial changes and exhibit resilience similar to natural MAS and classic distributed systems to prevent catastrophic failure. Due to the high complexity in MAS, there are many possibilities for agents to deviate from their original policy with potentially negative global impact which cannot be considered exhaustively a priori. We take inspiration from engineering of classic distributed systems, which focuses on worst-case scenarios to ensure dependability through resilience, e.g., reliable and secure commu-

nication or permanent availability of data and services, without exhaustive consideration of partial changes [79]. Thus, we focus on worst-case scenarios in MARL using adversarial learning techniques.
In the following sections, we propose two adversarial MARL approaches to improve resilience in cooperative games. We also provide testing methods to evaluate resilience w.r.t. partial changes based on dedicated agent test sets.

5.1. Antagonist-Ratio Training Scheme

The goal is to build cooperative MAS which are resilient against partial failure. Our *target system* represents a cooperative game $M_C = M$ with N agents $i \in \mathcal{D}$ and a common reward $r_t = \mathcal{R}_i(s_t, a_t)$ for all $i \in \mathcal{D}$ as defined in Section 2.1
For learning and testing, we reformulate the target system as *mixed (cooperative-competitive) game* M_X with N agents $i \in \mathcal{D}_X = \mathcal{D}_{pro} \cup \mathcal{D}_{ant}$ and $\mathcal{D}_{pro} \cap \mathcal{D}_{ant} = \emptyset$. $\mathcal{D}_{pro} \subseteq \mathcal{D}$ is a team of *protagonists* and represents a set of functional agents of the target system M_C. $\mathcal{D}_{ant}$ is a team of *antagonists* and represents a set of adversarial agents, i.e., some partial change in the MAS. The protagonists observe the original reward $r_{t,pro} = r_t$ of M_C, while the antagonist reward is defined by $r_{t,ant} = -r_{t,pro}$. M_X represents a zero-sum game between two agent teams $\mathcal{D}_{pro}$ and $\mathcal{D}_{ant}$, where each team consists of internally cooperating agents [34].
We define the *antagonist-ratio* $R_{ant} = \frac{|\mathcal{D}_{ant}|}{|\mathcal{D}|}$ as adversarial fraction in the mixed-game M_X. If $R_{ant} = 0$, then M_X reduces to the target system M_C.
Based on this setting, we propose an *Antagonist-Ratio Training Scheme (ARTS)* [51]. Since all agents of the target system may be subject to partial change, we maintain a pool of N protagonist policies $\hat{\pi}_{pro} = \langle \hat{\pi}_{i,pro} \rangle_{i \in \mathcal{D}}$ and a pool of N antagonist policies $\hat{\pi}_{ant} = \langle \hat{\pi}_{i,ant} \rangle_{i \in \mathcal{D}}$ to model an adversarial counterpart for each agent or protagonist $i \in \mathcal{D}$ of the target system. We therefore aim to approximate minimax optimal joint policies $\hat{\pi}_{pro} \approx \pi^+$ for various M_X rather than an optimal joint policy π^* for the target system M_C.
Given a fixed R_{ant}, there are theoretically $\binom{N-1}{N \cdot R_{ant}}$ possible antagonist team compositions per protagonist $i \in \mathcal{D}$. To avoid exhaustive exposure of all antagonist team compositions to all agents, we train $\hat{\pi}_{pro}$ and $\hat{\pi}_{ant}$ on randomly sampled mixed-games M_X, where π_X is composed of a random selection of protagonist and antagonist policies according to R_{ant}. A centralized value function $\hat{Q}(\tau_t, a_t) \approx \mathbb{E}_{M_X, \pi_X}[Q^{\pi_X}(\tau_t, a_t)]$ is learned via CTDE based on the protagonist rewards $r_{t,pro}$. The policies of $\hat{\pi}_{pro}$ are then updated to maximize $\hat{Q}$, while the policies of $\hat{\pi}_{ant}$ are updated to minimize $\hat{Q}$ according to the minimax principle [33, 66]. Figure 5.1 gives an overview of ARTS.
ARTS can also be used to test resilience of MAS by omitting the policy updates. Given some cooperative MAS with joint policy $\pi = \hat{\pi}_{pro}$ and an antagonist policy pool $\hat{\pi}_{ant}$, e.g., provided by domain experts or a different MARL process, the scheme from Figure 5.1 can be used to evaluate performance w.r.t. different

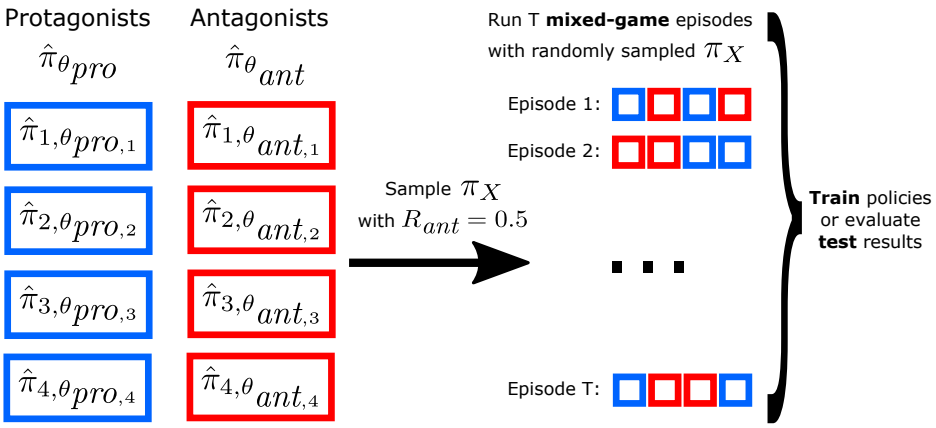
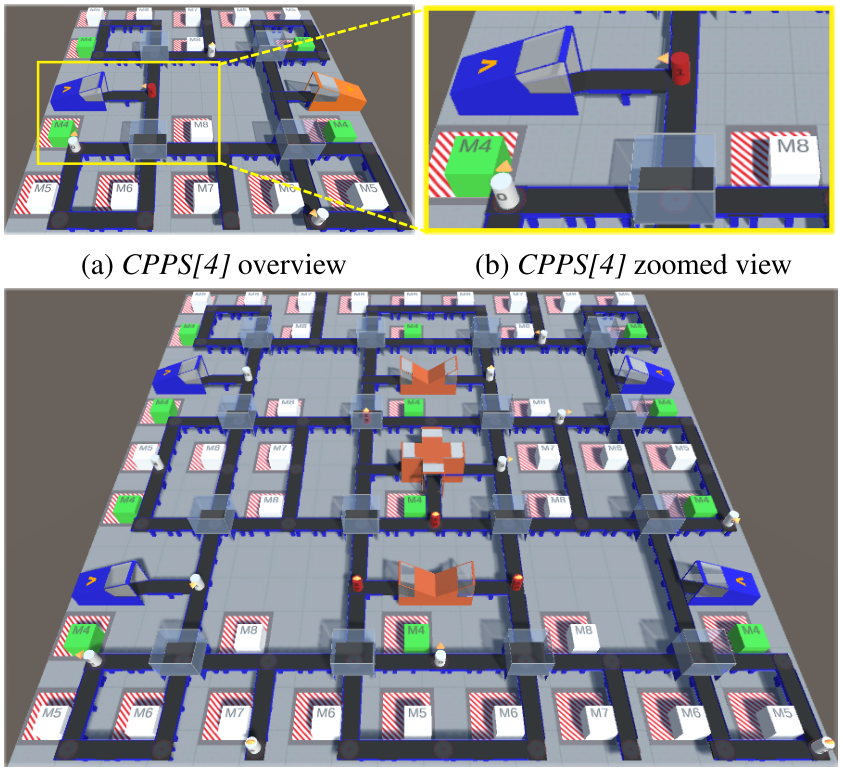


Figure 5.1.: Overview of the components and an example process of ARTS with $R_{ant} = 0.5$ and $N = 4$. Image reproduced from [51].

partial change scenarios depending on R_{ant}. With ARTS, all agents of a target system can be evaluated with opponents or antagonists from *different* training processes to truly assess the ability of a MAS to withstand partial changes. Testing π with $R_{ant} = 0$ corresponds to evaluating the best-case performance similarly to most state-of-the-art works on MARL.
We evaluated ARTS in a simulated *cyber physical production system (CPPS)* as shown in Figure 5.2a using *DQN(R_{ant})*, *QMIX(R_{ant})*, and *QMixMax* as *QMIX($\frac{|\mathcal{D}|-1}{|\mathcal{D}|}$)*. The completion rate of produced items is used as performance measure. Our results in Figure 5.3a show that antagonist-based training with $R_{ant} > 0$ is competitive to idealized MARL w.r.t. best-case performance for adequately chosen R_{ant}. Figure 5.3b shows the results of a cross-validation, where protagonists and antagonists from independent *QMIX(R_{ant})* training processes with different R_{ant} are evaluated in random mixed-games using ARTS. We can see a tradeoff between resilience and best-case performance, where training with large R_{ant} leads to better resilience against more antagonists but rather poor performance in actually cooperative settings. Training with $R_{ant} > 0$ always improves resilience compared to idealized cooperative MARL except *QMixMax*, which focuses on extreme cases, where a single protagonist faces $N - 1$ antagonists and always performs poorly. However, the antagonists of *QMixMax* are able to easily uncover flaws in MAS therefore being well suited for testing [51].

5.2. Adversarial Value Decomposition

Adversarial MARL approaches like ARTS often focus on specialized settings, assuming adversaries with a fixed strength defined by a hyperparameter like R_{ant} [32, 51]. The resulting performance and resilience therefore depends on the choice of R_{ant}, which must be either known a priori or extensively tuned.



(a) *CPPS[4]* overview (b) *CPPS[4]* zoomed view

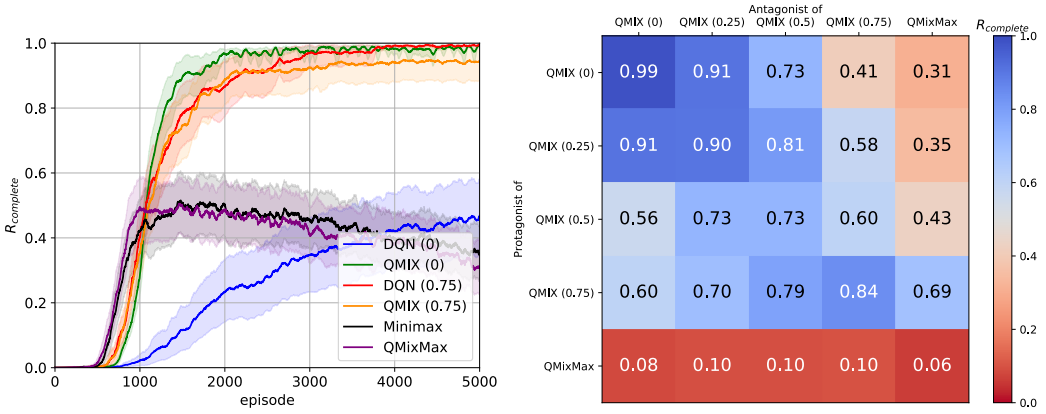
(c) *CPPS[16]* overview

Figure 5.2.: Two CPPS instances with $R_{ant} = \frac{1}{4}$. The white and red cylinders represent protagonists and antagonists respectively. (a, b) *CPPS[4]* as 4-agent setting with 1 antagonist. (c) *CPPS[16]* as 16-agent setting with 4 antagonists. Image taken from [47].

Having a fixed R_{ant} can lead to inflexible policies, which either perform well in purely cooperative settings or in mixed-games with many antagonists, but rarely in both. Furthermore, R_{ant} may affect the training quality, e.g., if R_{ant} is too large, the sampled mixed-games become too difficult to learn any meaningful protagonist policy as indictated by *QMixMax* in Figure 5.3.

To this end, we propose *Resilient Adversarial value Decomposition with Antagonist-Ratios (RADAR)* to flexibly train protagonists and antagonists with variable R_{ant} hence overcoming the limitations of ARTS [47]. RADAR is a CTDE approach based on the same setting as ARTS, assuming a cooperative target system M_C, which is trained via random mixed-games M_X using protagonist and antagonist policy pools.

In RADAR, $R_{ant} \in [0, 1)$ is iteratively sampled from a uniform distribution during training thus enabling training with mixed-games of varying number of protagonists and antagonists. RADAR uses VDN-based value function factorization, where two value functions $\hat{Q}_{pro}$ and $\hat{Q}_{ant}$ are approximated for protagonists and antagonists respectively:



(a) *Learning progress* (b) *Cross-validation*

Figure 5.3.: (a) Learning progress of 50 runs of DQN and QMIX using ARTS, Minimax, and QMixMax in CPPS[4]. (b) Cross-validation of protagonists and antagonists trained with different antagonist-ratios R_{ant} using QMIX or QMixMax. Results taken from [51].

$$\hat{Q}_{pro}(\tau_{t,pro}, a_{t,pro}) = \sum_{i \in \mathcal{D}_{pro}} \hat{Q}_{i,pro}(\tau_{t,i}, a_{t,i}) \quad = \mathbb{E}_{\pi,b_0,\mathcal{P}}\left[\frac{|\mathcal{D}_{pro}|}{N} G_{t,pro} \middle| \tau_t, a_t\right] \tag{5.1}$$

$$\hat{Q}_{ant}(\tau_{t,ant}, a_{t,ant}) = \sum_{j \in \mathcal{D}_{ant}} \hat{Q}_{j,ant}(\tau_{t,j}, a_{t,j}) \quad = -\mathbb{E}_{\pi,b_0,\mathcal{P}}\left[\frac{|\mathcal{D}_{pro}|}{N} G_{t,pro} \middle| \tau_t, a_t\right] \tag{5.2}$$

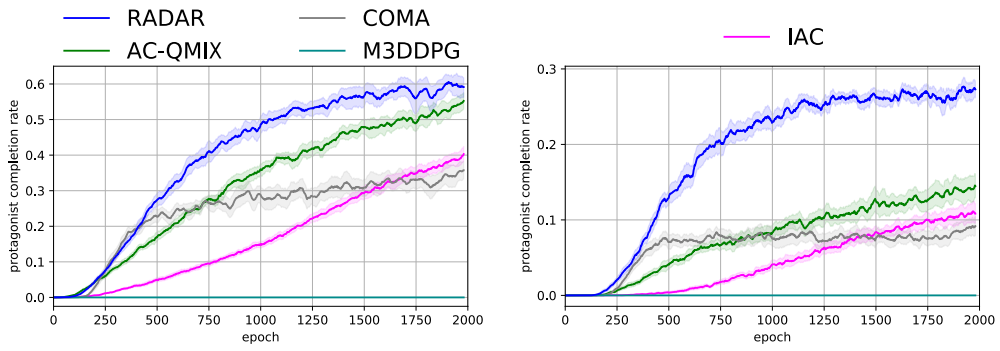
where $G_{t,pro}$ is the return of protagonist rewards $r_{t,pro}$ and $\frac{|\mathcal{D}_{pro}|}{N}$ is used to normalize $G_{t,pro}$ w.r.t. the current number of participating protagonists as the scale of $G_{t,pro}$ otherwise gives more weight to settings, where R_{ant} is small.
$\hat{Q}_{pro}$ and $\hat{Q}_{ant}$ can be approximated via end-to-end learning using backpropagation on $\langle \hat{Q}_{i,pro} \rangle_{i \in \mathcal{D}_{pro}}$ and $\langle \hat{Q}_{j,ant} \rangle_{j \in \mathcal{D}_{ant}}$ respectively [75]. Local policies can be derived from $\hat{Q}_{i,pro}$ and $\hat{Q}_{j,ant}$ using, e.g., multi-armed bandits or actor-critic methods. The striking simplicity of VDN compared to alternative non-linear factorization methods is beneficial when training teams of variable sizes as mentioned in Section 4.2.
Many works on MARL simply evaluate the best-case using only the exact same agents as seen during training. Hence, it is unclear if the MAS would still behave as intended if some agents were replaced, e.g., by "newer versions" in case of an update, or by adversarially behaving agents in case of manipulation or flaws. Therefore, in addition to RADAR, we propose an *agent-based test scheme* to evaluate performance and resilience in MAS in a fair way [47]. Before training the target system, we prepare some *test suites* $\mathcal{T}_{cooperation}$ and $\mathcal{T}_{failure,R_{ant}}$ which consist of pretrained protagonist or antagonist agents, e.g.,

created by domain experts or any kind of adversarial MARL like ARTS or RADAR. During training, the target system can be tested with $\mathcal{T}_{cooperation}$ and $\mathcal{T}_{failure,R_{ant}}$ and some normalized performance measure $\overline{g}$ like the average return or some domain specific value to evaluate the following measures:

Cooperation performance which estimates $\mathbb{E}_{\mathcal{T}_{cooperation}}[\overline{g}]$ to assess the ability of a MAS to collaborate with new agents. Note that unlike in the best-case, protagonists from *different* training processes than the target system are used as $\hat{\pi}_{ant}$ for testing.

Worst case performance which estimates $min_{c \in \mathcal{T}_{cooperation} \cup (\bigcup_{\chi} \mathcal{T}_{failure,\chi})}\{\overline{g}\}$ to assess the resilience of a MAS against a variety of partial changes.

With $\mathcal{T}_{cooperation}$ and $\mathcal{T}_{failure,R_{ant}}$, we can compare the cooperation and worst-case performance of different MARL approaches in a fair way because the trained MAS are evaluated (but not trained) against the *same* set of test agents. Hence, the agents of a MAS can no longer rely on the assumption of only meeting fellow agents from training during evaluation.
We evaluated RADAR in a predator-prey domain and the simulated CPPS shown in Figure 5.2c. The CPPS results are shown in Figure 5.4 and demonstrate the effectiveness of RADAR in terms of cooperation and worst-case performance, where all approaches are evaluated with the same agent-based test suites. Ablation studies show the flexibility and superiority of RADAR compared to variants with fixed R_{ant} and without VDN for adversarial value decomposition [47].



(a) cooperation performance (b) worst-case performance

Figure 5.4.: Cooperation performance and worst-case performance of RADAR and state-of-the-art MARL for *CPPS[16]*. Results taken from [47].

6. Further Topics

> The tradeoff between complexity and optimality is very common in real life.
>
> *Modeling Bounded Rationality*
> Ariel Rubinstein

6.1. Resource Efficiency

The increasing complexity of potential AI domains often leads to more complex algorithms which require tremendous effort on tuning and execution [24]. Such developments can limit actual progress in the long run due to the physical limits regarding compute and memory, especially in low resource environments. Furthermore, increasing demand for computational and memory resources has potential negative societal impacts, e.g., regarding energy consumption [88].
We focus on model-based decision making in particular, where *Monte Carlo planning (MCP)* seems promising to solve complex problems via statistical sampling [27, 71, 86]. MCTS represents the current state-of-the-art in MCP, which constructs sparse closed-loop trees over (belief) states and actions for efficient policy search. Despite avoiding exhaustive search, the closed-loop trees can still become arbitrarily large for highly complex domains with large branching factors, e.g., due to large (joint) action spaces. In environments with highly restricted memory resources, MCTS could become either infeasible or perform very poorly – regardless of the provided computation time.
As mentioned in Section 4.1, open-loop planning can significantly reduce the branching factor by discarding (belief) state information and merely focusing on action sequences [29, 44, 86]. Especially in low resource environments, open-loop planning seems promising to find good solutions in a much more scalable way than closed-loop planning. Focusing on games with $N = 1$[1], we propose memory bounded open-loop MCP algorithms using stacks of Thompson Sampling bandits. Each bandit represents a decision rule for a particular time step t regardless of the actual simulated state, observation, or preceding actions, in order to generate and evaluate open-loop plans [48, 50]. These bandits are updated recursively according to the corresponding returns G_t. The stack size

[1] Such games are known as *games against nature* or POMDPs, where a single decision making agent faces the environment as "stationary opponent" [43].

can be either fixed or dynamically adapted, depending on the convergence of bandits.
Our stack-based algorithms are highly efficient w.r.t. compute and memory being competitive against tree-based (open- and closed-loop) MCP in domains with large action spaces and significantly outperform tree-based MCP when memory resources are restricted. Our approaches can be used in hybrid methods like STEP from Section 4.1 to enable resource efficient planning supported by emergent policy and value functions.

6.2. State Uncertainty

Decentralized partially observable Markov decision processes (Dec-POMDP) are more general variants of cooperative games, where the observation function Ω is stochastic to model noisy sensors [25, 41]. Dec-POMDPs often exhibit high degrees of state uncertainty, which pose a major challenge for decentralized coordination [42]. *State-based CTDE* tackles Dec-POMDPs by exploiting state information to learn a centralized value function in order to derive coordinated local policies [15, 53, 60]. Due to its empirical effectiveness in the *StarCraft Multi-Agent Challenge (SMAC)* as the current de facto standard for MARL evaluation [63], state-based CTDE has become very popular and is widely considered an adequate approach to general Dec-POMDPs [15, 36, 37].
However, state-based CTDE is generally insufficient to learn proper value functions under state uncertainty, since all agents make decisions on a completely different basis, i.e., individual histories of noisy observations and actions. Furthermore, SMAC has very limited state uncertainty due to deterministic observations and low variance in initial states, being insufficient for assessing practicability of MARL. Thus, merely relying on state-based CTDE and SMAC in MARL research can be a pitfall in practice as state uncertainty is largely neglected – despite being an important aspect in Dec-POMDPs.
To this end, we propose *Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL)*, which replaces the true state with a learned representation of multi-agent recurrence as illustrated in Figure 6.1 [52]. By leveraging the memory representations of all agents' recurrent functions, AERIAL considers more accurate closed-loop information about decentralized agent decisions than state-based CTDE to learn proper value functions under state uncertainty [9, 20, 21].
We also introduce MessySMAC which extends SMAC with *observation stochasticity* w.r.t. Ω, where all measured values of observation $z_{t,i}$ are negated with a probability of $\phi \in [0, 1)$, and *initialization stochasticity* w.r.t. b_0, where K random steps are initially performed before officially starting an episode [52]. MessySMAC represents a more general Dec-POMDP benchmark than SMAC, enabling systematic evaluation under different configurations of state uncertainty according to ϕ and K. Figure 6.2 shows the PCA visualization of joint observations in two SMAC and MessySMAC maps within the first 5 steps of

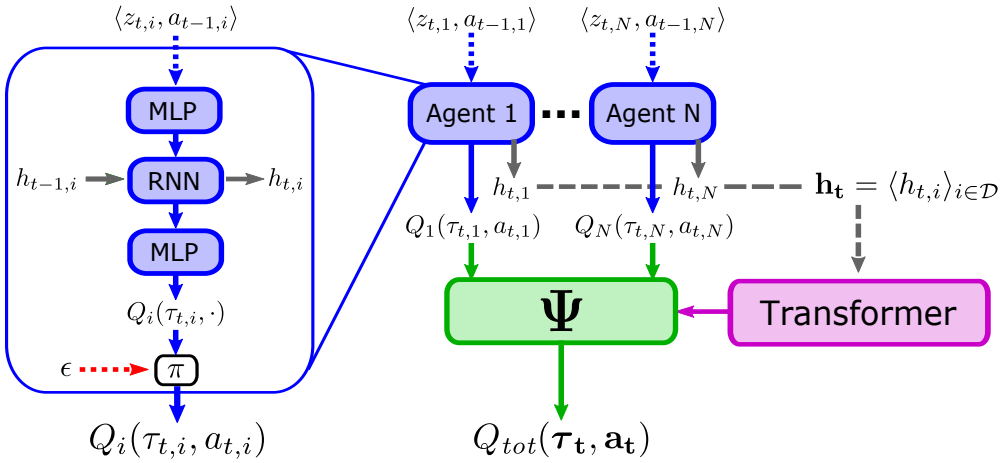


Figure 6.1.: Illustration of the AERIAL setup. *Left:* Recurrent agent network structure with memory representations $h_{t-1,i}$ and $h_{t,i}$. *Right:* Value function factorization via factorization operator Ψ using the joint memory representation $\mathbf{h_t} = \langle h_{t,i} \rangle_{i \in \mathcal{D}}$ of all agents' RNNs instead of true states s_t. All memory representations $h_{t,i}$ are detached from the computation graph to avoid additional differentiation (indicated by the dashed gray arrows) and passed through a simplified transformer before being used by Ψ for value function factorization. Image taken from [52].

1,000 episodes, indicating higher state uncertainty in MessySMAC maps.
Our experiments demonstrate that AERIAL is competitive against state-based CTDE in SMAC and superior in MessySMAC, confirming the effectiveness of exploiting multi-agent recurrence instead of true states in centralized training. Considering state uncertainty in MARL is important to avoid emergence of uncoordinated behavior and to improve resilience as the real world is generally messy and only observable through noisy sensors [25, 41, 71].

6.3. Non-Cooperative Emergence

In Chapter 3, we studied how emergent cooperation can be incentivized in general-sum games using local communication. In Chapter 4, we demonstrated how coordination can be improved by exploiting emergent properties in CTDE. In both chapters, the resulting emergent behavior improves social welfare over the course of training. In the real world, not all emergent phenomena improve social welfare though. Traffic jams or panic buying are examples of *non-cooperative emergence*, where individuals act greedily regardless of others leading to worse social welfare than is possible.
In our study, we conducted predator-prey simulations of N learning agents representing fishes as prey in a general-sum game along with a shark as predator, which greedily hunts fishes within close vicinity [18]. The goal of all agents

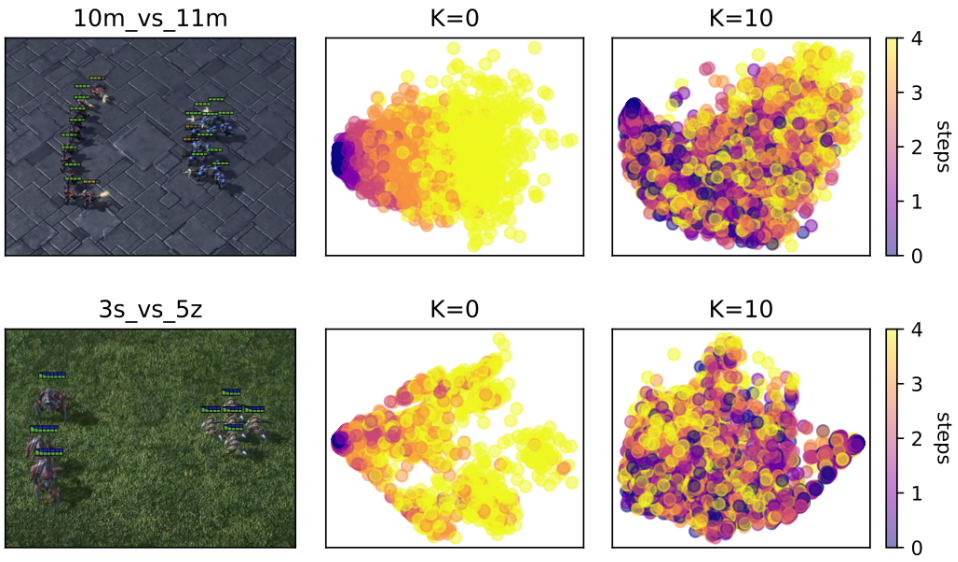


Figure 6.2.: *Left:* Screenshot of two SMAC maps. *Middle:* PCA visualization of the joint observations in original SMAC within the first 5 steps of 1,000 episodes using a random policy with $K = 0$ initial random steps. *Right:* Analogous PCA visualization for MessySMAC with $K = 10$ initial random steps. For visual comparability, the observations are deterministic here. Image taken from [52].

is to survive regardless of others which is individually rewarded with +1 for each time step. However, an agent gets penalized with −1000 and removed from the game when being caught by the shark.
As expected from nature, the agents learned to form swarms – despite not being explicitly rewarded to do so – effectively distracting the shark, which consequently focused on hunting "outlier" agents instead. At first glance, one might assume the resulting swarm to be an emergent cooperative structure. However, further analysis revealed that swarm forming policies are rather inefficient w.r.t. the expected agent life time, which is a surprising finding given that agents are actually rewarded for surviving [18]. We found that if all agents simply escaped from the shark independently of each other, their expected life time would increase significantly. However if a swarm emerges, then independent escaping agents tend to be caught first, causing the swarm to occasionally "sacrifice" some agents by isolating them as bait for the shark. Therefore, the whole swarm can be hypothesized to be an emergent selfish but stable structure which has a long expected life time "as a whole" while its individuals are rather short lived. Extending such analyses to more domains could provide valuable insights to devise better MARL algorithms that consider emergence and resilience.

7. Conclusion

> Fortunately, humans do have foresight and use it to speed up what would otherwise be a blind process of evolution.
>
> *The Evolution of Cooperation*
> Robert Axelrod

7.1. Summary

Due to the rapid progress in MARL, artificial agents are going to become more and more omnipresent in our everyday life, influencing our world through their decisions like natural MAS already do [14, 17]. In this thesis, we investigated the explicit consideration of emergence and resilience in MARL and presented methods that overcome common limitations of the state-of-the-art regarding various aspects relevant to the real world. Our work is based on common MAS settings and paradigms as described in Chapter 2.
We first studied how cooperation can emerge in general-sum games from local interaction. Agents are able to incentivize each other to adopt more cooperative policies via locally exchanged rewards using a two-phase communication protocol. To avoid destabilization of cooperation by social pressure, agents need the ability to reciprocate via adequate responses. The locality of information and communication naturally offers some degree of resilience against external disturbances like random protocol defections and communication failures, which are common challenges in the real world [79].
While the emergence of cooperation in general-sum games is desirable for potential future multi-agent scenarios, many distributed real-world applications can be straightforwardly modeled as cooperative games, where all agents share a common goal and are trainable via CTDE. Agents of a cooperative game need to make individual decisions to not just achieve emergent cooperation but to improve overall coordination. Despite state-of-the-art CTDE commonly using states and joint actions to learn coordinated policies, we hypothesized that emergent phenomena are not sufficiently considered during training therefore limiting performance and scalability. To this end, we proposed MARL approaches that explicitly consider emergence in form of globally tracked models for prediction and dynamic team structures for hierarchization to overcome these limitations.

MAS are supposed to withstand internal disturbances or partial changes, where agents themselves can unexpectedly alter their policies due to updates, manipulation, or flaws. Even if the underlying setting is purely cooperative, a MAS should *always* be prepared for partial changes and exhibit resilience similar to natural MAS and classic distributed systems in order to prevent catastrophic failure. We therefore reformulated originally cooperative games as zero-sum games of agent teams and devised antagonist-based methods, where all agents can potentially change their behavior adversarially, to learn resilient policies. We also proposed testing methods to evaluate resilience w.r.t. partial changes based on dedicated agent test sets.
At last, we briefly presented further topics that are also relevant to building artificial MAS. We addressed the problem of resource efficiency, which is important regarding low resource environments and societal impact, by proposing memory bounded open-loop MCP for effective decision making in a computationally and memory efficient way. As state uncertainty is largely neglected in MARL research, we introduced a configurable benchmark for adequate evaluation as well as a recurrence-based approach to MARL which implicitly considers state uncertainty to advance MARL towards more general settings. We also investigated non-cooperative emergence in form of selfish swarms, where individual agents are occasionally "sacrificed" by the swarm to ensure its survival. A better understanding of such phenomena could be useful to further improve algorithms considering emergence and resilience.
Returning to our research questions formulated in Section 1.2, we can now give the following answers:

(Q1) Can cooperation emerge in MARL from local interaction? In Chapter 3, we proposed a peer incentivization approach based on mutual acknowledgments, where self-interested agents are able to exchange rewards and penalties. Agent interaction is completely local due to partial observability and neighborhood communication therefore offering scalability and better applicability to real-world scenarios. Compared to prior methods, which mostly assume global information, our approach is able to achieve superior cooperation, emerging from local interaction.

(Q2) How to maintain emergent cooperation in MARL? Our mutual acknowledgment approach is able to stably maintain cooperation even under social pressure due to our penalization mechanism which enables reciprocity. In Chapter 3, we additionally presented a resilience evaluation regarding external disturbances like random protocol defections and communication failures. Our approach naturally offers some degree of resilience due to the locality of information and communication as it is able to maintain its superior cooperation in large and partially observable domains in contrast to prior methods that assume global information.

(Q3) How to consider emergence to improve performance in MARL? State-of-the-art CTDE is currently limited to domains with a handful of

agents despite using global information like states and joint actions to learn coordinated policies. In Chapter 4, we hypothesized that explicit consideration of emergence beyond mere states and joint actions is needed to improve overall coordination in large MAS. We first presented a model-based hybrid approach, integrating learned policy and value functions into decentralized open-loop planning to efficiently predict emergent effects for agent-wise policy learning. We then presented a model-free approach which is able to exploit the emergence of dynamic team structures to form an optimized hierarchy for scalable value function factorization. Both approaches are able to outperform state-of-the-art CTDE approaches that do not explicitly consider emergence, especially in large-scale domains with many agents.

(Q4) How to improve resilience in MARL against partial changes? Despite resilience being a main motivation for artificial MAS for decades, many works on MARL focus on optimizing the best-case performance under idealized conditions, assuming no partial changes at all. In Chapter 5, we proposed two antagonist-based methods, where all agents can change their behavior, i.e., through random replacement by adversarial counterparts, during training to learn resilient policies. The resulting MAS outperform idealized MARL in the presence of partial changes, indicating that antagonist-based training can produce resilient MAS that are able to survive internal disturbances.

(Q5) How to evaluate resilience in MARL against partial changes? Since many works on MARL assume idealized conditions, the evaluation of a MAS is typically conducted with the exact same agents as seen during training. In addition to our antagonist-based training methods, we introduce testing methods in Chapter 5 to evaluate resilience w.r.t. partial changes based on dedicated agent test sets. These test sets can be used for cross-validation at the end of training or as pretrained test suites to evaluate the target system during training. By using the same set of test suites or agents to evaluate different MARL approaches, we can ensure a fair comparison w.r.t. cooperation and worst-case performance regarding any kind of partial change.

(Q6) What are further relevant topics? Beside emergence and resilience, there are many other topics that are also relevant to building artificial MAS. In Chapter 6, we briefly addressed resource efficiency, state uncertainty, and non-cooperative emergence. Resource efficiency is important regarding low resource environments and societal impact. Considering state uncertainty is important to avoid emergence of uncoordinated behavior and to improve resilience in a messy world that is only observable through noisy sensors. Studying the emergence of selfish structures in large MAS like swarms could provide a starting point for devising better algorithms considering emergence and resilience.

Based on the findings of our work and the provided answers to the research questions, we conclude that the explicit consideration of emergence and resilience in MARL is important to build artificial MAS that are able to solve complex tasks while withstanding disturbances thus confirming our central hypothesis 1.1. As our world is steadily expanding towards more artificial agents and MAS that coexist with natural MAS, considering emergence and resilience is going to become an increasingly important task.

7.2. Outlook

In this thesis, we assumed all agents to be purely rational. In the real world, where AI systems are supposed to coexist with humans and other living organisms, pure rationality cannot always be assumed (from an artificial agent's perspective) though [62]. Modeling *non-rational* or *bounded rational* agents can help MARL to better align with human behavior and preferences to improve safety, trustworthiness, and acceptance in wide areas. Assuming bounded rationality in agents could also improve resilient MARL, where disturbances can occur due to irrational actions, e.g., human intervention or mistakes, without any adversarial intention.
Social abstraction may be useful to reduce complexity in large MAS with many agents. In Section 3.2 and 4.2, we found that locality and optimized grouping of agents can improve performance and resilience in MAS. However, we only assumed a single level of abstraction, where agents are merely part of some sub-team or neighborhood without further structure or hierarchization. In real life, humans and animals can have multiple levels of abstractions to classify others, e.g., as close relatives, friends, rivals, or completely unrelated individuals. These abstractions enable efficient prioritization of actions without needing to know each opponent in full detail. Social abstraction in MARL could improve cooperation and resilience in large MAS by integrating emergent social structures, e.g., to cooperate based on social relationships or to adversarially attack important relationships in order to evaluate resilience.
As mentioned in Section 6.1, the complexity of algorithms w.r.t. hyperparameters and mechanisms tends to increase as AI advances towards more complex domains. Increasing complexity can be prohibitive for actual progress because more effort is required for implementation, execution, tuning, and understanding, which has potential negative societal impacts. Therefore, addressing complexity with *simplicity* is a more sensible goal than vice versa. In our research, we put some effort in introducing rather simple methods with very few additional hyperparameters that are able to overcome limitations of more complex state-of-the-art methods. However, future work should actually aim to *reduce* the number of hyperparameters instead of introducing new ones [24].

Bibliography

[1] Christopher Amato and Frans Oliehoek. "Scalable Planning and Learning for Multiagent POMDPs". *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), February 2015. doi:https://doi.org/10.1609/aaai.v29i1.9439.

[2] Thomas Anthony, Zheng Tian, and David Barber. "Thinking Fast and Slow with Deep Learning and Tree Search". In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. doi:https://dl.acm.org/doi/abs/10.5555/3295222.3295288.

[3] Robert Axelrod. *"The Evolution Of Cooperation"*. Basic Books, 1984. doi:https://doi.org/10.1007/978-3-319-16999-6_1220-1.

[4] Robert Axelrod and William D. Hamilton. "The Evolution of Cooperation". *Science*, 211(4489):1390–1396, 1981. doi:https://doi.org/10.1126/science.7466396.

[5] Wendelin Boehmer, Vitaly Kurin, and Shimon Whiteson. "Deep Coordination Graphs". In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 980–991. PMLR, 13–18 July 2020. doi:https://dl.acm.org/doi/abs/10.5555/3524938.3525030.

[6] Craig Boutilier. "Planning, Learning and Coordination in Multiagent Decision Processes". In *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '96, pages 195–210, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. doi:https://dl.acm.org/doi/10.5555/1029693.1029710.

[7] Robert Boyd and Peter J. Richerson. "The Evolution of Indirect Reciprocity". *Social networks*, 11(3):213–236, 1989. doi:https://doi.org/10.1016/0378-8733(89)90003-8.

[8] Yu-Han Chang, Tracey Ho, and Leslie P Kaelbling. "All Learning is Local: Multi-Agent Learning in Global Reward Games". In *Advances in Neural Information Processing Systems*, pages 807–814, 2004. doi:https://dl.acm.org/doi/10.5555/2981345.2981446.

[9] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". In *Proceedings of SSST-8, 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

[10] Daniel Claes, Frans Oliehoek, Hendrik Baier, and Karl Tuyls. "Decentralised Online Planning for Multi-Robot Warehouse Commissioning". In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, pages 492–500. International Foundation for Autonomous Agents and Multiagent Systems, 2017. doi:https://dl.acm.org/doi/abs/10.5555/3091125.3091198.

[11] S. Marc Cohen and C. D. C. Reeve. "Aristotle's Metaphysics". In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[12] R. Dawkins. *"The Selfish Gene: 40th Anniversary Edition"*. Oxford Landmark Science. OUP Oxford, 2016.

[13] Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. "Approximate Solutions for Partially Observable Stochastic Games with Common Payoffs". AAMAS '04, pages 136–143, USA, 2004. IEEE Computer Society. doi:https://dl.acm.org/doi/10.5555/1018409.1018741.

[14] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. "Learning with Opponent-Learning Awareness". In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '18, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018. doi:https://dl.acm.org/doi/10.5555/3237383.3237408.

[15] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. "Counterfactual Multi-Agent Policy Gradients". *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi:https://doi.org/10.1609/aaai.v32i1.11794.

[16] Thomas Gabor, Andreas Sedlmeier, Thomy Phan, Fabian Ritz, Marie Kiermeier, Lenz Belzner, Bernhard Kempter, Cornel Klein, Horst Sauer, Reiner Schmid, Jan Wieghardt, Marc Zeller, and Claudia Linnhoff-Popien. "The Scenario Coevolution Paradigm: Adaptive Quality Assurance for Adaptive Systems". *International Journal on Software Tools for Technology Transfer*, 22(4):457–476, 2020. doi:https://doi.org/10.1007/s10009-020-00560-5.

[17] Balint Gucsi, Danesh S. Tarapore, William Yeoh, Christopher Amato, and Long Tran-Thanh. "To Ask or Not to Ask: A User Annoyance Aware Preference Elicitation Framework for Social Robots". In

2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 7935–7940, 2020. doi:https://doi.org/10.1109/IROS45743.2020.9341607.

[18] Carsten Hahn, Thomy Phan, Thomas Gabor, Lenz Belzner, and Claudia Linnhoff-Popien. "Emergent Escape-Based Flocking Behavior using Multi-Agent Reinforcement Learning". In *The 2021 Conference on Artificial Life*, ALIFE 2021, pages 598–605, July 2019. doi:https://doi.org/10.1162/isal_a_00226.

[19] Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. "Dynamic Programming for Partially Observable Stochastic Games". In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 709–715. AAAI Press, 2004. doi:https://dl.acm.org/doi/abs/10.5555/1597148.1597262.

[20] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". *Neural Computation*, 9(8):1735–1780, 1997. doi:https://doi.org/10.1162/neco.1997.9.8.1735.

[21] Hengyuan Hu and Jakob N Foerster. "Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning". In *International Conference on Learning Representations*, 2019.

[22] Matthew O. Jackson and Simon Wilkie. "Endogenous Games and Mechanisms: Side Payments Among Players". *The Review of Economic Studies*, 72(2):543–566, 04 2005. doi:https://doi.org/10.1111/j.1467-937X.2005.00342.x.

[23] Steven Johnson. *"Emergence: The Connected Lives of Ants, Brains, Cities, and Software"*. Scribner, 2002.

[24] Scott Jordan, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip Thomas. "Evaluating the Performance of Reinforcement Learning Algorithms". In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4962–4973. PMLR, 2020. doi:https://dl.acm.org/doi/abs/10.5555/3524938.3525399.

[25] Leslie Pack Kaelbling, Michael L Littman, and Anthony R. Cassandra. "Planning and Acting in Partially Observable Stochastic Domains". *Artificial intelligence*, 101(1-2):99–134, 1998. doi:https://doi.org/10.1016/S0004-3702(98)00023-X.

[26] Daniel Kahneman. *"Thinking, Fast and Slow"*. Farrar, Straus and Giroux, New York, 2011. doi:https://doi.org/10.1007/s00362-013-0533-y.

[27] Levente Kocsis and Csaba Szepesvári. "Bandit Based Monte-Carlo Planning". In *European Conference on Machine Learning*, pages 282–293. Springer, 2006. doi:https://doi.org/10.1007/11871842_29.

[28] Guillaume J. Laurent, Laëtitia Matignon, and Le Fort-Piat. "The World of Independent Learners is Not Markovian". *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64, 2011. doi:https://dl.acm.org/doi/10.5555/1971886.1971887.

[29] Erwan Lecarpentier, Guillaume Infantes, Charles Lesire, and Emmanuel Rachelson. "Open Loop Execution of Tree-Search Algorithms". In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2362–2368. IJCAI Organization, July 2018. doi:https://doi.org/10.24963/ijcai.2018/327.

[30] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. "Multi-Agent Reinforcement Learning in Sequential Social Dilemmas". In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, AAMAS '17, pages 464–473. International Foundation for Autonomous Agents and Multiagent Systems, 2017. doi:https://dl.acm.org/doi/10.5555/3091125.3091194.

[31] Jiaoyang Li, Andrew Tinka, Scott Kiesel, Joseph W. Durham, T. K. Satish Kumar, and Sven Koenig. "Lifelong Multi-Agent Path Finding in Large-Scale Warehouses". *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11272–11281, May 2021. doi:https://doi.org/10.1609/aaai.v35i13.17344.

[32] Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. "Robust Multi-Agent Reinforcement Learning via Minimax Deep Deterministic Policy Gradient". In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4213–4220, 2019. doi:https://doi.org/10.1609/aaai.v33i01.33014213.

[33] Michael L. Littman. "Markov Games as a Framework for Multi-Agent Reinforcement Learning". In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 157–163. Morgan Kaufmann, San Francisco (CA), 1994. doi:https://doi.org/10.1016/B978-1-55860-335-6.50027-1.

[34] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments". In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017. doi:https://dl.acm.org/doi/abs/10.5555/3295222.3295385.

[35] Andrei Lupu and Doina Precup. "Gifting in Multi-Agent Reinforcement Learning". In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, pages 789–797, 2020. doi:https://dl.acm.org/doi/abs/10.5555/3398761.3398855.

[36] Xueguang Lyu, Andrea Baisero, Yuchen Xiao, and Christopher Amato. "A Deeper Understanding of State-Based Critics in Multi-Agent Reinforcement Learning". *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9396–9404, Jun. 2022. doi:https://doi.org/10.1609/aaai.v36i9.21171.

[37] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. "Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning". In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, pages 844–852, 2021. doi:https://dl.acm.org/doi/abs/10.5555/3463952.3464053.

[38] Alexander Mey and Frans A. Oliehoek. "Environment Shift Games: Are Multiple Agents the Solution, and not the Problem to Non-Stationarity?". AAMAS '21, pages 23–27. International Foundation for Autonomous Agents and Multiagent Systems, 2021. doi:https://dl.acm.org/doi/10.5555/3463952.3463958.

[39] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. "Human-Level Control through Deep Reinforcement Learning". *Nature*, 518(7540):529–533, 2015. doi:https://doi.org/10.1038/nature14236.

[40] Ranjit Nair, Milind Tambe, Makoto Yokoo, David Pynadath, and Stacy Marsella. "Taming Decentralized POMDPs: Towards Efficient Policy Computation for Multiagent Settings". In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 705–711, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc. doi:https://dl.acm.org/doi/abs/10.5555/1630659.1630762.

[41] Frans A. Oliehoek and Christopher Amato. *"A Concise Introduction to Decentralized POMDPs"*. Springer, 2016. doi:https://doi.org/10.1007/978-3-319-28929-8.

[42] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. "Optimal and Approximate Q-Value Functions for Decentralized POMDPs". *Journal of Artificial Intelligence Research*, 32:289–353, 2008. doi:https://doi.org/10.1613/jair.2447.

[43] Christos H. Papadimitriou. "Games against Nature". *Journal of Computer and System Sciences*, 31(2):288–301, 1985. doi:https://doi.org/10.1016/0022-0000(85)90045-5.

[44] Diego Perez Liebana, Jens Dieskau, Martin Hunermund, Sanaz Mostaghim, and Simon Lucas. "Open Loop Search for General Video Game Playing". In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 337–344. ACM, 2015. doi:https://dl.acm.org/doi/10.1145/2739480.2754811.

[45] Julien Perolat, Joel Z. Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. "A Multi-Agent Reinforcement Learning Model of Common-Pool Resource Appropriation". In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3646–3655, Red Hook, NY, USA, 2017. Curran Associates Inc. doi:https://dl.acm.org/doi/abs/10.5555/3294996.3295122.

[46] Thomy Phan, Lenz Belzner, Thomas Gabor, and Kyrill Schmid. "Leveraging Statistical Multi-Agent Online Planning with Emergent Value Function Approximation". In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 730–738. International Foundation for Autonomous Agents and Multiagent Systems, 2018. doi:https://dl.acm.org/doi/10.5555/3237383.3237491.

[47] Thomy Phan, Lenz Belzner, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, and Claudia Linnhoff-Popien. "Resilient Multi-Agent Reinforcement Learning with Adversarial Value Decomposition". *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11308–11316, May 2021. doi:https://doi.org/10.1609/aaai.v35i13.17348.

[48] Thomy Phan, Lenz Belzner, Marie Kiermeier, Markus Friedrich, Kyrill Schmid, and Claudia Linnhoff-Popien. "Memory Bounded Open-Loop Planning in Large POMDPs using Thompson Sampling". *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7941–7948, July 2019. doi:https://doi.org/10.1609/aaai.v33i01.33017941.

[49] Thomy Phan, Lenz Belzner, Kyrill Schmid, Thomas Gabor, Fabian Ritz, Sebastian Feld, and Claudia Linnhoff-Popien. "A Distributed Policy Iteration Scheme for Cooperative Multi-Agent Policy Approximation". In *12th Adaptive and Learning Agents Workshop*, ALA '20, 2020.

[50] Thomy Phan, Thomas Gabor, Robert Müller, Christoph Roch, and Claudia Linnhoff-Popien. "Adaptive Thompson Sampling Stacks for Memory Bounded Open-Loop Planning". In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5607–5613.

International Joint Conferences on Artificial Intelligence Organization, July 2019. doi:https://doi.org/10.24963/ijcai.2019/778.

[51] Thomy Phan, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, Bernhard Kempter, Cornel Klein, Horst Sauer, Reiner Schmid, Jan Wieghardt, Marc Zeller, and Claudia Linnhoff-Popien. "Learning and Testing Resilience in Cooperative Multi-Agent Systems". In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1055–1063. International Foundation for Autonomous Agents and Multiagent Systems, 2020. doi:https://dl.acm.org/doi/10.5555/3398761.3398884.

[52] Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023. To appear.

[53] Thomy Phan, Fabian Ritz, Lenz Belzner, Philipp Altmann, Thomas Gabor, and Claudia Linnhoff-Popien. "VAST: Value Function Factorization with Variable Agent Sub-Teams". In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24018–24032. Curran Associates, Inc., 2021.

[54] Thomy Phan, Kyrill Schmid, Lenz Belzner, Thomas Gabor, Sebastian Feld, and Claudia Linnhoff-Popien. "Distributed Policy Iteration for Scalable Approximation of Cooperative Multi-Agent Policies". In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 2162–2164. International Foundation for Autonomous Agents and Multiagent Systems, 2019. doi:https://dl.acm.org/doi/10.5555/3306127.3332044.

[55] Thomy Phan, Felix Sommer, Philipp Altmann, Fabian Ritz, Lenz Belzner, and Claudia Linnhoff-Popien. "Emergent Cooperation from Mutual Acknowledgment Exchange". In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '22, pages 1047–1055. International Foundation for Autonomous Agents and Multiagent Systems, 2022. doi:https://dl.acm.org/doi/10.5555/3535850.3535967.

[56] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. "Robust Adversarial Reinforcement Learning". In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learn-*

ing Research, pages 2817–2826. PMLR, 06–11 August 2017. doi:https://dl.acm.org/doi/10.5555/3305890.3305972.

[57] David Premack and Guy Woodruff. "Does the Chimpanzee Have a Theory of Mind?". *Behavioral and brain sciences*, 1(4):515–526, 1978. doi:https://doi.org/10.1017/S0140525X00076512.

[58] Martin L. Puterman. *"Markov Decision Processes: Discrete Stochastic Dynamic Programming"*. Wiley Series in Probability and Statistics. Wiley, 2014. doi:https://dx.doi.org/10.1002/9780470316887.

[59] Anatol Rapoport. "Prisoner's Dilemma — Recollections and Observations". In *Game Theory as a Theory of a Conflict Resolution*, pages 17–34. Springer, 1974. doi:https://doi.org/10.1007/978-94-010-2161-6_2.

[60] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. "QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning". In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4295–4304. PMLR, 2018.

[61] Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-Draa. "Online Planning Algorithms for POMDPs". *Journal of Artificial Intelligence Research*, 32:663–704, 2008. doi:https://doi.org/10.1613/jair.2567.

[62] Ariel Rubinstein. *"Modeling Bounded Rationality"*. The MIT Press, 1998. doi:https://doi.org/10.1023/A:1008058417088.

[63] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. "The StarCraft Multi-Agent Challenge". In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '19, pages 2186–2188. International Foundation for Autonomous Agents and Multiagent Systems, 2019. doi:https://dl.acm.org/doi/abs/10.5555/3306127.3332052.

[64] Kyrill Schmid, Lenz Belzner, Robert Müller, Johannes Tochtermann, and Claudia Linnhoff-Popien. "Stochastic Market Games". In Zhi-Hua Zhou, editor, *Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 384–390. International Joint Conferences on Artificial Intelligence Organization, August 2021. doi:https://doi.org/10.24963/ijcai.2021/54.

[65] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model". *Nature*, 588(7839):604–609, 2020. doi:https://doi.org/10.1038/s41586-020-03051-4.

[66] Yoav Shoham and Kevin Leyton-Brown. *"Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations"*. Cambridge University Press, 2008. doi:https://doi.org/10.1017/CBO9780511811654.

[67] Yoav Shoham, Rob Powers, and Trond Grenager. "If Multi-Agent Learning is the Answer, What is the Question?". *Artificial Intelligence*, 171(7):365–377, 2007. Foundations of Multi-Agent Learning. doi:https://doi.org/10.1016/j.artint.2006.02.006.

[68] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. "Mastering the Game of Go with Deep Neural Networks and Tree Search". *Nature*, 529(7587):484–489, 2016. doi:https://doi.org/10.1038/nature16961.

[69] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. "A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-Play". *Science*, 362(6419):1140–1144, 2018. doi:https://doi.org/10.1126/science.aar6404.

[70] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. "Mastering the Game of Go without Human Knowledge". *Nature*, 550(7676):354–359, 2017. doi:https://doi.org/10.1038/nature24270.

[71] David Silver and Joel Veness. "Monte-Carlo Planning in Large POMDPs". In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. doi:https://dl.acm.org/doi/abs/10.5555/2997046.2997137.

[72] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. "QTRAN: Learning to Factorize with Transformation for Co-

operative Multi-Agent Reinforcement Learning". In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5887–5896. PMLR, 09–15 Jun 2019.

[73] Peter Stone and Michael L. Littman. "Implicit Negotiation in Repeated Games". In *Pre-proceedings of the 8th International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, pages 96–105, 2001.

[74] Peter Stone and Manuela Veloso. "Multiagent Systems: A Survey from a Machine Learning Perspective". *Autonomous Robots*, 8(3):345–383, July 2000. doi:https://doi.org/10.1023/A:1008942012299.

[75] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and Thore Graepel. "Value-Decomposition Networks for Cooperative Multi-Agent Learning based on Team Reward". In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (Extended Abstract)*, AAMAS '18, pages 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems, 2018. doi:https://dl.acm.org/doi/10.5555/3237383.3238080.

[76] Richard S. Sutton and Andrew G. Barto. *"Reinforcement Learning: An Introduction"*. A Bradford Book, Cambridge, MA, USA, 2018.

[77] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063. MIT Press, 2000. doi:https://dl.acm.org/doi/abs/10.5555/3009657.3009806.

[78] Ming Tan. "Multi-Agent Reinforcement Learning: Independent versus Cooperative Agents". In *Proceedings of the 10th International Conference on International Conference on Machine Learning*, pages 330–337. Morgan Kaufmann Publishers Inc., 1993. doi:https://dl.acm.org/doi/10.5555/3091529.3091572.

[79] Andrew S. Tanenbaum and Maarten Van Steen. *"Distributed Systems: Principles and Paradigms"*. Prentice-Hall, 2007.

[80] Gerald Tesauro. "Temporal Difference Learning and TD-Gammon". *Communications of the ACM*, 38(3):58–68, March 1995. doi:https://doi.org/10.1145/203330.203343.

[81] Robert L. Trivers. "The Evolution of Reciprocal Altruism". *The Quarterly Review of Biology*, 46(1):35–57, 1971. doi:https://doi.org/10.1086/406755.

[82] Jonathan Uesato, Ananya Kumar, Csaba Szepesvari, Tom Erez, Avraham Ruderman, Keith Anderson, Krishmamurthy (Dj) Dvijotham, Nicolas Heess, and Pushmeet Kohli. "Rigorous Agent Evaluation: An Adversarial Approach to Uncover Catastrophic Failures". *International Conference on Learning Representations*, 2019.

[83] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, and Joel Z Leibo. "A Learning Agent that Acquires Social Norms from Public Sanctions in Decentralized Multi-Agent Settings". *arXiv preprint arXiv:2106.09012*, 2021.

[84] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. "Grandmaster Level in StarCraft II using Multi-Agent Reinforcement Learning". *Nature*, pages 1–5, 2019. doi:https://doi.org/10.1038/s41586-019-1724-z.

[85] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. "QPLEX: Duplex Dueling Multi-Agent Q-Learning". In *International Conference on Learning Representations*, 2021.

[86] Ari Weinstein and Michael L Littman. "Open-loop Planning in Large-Scale Stochastic Domains". In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1436–1442. AAAI Press, 2013. doi:https://doi.org/10.1609/aaai.v27i1.8547.

[87] Gerhard Weiß, editor. *"Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence"*. MIT Press, Cambridge, MA, USA, 1999. doi:https://dl.acm.org/doi/10.5555/305606.

[88] Jess Whittlestone, Kai Arulkumaran, and Matthew Crosby. "The Societal Implications of Deep Reinforcement Learning". *Journal of Artificial Intelligence Research*, 70:1003–1030, May 2021. doi:https://doi.org/10.1613/jair.1.12360.

[89] David H Wolpert and Kagan Tumer. "Optimal Payoff Functions for Members of Collectives". In *Modeling Complexity in Economic and Social Systems*, pages 355–369. World Scientific, 2002. doi:https://doi.org/10.1142/9789812777263_0020.

[90] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. "Learning to Incentivize Other Learning Agents". *Advances in Neural Information Processing Systems*, 33, 2020. doi:https://dl.acm.org/doi/abs/10.5555/3495724.3496999.

[91] Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, IMAN FADAKAR, Zheng Chen, Chongxi Huang, Ying Wen, Kimia Hassanzadeh, Daniel Graves, Zhengbang Zhu, Yihan Ni, Nhat Nguyen, Mohamed Elsayed, Haitham Ammar, Alexander Cowen-Rivers, Sanjeevan Ahilan, Zheng Tian, Daniel Palenicek, Kasra Rezaee, Peyman Yadmellat, Kun Shao, dong chen, Baokuan Zhang, Hongbo Zhang, Jianye Hao, Wulong Liu, and Jun Wang. "SMARTS: An Open-Source Scalable Multi-Agent RL Training School for Autonomous Driving". In Jens Kober, Fabio Ramos, and Claire Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 264–285. PMLR, 16–18 Nov 2021.

A. Publications

We list all publications that contain the core findings of the thesis and provide the core contributions as well as the credit in accordance with LMU Munich regulations. We reprint Table 1.1 below as an overview for easier reference.

Thesis chapter	Questions	Publications
3. Emergent Cooperation in General-Sum Games	**Q1**, **Q2**	[55]
4. Emergence in Cooperative Multi-Agent RL	**Q3**	[49, 53, 54]
5. Resilience in Cooperative Multi-Agent RL	**Q4**, **Q5**	[47, 51]
6. Further Topics	**Q6**	[18, 48, 50, 52]

A.1. Emergent Cooperation from Mutual Acknowledgment Exchange

Publication

Thomy Phan, Felix Sommer, Philipp Altmann, Fabian Ritz, Lenz Belzner, and Claudia Linnhoff-Popien. "Emergent Cooperation from Mutual Acknowledgment Exchange". In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1047–1055. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

DOI: `https://dl.acm.org/doi/10.5555/3535850.3535967`
URL: `https://ifaamas.org/Proceedings/aamas2022/pdfs/p1047.pdf`

Contributions

1. A two-phase communication protocol for reciprocal peer incentivization.
2. Robustness evaluation w.r.t. random protocol defectors and communication failures.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Sommer conceived and conducted a preliminary study of a centralized MARL approach based on trust metrics supervised by Phan and Altmann. Ritz supported the robustness evaluation. Altmann, Ritz, and Belzner discussed and reviewed the results. Linnhoff-Popien consulted the process and reviewed the results. This publication is based on Sommer's thesis to achieve the degree Master of Science.

Purpose

Main focus of Chapter 3 regarding the research questions **Q1** and **Q2** from Section 1.2.

A.2. Distributed Policy Iteration for Scalable Approximation of Cooperative Multi-Agent Policies

Publication

Thomy Phan, Kyrill Schmid, Lenz Belzner, Thomas Gabor, Sebastian Feld, and Claudia Linnhoff-Popien. "Distributed Policy Iteration for Scalable Approximation of Cooperative Multi-Agent Policies". In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 2162–2164. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

DOI: `https://dl.acm.org/doi/10.5555/3306127.3332044`
URL: `https://ifaamas.org/Proceedings/aamas2019/pdfs/p2162.pdf`

Contributions

1. A distributed policy iteration scheme combining planning and RL for fully observable multi-agent domains.

2. An open-loop planning algorithm integrating learned multi-agent policies and value functions for efficient search.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Schmid and Belzner discussed the concepts and results, and provided feedback on related work. Gabor and Feld discussed and reviewed the results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 4 regarding the research question **Q3** from Section 1.2.

A.3. A Distributed Policy Iteration Scheme for Cooperative Multi-Agent Policy Approximation

Publication

Thomy Phan, Lenz Belzner, Kyrill Schmid, Thomas Gabor, Fabian Ritz, Sebastian Feld, and Claudia Linnhoff-Popien. "A Distributed Policy Iteration Scheme for Cooperative Multi-Agent Policy Approximation". In *12th Adaptive and Learning Agents Workshop (ALA)*, 2020.

URL: `https://ala2020.vub.ac.be/papers/ALA2020_paper_36.pdf`

Contributions

1. Extension of [54] to partially observable multi-agent domains.
2. A centralized training and decentralized planning scheme to learn effective local policies for large-scale multi-agent scenarios.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Belzner and Schmid discussed the concepts and results, and provided feedback on related work. Gabor, Ritz, and Feld discussed and reviewed the results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 4 regarding the research question **Q3** from Section 1.2.

A.4. VAST: Value Function Factorization with Variable Agent Sub-Teams

Publication

Thomy Phan, Fabian Ritz, Lenz Belzner, Philipp Altmann, Thomas Gabor, and Claudia Linnhoff-Popien. "VAST: Value Function Factorization with Variable Agent Sub-Teams". In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 24018–24032. Curran Associates, Inc., 2021.

URL: `https://proceedings.neurips.cc/paper/2021/hash/c97e7a5153badb6576d8939469f58336-Abstract.html`

Contributions

1. Hierarchical value function factorization with variable agent sub-teams.
2. Meta-gradient optimization of sub-teams.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Ritz and Belzner discussed the concepts and results, and provided support on the sub-team evaluation. Altmann and Gabor discussed and reviewed the results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 4 regarding the research question **Q3** from Section 1.2.

A.5. Learning and Testing Resilience in Cooperative Multi-Agent Systems

Publication

Thomy Phan, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, Bernhard Kempter, Cornel Klein, Horst Sauer, Reiner Schmid, Jan Wieghardt, Marc Zeller, and Claudia Linnhoff-Popien. "Learning and Testing Resilience in Cooperative Multi-Agent Systems". In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1055–1063. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

DOI: `https://dl.acm.org/doi/10.5555/3398761.3398884`
URL: `https://ifaamas.org/Proceedings/aamas2020/pdfs/p1055.pdf`

Contributions

1. Antagonist-based training and test scheme.
2. Resilience evaluation based on agent cross-validation.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Gabor discussed the concepts and results, and provided feedback on related work. Sedlmeier and Ritz discussed and reviewed the results, and supported the setup implementation. Kempter, Klein, Sauer, Schmid, Wieghardt, and Zeller reviewed and discussed the concepts, and provided feedback on the setup. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 5 regarding the research questions **Q4** and **Q5** from Section 1.2.

A.6. Resilient Multi-Agent Reinforcement Learning with Adversarial Value Decomposition

Publication

Thomy Phan, Lenz Belzner, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, and Claudia Linnhoff-Popien. "Resilient Multi-Agent Reinforcement Learning with Adversarial Value Decomposition". In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35(13), pages 11308–11316, 2021.

DOI: `https://doi.org/10.1609/aaai.v35i13.17348`
URL: `https://ojs.aaai.org/index.php/AAAI/article/view/17348`

Contributions

1. Antagonist-based training with variable sized antagonist teams.
2. Resilience evaluation based on cooperation and worst-case performance using dedicated test sets of agents.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Belzner and Gabor discussed the concepts and results, and provided feedback on related work. Sedlmeier supported the setup implementation. Ritz reviewed the results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 5 regarding the research questions **Q4** and **Q5** from Section 1.2.

A.7. Memory Bounded Open-Loop Planning in Large POMDPs using Thompson Sampling

Publication

Thomy Phan, Lenz Belzner, Marie Kiermeier, Markus Friedrich, Kyrill Schmid, and Claudia Linnhoff-Popien. "Memory Bounded Open-Loop Planning in Large POMDPs using Thompson Sampling". In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33(01), pages 7941–7948, 2019.

DOI: `https://doi.org/10.1609/aaai.v33i01.33017941`
URL: `https://ojs.aaai.org/index.php/AAAI/article/view/4794`

Contributions

1. Memory bounded open-loop planning with fixed Thompson Sampling stacks.

2. Memory efficiency evaluation of open- and closed-loop Monte Carlo planning algorithms.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Belzner co-conceived the concepts and provided feedback on implementing Thompson Sampling. Kiermeier, Friedrich, and Schmid discussed and reviewed the results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 6 regarding the research question **Q6** from Section 1.2.

A.8. Adaptive Thompson Sampling Stacks for Memory Bounded Open-Loop Planning

Publication

Thomy Phan, Thomas Gabor, Robert Müller, Christoph Roch, and Claudia Linnhoff-Popien. "Adaptive Thompson Sampling Stacks for Memory Bounded Open-Loop Planning". In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5607–5613. International Joint Conferences on Artificial Intelligence Organization, 2019.

DOI: `https://doi.org/10.24963/ijcai.2019/778`
URL: `https://www.ijcai.org/proceedings/2019/778`

Contributions

Memory bounded open-loop planning with adaptive Thompson Sampling stacks based on bandit convergence.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Gabor, Müller, and Roch discussed the concepts and results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 6 regarding the research question **Q6** from Section 1.2.

A.9. Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability

Publication

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

Contributions

1. Value function factorization scheme, which exploits the recurrence of all agents using self-attention instead of true states.
2. General and configurable Dec-POMDP benchmark, extending StarCraft Multi-Agent Challenge with stochastic observations and higher variance in initial states.

Credit

Phan conceived the original concepts and conducted the empirical analysis. Ritz, Altmann, Zorn, Nüßlein, Kölle, and Gabor discussed the concepts and results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 6 regarding the research question **Q6** from Section 1.2.

Preprint attached below.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability

Thomy Phan [1 †] **Fabian Ritz** [2] **Philipp Altmann** [2] **Maximilian Zorn** [2] **Jonas Nüßlein** [2] **Michael Kölle** [2] **Thomas Gabor** [2] **Claudia Linnhoff-Popien** [2]

Abstract

Stochastic partial observability poses a major challenge for decentralized coordination in multi-agent reinforcement learning but is largely neglected in state-of-the-art research due to a strong focus on state-based *centralized training for decentralized execution (CTDE)* and benchmarks that lack sufficient stochasticity like *StarCraft Multi-Agent Challenge (SMAC)*. In this paper, we propose *Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL)* to approximate value functions under stochastic partial observability. AERIAL replaces the true state with a learned representation of multi-agent recurrence, considering more accurate information about decentralized agent decisions than state-based CTDE. We then introduce *MessySMAC*, a modified version of SMAC with stochastic observations and higher variance in initial states, to provide a more general and configurable benchmark regarding stochastic partial observability. We evaluate AERIAL in Dec-Tiger as well as in a variety of SMAC and MessySMAC maps, and compare the results with state-based CTDE. Furthermore, we evaluate the robustness of AERIAL and state-based CTDE against various stochasticity configurations in MessySMAC.

1. Introduction

A wide range of real-world applications like fleet management, industry 4.0, or communication networks can be formulated as *decentralized partially observable Markov decision process (Dec-POMDP)* representing a cooperative *multi-agent system (MAS)*, where multiple agents have to coordinate to achieve a common goal (Oliehoek & Amato, 2016). *Stochastic partial observability* poses a major challenge for decentralized coordination in Dec-POMDPs due to noisy sensors and potentially high variance in initial states which are common in the real world (Kaelbling et al., 1998; Oliehoek & Amato, 2016).

Multi-agent reinforcement learning (MARL) is a general approach to tackle Dec-POMDPs with remarkable progress in recent years (Wang et al., 2021; Wen et al., 2022). State-of-the-art MARL is based on *centralized training for decentralized execution (CTDE)*, where training takes place in a laboratory or a simulator with access to global information (Lowe et al., 2017; Foerster et al., 2018). For example, *state-based CTDE* exploits true state information to learn a centralized value function in order to derive coordinated policies for decentralized decision making (Rashid et al., 2018; Yu et al., 2022). Due to its effectiveness in the *StarCraft Multi-Agent Challenge (SMAC)* as the current de facto standard for MARL evaluation, state-based CTDE has become very popular and is widely considered an adequate approach to general Dec-POMDPs for more than half a decade, leading to the development of many increasingly complex algorithms (Lyu et al., 2021; 2022).

However, merely relying on state-based CTDE and SMAC in MARL research can be a pitfall in practice as stochastic partial observability is largely neglected – despite being an important aspect in Dec-POMDPs (Lyu et al., 2022):

From an *algorithm perspective*, purely state-based value functions are insufficient to evaluate and adapt multi-agent behavior, since all agents make decisions on a completely different basis, i.e., individual histories of noisy observations and actions. True Dec-POMDP value functions consider more accurate closed-loop information about decentralized agent decisions though (Oliehoek et al., 2008). Furthermore, the optimal state-based value function represents an upper-bound of the true optimal Dec-POMDP value function thus state-based CTDE can result in overly optimistic behavior in general Dec-POMDPs (Lyu et al., 2022).

From a *benchmark perspective*, SMAC has very limited stochastic partial observability due to deterministic observations and low variance in initial states (Ellis et al., 2022).

[1]University of Southern California, USA. [†]Work done at LMU Munich [2]LMU Munich, Germany. This paper is an extension of (Phan et al., 2023). Correspondence to: Thomy Phan <thomy.phan@ifi.lmu.de>.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

Therefore, SMAC scenarios only represent simplified special cases rather than general Dec-POMDP challenges, being insufficient for assessing practicability of MARL.

In this paper, we propose *Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL)* to approximate value functions under agent-wise stochastic partial observability. AERIAL replaces the true state with a learned representation of multi-agent recurrence, considering more accurate closed-loop information about decentralized agent decisions than state-based CTDE. We then introduce *MessyS-MAC*, a modified version of SMAC with stochastic observations and higher variance in initial states, to provide a more general and configurable Dec-POMDP benchmark for more adequate evaluation. Our contributions are as follows:

- We formulate and discuss the concepts of AERIAL w.r.t. stochastic partial observability in Dec-POMDPs.
- We introduce MessySMAC to enable systematic evaluation under various stochasticity configurations.
- We evaluate AERIAL in Dec-Tiger, a small and traditional Dec-POMDP benchmark, as well as in a variety of original SMAC and MessySMAC maps, and compare the results with state-based CTDE. Our results show that AERIAL achieves competitive performance in original SMAC, and superior performance in Dec-Tiger and MessySMAC. Furthermore, we evaluate the robustness of AERIAL and state-based CTDE against various stochasticity configurations in MessySMAC.

2. Background

2.1. Decentralized POMDPs

We formulate cooperative MAS problems as *Dec-POMDP* $M = \langle \mathcal{D}, \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{Z}, \Omega, b_0 \rangle$, where $\mathcal{D} = \{1, ..., N\}$ is a set of agents i, $\mathcal{S}$ is a set of (true) states s_t at time step t, $\mathcal{A} = \langle \mathcal{A}_i \rangle_{i \in \mathcal{D}}$ is the set of joint actions $\mathbf{a_t} = \langle a_{t,1}, ..., a_{t,N} \rangle = \langle a_{t,i} \rangle_{i \in \mathcal{D}}$, $\mathcal{T}(s_{t+1}|s_t, \mathbf{a_t})$ is the state transition probability, $r_t = \mathcal{R}(s_t, \mathbf{a_t}) \in \mathbb{R}$ is the shared reward, $\mathcal{Z}$ is a set of local observations $z_{t,i}$ for each agent $i \in \mathcal{D}$, $\Omega(\mathbf{z_{t+1}}|\mathbf{a_t}, s_{t+1})$ is the probability of joint observation $\mathbf{z_{t+1}} = \langle z_{t+1,i} \rangle_{i \in \mathcal{D}} \in \mathcal{Z}^N$, and b_0 is the probability distribution over initial states s_0 (Oliehoek & Amato, 2016). Each agent i maintains a *local history* $\tau_{t,i} \in (\mathcal{Z} \times \mathcal{A}_i)^t$ and $\boldsymbol{\tau_t} = \langle \tau_{t,i} \rangle_{i \in \mathcal{D}}$ is the *joint history*. A *belief state* $b(s_t|\boldsymbol{\tau_t})$ is a sufficient statistic for joint history $\boldsymbol{\tau_t}$ and defines a probability distribution over true states s_t, updatable by Bayes' theorem (Kaelbling et al., 1998). Joint quantities are written in bold face.

Stochastic partial observability in M is given by *observation* and *initialization stochasticity* w.r.t. Ω and b_0 respectively.

A *joint policy* $\boldsymbol{\pi} = \langle \pi_i \rangle_{i \in \mathcal{D}}$ with *decentralized* or *local policies* π_i defines a deterministic mapping from joint histories to joint actions $\boldsymbol{\pi}(\boldsymbol{\tau_t}) = \langle \pi_i(\tau_{t,i}) \rangle_{i \in \mathcal{D}} \in \mathcal{A}$. The *return* is defined by $G_t = \sum_{c=0}^{T-1} \gamma^c r_{t+c}$, where T is the *horizon* and $\gamma \in [0, 1]$ is the *discount factor*. $\boldsymbol{\pi}$ can be evaluated with a *value function* $Q^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}, \mathbf{a_t}) = \mathbb{E}_{b_0, \mathcal{T}, \Omega}[G_t | \boldsymbol{\tau_t}, \mathbf{a_t}, \boldsymbol{\pi}]$. The goal is to find an *optimal joint policy* $\boldsymbol{\pi}^*$ with *optimal value function* $Q^{\boldsymbol{\pi}^*} = Q^*$ as defined in the next section.

2.2. Optimal Value Functions and Policies

Fully Observable MAS In MDP-like settings with a centralized controller, the optimal value function Q^*_{MDP} is defined by (Watkins & Dayan, 1992; Boutilier, 1996):

$$Q^*_{MDP}(s_t, \mathbf{a_t}) = r_t + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{X} \tag{1}$$

where $\mathcal{X} = \mathcal{T}(s_{t+1}|s_t, \mathbf{a_t}) max_{\mathbf{a_{t+1}} \in A} Q^*_{MDP}(s_{t+1}, \mathbf{a_{t+1}})$.

Due to full observability, Q^*_{MDP} does not depend on $\boldsymbol{\tau_t}$ but on s_t. Thus, decentralized observations $z_{t,i}$ and probabilities according to Ω and b_0 are not considered at all. An optimal (joint) policy $\boldsymbol{\pi}^*_{\mathbf{MDP}}$ of the centralized controller simply maximizes Q^*_{MDP} for all s_t (Watkins & Dayan, 1992):

$$\boldsymbol{\pi}^*_{\mathbf{MDP}} = argmax_{\boldsymbol{\pi}_{\mathbf{MDP}}} \sum_{s_t \in \mathcal{S}} Q^*_{MDP}(s_t, \boldsymbol{\pi}_{\mathbf{MDP}}(s_t)) \tag{2}$$

Partially Observable MAS In general Dec-POMDPs, where true states are not fully observable and only decentralized controllers or agents exist, the optimal value function Q^* is defined by (Oliehoek et al., 2008):

$$Q^*(\boldsymbol{\tau_t}, \mathbf{a_t}) = \sum_{s_t \in \mathcal{S}} b(s_t|\boldsymbol{\tau_t}) \left(r_t + \gamma \sum_{s_{t+1} \in \mathcal{S}} \sum_{\mathbf{z_{t+1}} \in \mathcal{Z}^N} \mathcal{X} \right) \tag{3}$$

where $\mathcal{X} = \mathcal{T}(s_{t+1}|s_t, \mathbf{a_t}) \Omega(\mathbf{z_{t+1}}|\mathbf{a_t}, s_{t+1}) Q^*(\boldsymbol{\tau_{t+1}}, \boldsymbol{\pi}^*(\boldsymbol{\tau_{t+1}}))$ with $\boldsymbol{\tau_{t+1}} = \langle \boldsymbol{\tau_t}, \mathbf{a_t}, \mathbf{z_{t+1}} \rangle$.

An optimal joint policy $\boldsymbol{\pi}^*$ for decentralized execution maximizes the expectation of Q^* for all joint histories $\boldsymbol{\tau_t}$ (Emery-Montemerlo et al., 2004; Oliehoek et al., 2008):

$$\boldsymbol{\pi}^* = argmax_{\boldsymbol{\pi}} \sum_{t=0}^{T-1} \sum_{\boldsymbol{\tau_t} \in (\mathcal{Z}^N \times \mathcal{A})^t} \mathcal{C}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}) \mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0) Q^*(\cdot) \tag{4}$$

where $Q^*(\cdot) = Q^*(\boldsymbol{\tau_t}, \boldsymbol{\pi}(\boldsymbol{\tau_t}))$, indicator $\mathcal{C}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t})$ filters out joint histories $\boldsymbol{\tau_t}$ that are inconsistent with $\boldsymbol{\pi}$, and probability $\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0)$ represents the *recurrence* of all agents considering agent-wise stochastic partial observability w.r.t. decentralization of $\boldsymbol{\pi}$ and $\boldsymbol{\tau_t}$ (Oliehoek et al., 2008):

$$\begin{aligned} \mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0) &= \mathbf{P}(\mathbf{z_0}|b_0) \prod_{c=1}^{t} \mathbf{P}(\mathbf{z_c}|\boldsymbol{\tau_{c-1}}, \boldsymbol{\pi}) \\ &= \mathbf{P}(\mathbf{z_0}|b_0) \prod_{c=1}^{t} \sum_{s_c \in \mathcal{S}} \sum_{s_{c-1} \in \mathcal{S}} \mathcal{T}(\cdot) \Omega(\cdot) \end{aligned} \tag{5}$$

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

where $\mathcal{T}(\cdot) = \mathcal{T}(s_c|s_{c-1}, \boldsymbol{\pi}(\boldsymbol{\tau_{c-1}}))$ and $\Omega(\cdot) = \Omega(\mathbf{z_c}|\boldsymbol{\pi}(\boldsymbol{\tau_{c-1}}), s_c)$.

Since all agents act according to their local history $\tau_{t,i}$ without access to the complete joint history $\boldsymbol{\tau_t}$, recurrence $\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0)$ depends on more accurate *closed-loop information* than just true states s_t, i.e., all previous observations, actions, and probabilities according to b_0, $\mathcal{T}$, and Ω.

Q^*_{MDP} is proven to represent an upper bound of Q^* (Oliehoek et al., 2008). Thus, naively deriving local policies π_i from Q^*_{MDP} instead of Q^* can result in overly optimistic behavior as we will show in Section 4.1 and 6.

2.3. Multi-Agent Reinforcement Learning

Finding an optimal joint policy $\boldsymbol{\pi}^*$ via exhaustive computation of Q^* according to Eq. 3-5 is intractable in practice (Nair et al., 2003; Szer et al., 2005). MARL offers a scalable way to learn Q^* and $\boldsymbol{\pi}^*$ via function approximation, e.g., using CTDE, where training takes place in a laboratory or a simulator with access to global information (Lowe et al., 2017; Foerster et al., 2018). We focus on value-based MARL to learn a centralized value function $Q_{tot} \approx Q^*$, which can be factorized into *local utility functions* $\langle Q_i \rangle_{i \in \mathcal{D}}$ for decentralized decision making via $\pi_i(\tau_{t,i}) = argmax_{a_{t,i}} Q_i(\tau_{t,i}, a_{t,i})$. For that, a *factorization operator* Ψ is used (Phan et al., 2021):

$$Q_{tot}(\boldsymbol{\tau_t}, \mathbf{a_t}) = \Psi(Q_1(\tau_{t,1}, a_{t,1}), ..., Q_N(\tau_{t,N}, a_{t,N})) \quad (6)$$

In practice, Ψ is realized with deep neural networks, such that $\langle Q_i \rangle_{i \in \mathcal{D}}$ can be learned end-to-end via backpropagation by minimizing the mean squared *temporal difference (TD)* error (Rashid et al., 2018; Sunehag et al., 2018). A factorization operator Ψ is *decentralizable* when satisfying the *IGM (Individual-Global-Max)* such that (Son et al., 2019):

$$argmax_{\mathbf{a_t}} Q_{tot}(\boldsymbol{\tau_t}, \mathbf{a_t}) = \begin{pmatrix} argmax_{a_{t,1}} Q_1(\tau_{t,1}, a_{t,1}) \\ \vdots \\ argmax_{a_{t,N}} Q_N(\tau_{t,N}, a_{t,N}) \end{pmatrix} \quad (7)$$

There exists a variety of factorization operators Ψ which satisfy Eq. 7 using monotonicity like QMIX (Rashid et al., 2018), nonlinear transformation like QPLEX (Wang et al., 2021), or loss weighting like CW- and OW-QMIX (Rashid et al., 2020). Most approaches use state-based CTDE to learn Q^*_{MDP} according to Eq. 1 instead of Q^* (Eq. 3-5).

2.4. Recurrent Reinforcement Learning

In partially observable settings, the policy π_i of agent i conditions on the history $\tau_{t,i}$ of past observations and actions (Kaelbling et al., 1998; Oliehoek & Amato, 2016). In practice, *recurrent neural networks (RNNs)* like LSTMs or GRUs are used to learn a compact representation $h_{t,i}$ of $\tau_{t,i}$ and π_i known as *hidden state* or *memory representation*[1], which implicitly encodes the *individual recurrence* of agent i, i.e., the distribution $P_i^{\pi_i}$ over $\tau_{t,i}$ (Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Hu & Foerster, 2019):

$$P_i^{\pi_i}(\tau_{t,i}|b_0) = P_i(z_{0,i}|b_0) \prod_{c=1}^{t} P_i(z_{c,i}|\tau_{c-1,i}, \pi_i) \quad (8)$$

RNNs are commonly used for partially observable problems and have been empirically shown to be more effective than using raw observations $z_{t,i}$ or histories $\tau_{t,i}$ (Hausknecht & Stone, 2015; Samvelyan et al., 2019; Vinyals et al., 2019).

3. Related Work

Multi-Agent Reinforcement Learning In recent years, MARL has achieved remarkable progress in challenging domains (Gupta et al., 2017; Vinyals et al., 2019). State-of-the-art MARL is based on CTDE to learn a centralized value function Q_{tot} for actor-critic learning (Lowe et al., 2017; Foerster et al., 2018; Yu et al., 2022) or factorization (Rashid et al., 2018; 2020; Wang et al., 2021). However, the majority of works assumes a simplified Dec-POMDP setting, where Ω is deterministic, and uses true states to approximate Q^*_{MDP} according to Eq. 1 instead of Q^* (Eq. 3-5). Thus, state-based CTDE is possibly less effective in more general Dec-POMDP settings. Our approach addresses stochastic partial observability with a *learned representation* of multi-agent recurrence $\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0)$ according to Eq. 5 instead of s_t.

Weaknesses of State-Based CTDE Recent works investigated potential weaknesses of state-based CTDE for multi-agent actor-critic methods regarding bias and variance (Lyu et al., 2021; 2022). The experimental results show that state-based CTDE can surprisingly fail in very simple Dec-POMDP benchmarks that exhibit more stochasticity than SMAC. While these studies can be considered an important step towards general Dec-POMDPs, there is neither an approach which adequately addresses stochastic partial observability nor a benchmark to systematically evaluate such an approach yet. In this work, we focus on *value-based MARL*, where learning an accurate value function is important for meaningful factorization, and propose an *attention-based recurrence approach* to approximate value functions under stochastic partial observability. We also introduce a *modified* SMAC benchmark, which enables systematic evaluation under various stochasticity configurations.

Attention-Based CTDE Attention has been used in CTDE to process information of potentially variable length

[1] In this paper, we use the term *memory representation* to avoid confusion with the state terminology of the (Dec-)POMDP literature (Kaelbling et al., 1998; Oliehoek & Amato, 2016).

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

N, where joint observations $\mathbf{z_t}$, joint actions $\mathbf{a_t}$, or local utilities $\langle Q_i \rangle_{i \in \mathcal{D}}$ are weighted and aggregated to provide a meaningful representation for value function approximation (Iqbal & Sha, 2019; Wang et al., 2021; Iqbal et al., 2021; Wen et al., 2022; Khan et al., 2022). Most works focus on Markov games without observation stochasticity, which are special cases of the Dec-POMDP setting. In this work, we focus on *stochastic partial observability* and apply *self-attention* to the *memory representations* $h_{t,i}$ of all agents' RNNs instead of the raw observations $z_{t,i}$ to approximate Q^* for *general Dec-POMDPs* according to Eq. 3-5.

4. AERIAL

4.1. Limitation of State-Based CTDE

Most state-of-the-art works assume a simplified Dec-POMDP setting, where Ω is deterministic, and approximate Q^*_{MDP} according to Eq. 1 instead of Q^* (Eq. 3-5).

If there are only deterministic observations and initial states s_0 such that $b_0(s_0) = 1$ and $b_0(s') = 0$ if $s' \neq s_0$, then multi-agent recurrence $\mathbf{P^\pi}(\boldsymbol{\tau_t}|b_0)$ as defined in Eq. 5 would only depend on state transition probabilities $\mathcal{T}(s_{t+1}|s_t, \mathbf{a_t})$ which are purely state-based, ignoring decentralization of agents and observations (Oliehoek et al., 2008). In such scenarios, stochastic partial observability is very limited, especially if all π_i are deterministic. We hypothesize that this is one reason for the empirical success of state-based CTDE in original SMAC, whose scenarios seemingly have these simplifying properties (Ellis et al., 2022).

In the following, we regard a small example, where state-based CTDE can fail at finding an optimal joint policy $\boldsymbol{\pi}^*$.

Example *Dec-Tiger* is a traditional and simple Dec-POMDP benchmark with $N = 2$ agents facing two doors (Nair et al., 2003). A tiger is randomly placed behind the left (s_L) or right door (s_R) representing the true state. Both agents are able to listen (*li*) and open the left (o_L) or right door (o_R). The listening action *li* produces a noisy observation of either hearing the tiger to be left (z_L) or right (z_R), which correctly indicates the tiger's position with 85% chance and a cost of -1 per listening agent. If both agents open the same door, the episode terminates with a reward of -50 if opening the tiger door and +20 otherwise. If both agents open different doors, the episode ends with -100 reward and, if only one agent opens a door while the other agent is listening, the episode terminates with -101 if opening the tiger door and +9 otherwise.

Given a horizon of $T = 2$, the tiger being behind the right door (s_R), and both agents having listened in the first step, where agent 1 heard z_L and agent 2 heard z_R: Assuming that both agents learned to perform the same actions, e.g., due to CTDE and parameter sharing (Tan, 1993; Gupta et al., 2017), Q^*_{MDP} and Q^* would estimate the following values[2]:

$$Q^*_{MDP}(s_R, \langle li, li \rangle) = -2 \qquad Q^*(\boldsymbol{\tau_t}, \langle li, li \rangle) = -2$$
$$Q^*_{MDP}(s_R, \langle o_L, o_L \rangle) = 20 \qquad Q^*(\boldsymbol{\tau_t}, \langle o_L, o_L \rangle) = -15$$
$$Q^*_{MDP}(s_R, \langle o_R, o_R \rangle) = -50 \qquad Q^*(\boldsymbol{\tau_t}, \langle o_R, o_R \rangle) = -15$$

Any policy $\boldsymbol{\pi}^*_{\mathbf{MDP}}$ or decentralizable joint policy $\boldsymbol{\pi}$ w.r.t. IGM (Eq. 7) that maximizes Q^*_{MDP} according to Eq. 2 would optimistically recommend $\langle o_L, o_L \rangle$ based on the true state s_R, regardless of what the agents observed. However, any joint policy $\boldsymbol{\pi}^*$ that maximizes the expectation of Q^* according to Eq. 4 would consider agent-wise stochastic partial observability and recommend $\langle li, li \rangle$, which corresponds to the true optimal decision for $T = 2$ (Szer et al., 2005).

4.2. Attention-Based Embeddings of Recurrence

Preliminaries We now introduce *Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL)* to approximate optimal Dec-POMDP value functions Q^* according to Eq. 3-5. Our setup uses a factorization operator Ψ like QMIX or QPLEX according to Eq. 6-7. All agents process their local histories $\tau_{t,i}$ via RNNs as motivated in Section 2.4 and schematically shown in Fig. 1 (left).

Unlike Q^*_{MDP}, the true optimal Dec-POMDP value function Q^* considers more accurate closed-loop information about decentralized agent decisions through multi-agent recurrence $\mathbf{P^\pi}(\boldsymbol{\tau_t}|b_0)$ according to Eq. 5. Simply replacing s_t with $\boldsymbol{\tau_t}$ as suggested in (Lyu et al., 2022) is not sufficient because the resulting value function would assume a centralized controller with access to the complete joint history $\boldsymbol{\tau_t}$, in contrast to decentralized agents i which can only access their respective local history $\tau_{t,i}$ (Oliehoek et al., 2008).

Exploiting Multi-Agent Recurrence At first we propose to naively exploit all individual recurrences by simply replacing the true state s_t in CTDE with the *joint memory representation* $\mathbf{h_t} = \langle h_{t,i} \rangle_{i \in \mathcal{D}}$ of all agents' RNNs. Each memory representation $h_{t,i}$ implicitly encodes the individual recurrence $P^{\pi_i}_i(\tau_{t,i}|b_0)$ of agent i according to Eq. 8. Therefore, $\mathbf{h_t}$ provides more accurate closed-loop information about decentralized agent decisions than s_t.

This approach, called `AERIAL (no attention)`, can already be considered a sufficient solution if all individual recurrences $P^{\pi_i}_i(\tau_{t,i}|b_0)$ are statistically independent such that $\mathbf{P^\pi}(\boldsymbol{\tau_t}|b_0) = \prod_{i=1}^{N} P^{\pi_i}_i(\tau_{t,i}|b_0)$.

Attention-Based Recurrence While `AERIAL (no attention)` offers a simple way to address agent-wise stochastic partial observability, the independence assumption of all individual recurrences $P^{\pi_i}_i(\tau_{t,i}|b_0)$ does not hold

[2] The exact calculation is provided in the Appendix B.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

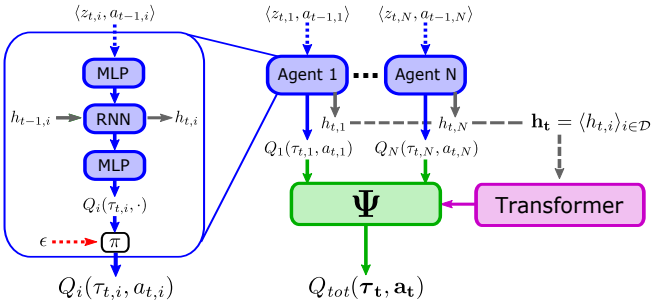


Figure 1. Illustration of the AERIAL setup. *Left:* Recurrent agent network structure with memory representations $h_{t-1,i}$ and $h_{t,i}$. *Right:* Value function factorization via factorization operator Ψ using the joint memory representation $\mathbf{h_t} = \langle h_{t,i} \rangle_{i \in \mathcal{D}}$ of all agents' RNNs instead of true states s_t. All memory representations $h_{t,i}$ are detached from the computation graph to avoid additional differentiation (indicated by the dashed gray arrows) and passed through a simplified transformer before being used by Ψ for value function factorization.

in practice due to correlations in observations and actions (Bernstein et al., 2005; Amato et al., 2007).

Given the Dec-Tiger example above, the individual recurrences according to Eq. 8 are $P_1^{\pi_1}(\tau_{t,1}|b_0) = P_2^{\pi_2}(\tau_{t,2}|b_0) = 0.5$ (Kaelbling et al., 1998). However, the actual multi-agent recurrence according to Eq. 5 is $\mathbf{P^\pi}(\boldsymbol{\tau}_\mathbf{t}|b_0) = 0.15 \cdot 0.85 \neq P_1^{\pi_1}(\tau_{t,1}|b_0) \cdot P_2^{\pi_2}(\tau_{t,2}|b_0)$, indicating that individual recurrences are not statistically independent in general (Oliehoek & Amato, 2016).

Therefore, we process $\mathbf{h_t}$ by a simplified *transformer* along the agent axis to automatically consider the latent dependencies of all memory representations $h_{t,i} \in \mathbf{h_t}$ through self-attention. The resulting approach, called `AERIAL`, is depicted in Fig. 1 and Algorithm 1 in Appendix C.

Our transformer does not use positional encoding or masking, since we assume no particular ordering among agents. The joint memory representation $\mathbf{h_t}$ is passed through a single *multi-head attention* layer with the output of each attention head c being defined by (Vaswani et al., 2017):

$$att_c(\mathbf{h_t}) = softmax\left(\frac{W_q^c(\mathbf{h_t})W_k^c(\mathbf{h_t})^\top}{\sqrt{d_{att}}}\right) W_v^c(\mathbf{h_t}) \quad (9)$$

where W_q^c, W_k^c, and W_v^c are *multi-layer perceptrons (MLP)* with an output dimensionality of d_{att}. All outputs $att_c(\mathbf{h_t})$ are summed and passed through a series of MLP layers before being fed into the factorization operator Ψ, effectively replacing the true state s_t by a learned representation of multi-agent recurrence $\mathbf{P^\pi}(\boldsymbol{\tau}_\mathbf{t}|b_0)$ according to Eq. 5.

To avoid additional differentation of $\mathbf{h_t}$ through Ψ or Eq. 9, we detach $\mathbf{h_t}$ from the computation graph. Thus, we make sure that $\mathbf{h_t}$ is only learned through agent RNNs.

4.3. Discussion of AERIAL

The strong focus on state-based CTDE in the last few years has led to the development of increasingly complex algorithms that largely neglect stochastic partial observability in general Dec-POMDPs (Lyu et al., 2021; 2022). In contrast, AERIAL offers a simple way to adjust factorization approaches by replacing the true state s_t with a learned representation of multi-agent recurrence $\mathbf{P^\pi}(\boldsymbol{\tau}_\mathbf{t}|b_0)$ to consider more accurate closed-loop information about decentralized agent decisions. The rest of the training scheme remains unchanged, which eases adjustment of existing approaches.

Since the naive independence assumption of individual memory representations $h_{t,i}$ does not hold in practice – despite decentralization – we use a simplified transformer to consider the latent dependencies of all $h_{t,i} \in \mathbf{h_t}$ along the agent axis to learn an adequate representation of multi-agent recurrence $\mathbf{P^\pi}(\boldsymbol{\tau}_\mathbf{t}|b_0)$ according to Eq. 5.

AERIAL does not depend on true states therefore requiring less overall information than state-based CTDE, since we assume $\mathbf{h_t}$ to be available in all CTDE setups anyway (Foerster et al., 2018; Rashid et al., 2020). Note that AERIAL does not necessarily require RNNs to obtain $\mathbf{h_t}$ as hidden layers of MLPs or decision transformers can be used to approximate $\mathbf{h_t}$ as well (Son et al., 2019; Chen et al., 2021).

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

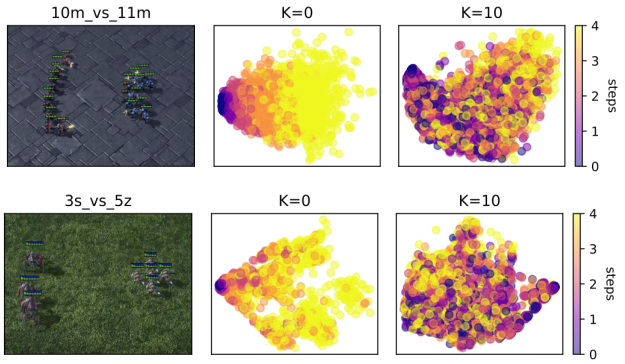


Figure 2. Left: Screenshot of two SMAC maps. *Middle:* PCA visualization of the joint observations in original SMAC within the first 5 steps of 1,000 episodes using a random policy with $K = 0$ initial random steps. *Right:* Analogous PCA visualization for MessySMAC with $K = 10$ initial random steps. For visual comparability, the observations are deterministic here.

5. MessySMAC

5.1. Limitation of SMAC as a Benchmark

StarCraft Multi-Agent Challenge (SMAC) provides a rich set of micromanagement tasks, where a team of learning agents has to fight against an enemy team, which acts according to handcrafted heuristics of the built-in StarCraft AI (Samvelyan et al., 2019). SMAC currently represents the de facto standard for MARL evaluation (Rashid et al., 2018; 2020; Wang et al., 2021). However, SMAC scenarios exhibit very limited stochastic partial observability due to deterministic observations and low variance in initial states therefore only representing simplified special cases rather than general Dec-POMDP challenges (Lyu et al., 2022; Ellis et al., 2022). To assess practicability of MARL, we need benchmarks with sufficient stochasticity as the real-world is generally messy and only observable through noisy sensors.

5.2. SMAC with Stochastic Partial Observability

MessySMAC is a modified version of SMAC with *observation stochasticity* w.r.t. Ω, where all measured values of observation $z_{t,i}$ are negated with a probability of $\phi \in [0, 1)$, and *initialization stochasticity* w.r.t. b_0, where K random steps are initially performed before officially starting an episode. During the initial phase, the agents can already be ambushed by the built-in AI, which further increases difficulty compared to the original SMAC maps if $K > 0$. MessySMAC represents a more general Dec-POMDP challenge which enables systematic evaluation under various stochasticity configurations according to ϕ and K.

Fig. 2 shows the PCA visualization of joint observations in two maps of original SMAC ($K = 0$) and MessySMAC ($K = 10$) within the first 5 steps of 1,000 episodes using a random policy. In original SMAC, the initial observations of s_0 (dark purple) are very similar and can be easily distinguished from subsequent observations by merely regarding time steps. Therefore, open-loop control might already be sufficient to solve these scenarios satisfactorily as hypothesized in (Ellis et al., 2022). However, the distinction of observations by time steps is more tricky in MessySMAC due to significantly higher entropy in b_0, indicating higher initialization stochasticity and a stronger requirement for closed-loop control, where agents need to explicitly consider their actual observations to make proper decisions.

5.3. Comparison with SMACv2

SMACv2 is an update to the original SMAC benchmark featuring initialization stochasticity w.r.t. position and unit types, as well as observation restrictions (Ellis et al., 2022). SMACv2 addresses similar issues as MessySMAC but MessySMAC additionally features *observation stochasticity* w.r.t. Ω according to the general Dec-POMDP formulation in Section 2.1. Unlike MessySMAC, SMACv2 does not support the *original SMAC maps* thus not enabling direct comparability w.r.t. stochasticity configurations.

Therefore, SMACv2 can be viewed as entirely new StarCraft II benchmark, while MessySMAC represents a *SMAC extension*, enabling systematic evaluation under various stochasticity configurations for the original SMAC maps.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

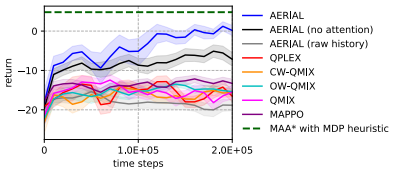


Figure 3. Average learning progress w.r.t. the return of AERIAL variants and state-of-the-art baselines in Dec-Tiger over 50 runs. Shaded areas show the 95% confidence interval.

6. Experiments

We use the state-based CTDE implementations of QPLEX, CW-QMIX, OW-QMIX, and QMIX from (Rashid et al., 2020) as state-of-the-art baselines with their default hyperparameters. We also integrate MAPPO from (Yu et al., 2022). For all experiments, we report the average performance and the 95% confidence interval over at least 20 runs.

AERIAL is implemented[3] using QMIX as factorization operator Ψ according to Fig. 1. We also experimented with QPLEX as alternative with no significant difference in performance. Thus, we stick with QMIX for efficiency due to fewer trainable parameters. The transformer of AERIAL has 4 heads with W_q^c, W_k^c, and W_v^c each having one hidden layer of $d_{att} = 64$ units with ReLU activation. The subsequent MLP layers have 64 units with ReLU activation.

For ablation study, we implement AERIAL (no attention), which trains Ψ directly on $\mathbf{h_t}$ without self-attention as described in Section 4.2, and AERIAL (raw history), which trains Ψ on the raw joint history $\boldsymbol{\tau_t}$ concatenated with the true state s_t as originally proposed for actor-critic methods (Lyu et al., 2022).

6.1. Dec-Tiger

Setting We use the Dec-Tiger problem described in Section 4.1 and (Nair et al., 2003) as simple proof-of-concept domain with $T = 4$ and $\gamma = 1$. We also provide the optimal value of 4.8 computed with MAA* (Szer et al., 2005).

Results The results are shown in Fig. 3. AERIAL comes closest to the optimum, achieving an average return of about zero. AERIAL (no attention) performs second best with an average return of about -8, while all other approaches achieve an average return of about -15.

Discussion The results confirm the example from Section 4.1 and the findings of (Oliehoek et al., 2008; Lyu et al., 2022). All state-based CTDE approaches and AERIAL (raw history) converge to a one-step policy, where both agents optimistically open the same door regardless of any agent observation. AERIAL (no attention) converges to a local optimum most of the time, where both agents only listen for all $T = 4$ time steps. AERIAL performs best due to considering the latent dependencies of all memory representations $h_{t,i} \in \mathbf{h_t}$ via self-attention to learn an adequate representation of multi-agent recurrence $\mathbf{P^\pi}(\boldsymbol{\tau_t}|b_0)$ according to Eq. 5.

6.2. Original SMAC

Setting We evaluate AERIAL in original SMAC using the maps 3s5z and 10m_vs_11m, which are classified as *easy*, as well as the *hard* maps 2c_vs_64zg, 3s_vs_5z, and 5m_vs_6m, and the *super hard* map 3s5z_vs_3s6z (Samvelyan et al., 2019).

Results The final average test win rates after 2 million steps of training are shown in Table 1. AERIAL is competitive to QPLEX and QMIX in the easy maps, while performing best in 3s_vs_5z and 5m_vs_6m. MAPPO performs best in 2c_vs_64zg and 3s5z_vs_3s6z with AERIAL being second best in the super hard map 3s5z_vs_3s6z.

Discussion AERIAL is competitive to state-of-the-art baselines in original SMAC, indicating that replacing the true state s_t with the joint memory representation $\mathbf{h_t}$ does not notably harm performance. Despite outperforming most baselines in some maps, we do not claim significant outperformance here, since we regard most SMAC maps as widely solved by the community anyway (Ellis et al., 2022).

6.3. MessySMAC

Setting We evaluate AERIAL in MessySMAC using the same maps as in Section 6.2. We set $\phi = 15\%$ and $K = 10$.

Results The results are shown in Fig. 4. AERIAL performs best in all maps with AERIAL (no attention) being second best except in 2c_vs_64zg. In 3s5z_vs_3s6z, only AERIAL and AERIAL (no attention) progress notably. AERIAL (raw history) performs worst in all maps. MAPPO only progresses notably in 2c_vs_64zg.

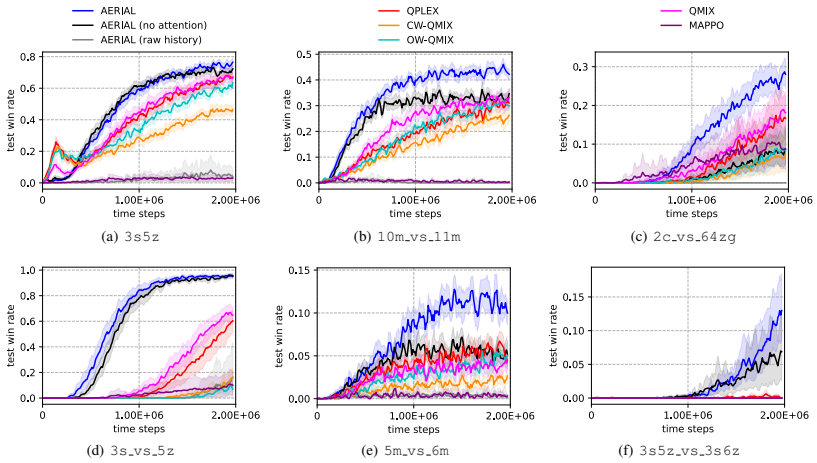
Discussion Similar to the Dec-Tiger experiment, the results confirm the benefit of exploiting more accurate closed-loop information in domains with stochastic partial observability. AERIAL consistently outperforms AERIAL (no attention), indicating that self-attention can correct for the naive independence assumption of all $h_{t,i} \in \mathbf{h_t}$. MAPPO performs especially poorly in MessySMAC due to

[3]Code is available at https://github.com/thomyphan/messy_smac. Further details are in Appendix D.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

Table 1. Average win rate of AERIAL and state-of-the-art baselines after 2 million time steps of training across 400 final test episodes for the original SMAC maps with the 95% confidence interval. The best results per map are highlighted in boldface and blue.

	AERIAL	QPLEX	CW-QMIX	OW-QMIX	QMIX	MAPPO
3s5z	**0.95 ± 0.01**	0.94 ± 0.01	0.87 ± 0.02	0.91 ± 0.02	**0.95 ± 0.01**	68.7 ± 0.94
10m_vs_11m	**0.97 ± 0.01**	0.90 ± 0.02	0.91 ± 0.02	0.96 ± 0.01	0.90 ± 0.02	77.3 ± 0.66
2c_vs_64zg	0.52 ± 0.11	0.29 ± 0.1	0.38 ± 0.12	0.55 ± 0.13	0.59 ± 0.11	**90.2 ± 0.24**
3s_vs_5z	**0.96 ± 0.02**	0.74 ± 0.11	0.18 ± 0.06	0.08 ± 0.04	0.81 ± 0.05	73.8 ± 0.44
5m_vs_6m	**0.77 ± 0.03**	0.66 ± 0.04	0.41 ± 0.04	0.55 ± 0.06	0.67 ± 0.05	60.6 ± 1.13
3s5z_vs_3s6z	0.18 ± 0.09	0.1 ± 0.03	0.0 ± 0.0	0.02 ± 0.01	0.02 ± 0.02	**20.5 ± 2.91**



(a) 3s5z (b) 10m_vs_11m (c) 2c_vs_64zg

(d) 3s_vs_5z (e) 5m_vs_6m (f) 3s5z_vs_3s6z

Figure 4. Average learning progress w.r.t. the win rate of AERIAL variants and state-of-the-art baselines in MessySMAC for 2 million steps over 20 runs. Shaded areas show the 95% confidence interval. The legend at the top applies across all plots.

its misleading dependence on true states without any credit assignment, confirming the findings of (Ellis et al., 2022).

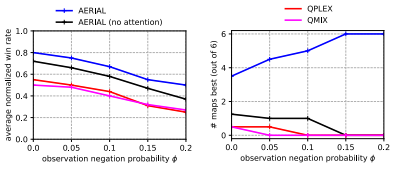
6.4. Robustness against Stochastic Partial Observability

Setting To evaluate the robustness of AERIAL and AERIAL (no attention) against various stochasticity configurations in MessySMAC, we manipulate Ω through the observation negation probability ϕ and b_0 through the number of initial random steps K as defined in Section 5.2. We compare the results with QMIX and QPLEX as the best performing state-of-the-art baselines in MessySMAC according to the results in Section 6.3. We present summarized plots, where the results are aggregated accross all maps from Section 6.3. To avoid that easy maps dominate the average win rate, since all approaches achieve high values there, we normalize the values by the maximum win rate achieved in the respective map for all tested configurations of ϕ and K. Thus, we ensure an equal weighting regardless of the particular difficulty level. If not mentioned otherwise, we set $\phi = 15\%$ and $K = 10$ as default parameters based on Section 6.3.

Results The results regarding observation stochasticity w.r.t. Ω and ϕ are shown in Fig. 5. Fig. 5(a) shows that the average win rates of all approaches decrease with increasing ϕ with AERIAL consistently achieving the highest average win rate in all configurations. Fig. 5(b) shows that AERIAL performs best in most MessySMAC maps, especially when $\phi \geq 15\%$. AERIAL (no attention) performs second best.

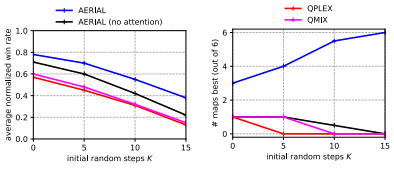
The results regarding initialization stochasticity w.r.t. b_0 and K are shown in Fig. 6. Analogously to Fig. 5, Fig. 6(a) shows that the average (normalized) win rates of all ap-

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.



(a) normalized test win rate (b) # maps best out of 6

Figure 5. Evaluation of `AERIAL`, `AERIAL (no attention)`, and the best MessySMAC baselines for different observation negation probabilities ϕ affecting observation stochasticity w.r.t. Ω (20 runs per configuration). (a) The average normalized test win rate accross all 6 MessySMAC maps from Section 6.3. (b) The number of maps best out of 6. The legend at the top applies across all plots.



(a) normalized test win rate (b) # maps best out of 6

Figure 6. Evaluation of `AERIAL`, `AERIAL (no attention)`, and the best MessySMAC baselines for different initial random steps K affecting initialization stochasticity w.r.t. b_0 (20 runs per configuration). (a) The average normalized test win rate accross all 6 MessySMAC maps from Section 6.3. (b) The number of maps best out of 6. The legend at the top applies across all plots.

proaches decrease with increasing K with `AERIAL` consistently achieving the highest average win rate in all configurations. Fig. 6(b) shows that `AERIAL` performs best in most MessySMAC maps, especially when $K \geq 10$. `AERIAL (no attention)` performs second best.

Discussion Our results systematically demonstrate the robustness of `AERIAL` and `AERIAL (no attention)` against various stochasticity configurations according to Ω and b_0. State-based CTDE is notably less effective in settings, where observation and initialization stochasticity is high. As `AERIAL` consistently performs best in all maps when $\phi \geq 15\%$ or $K \geq 10$, we conclude that providing an adequate representation of $\mathbf{P}^{\boldsymbol{\pi}}(\boldsymbol{\tau_t}|b_0)$ according to Eq. 5 that is learned, e.g., through $\mathbf{h_t}$ and self-attention, is more beneficial for CTDE than merely relying on true states when facing domains with high stochastic partial observability.

7. Conclusion and Future Work

To tackle general multi-agent problems, which are messy and only observable through noisy sensors, we need adequate algorithms and benchmarks that sufficiently consider stochastic partial observability.

In this paper, we proposed AERIAL to approximate value functions under stochastic partial observability with a learned representation of multi-agent recurrence, considering more accurate closed-loop information about decentralized agent decisions than state-based CTDE.

We then introduced *MessySMAC*, a modified version of SMAC with stochastic observations and higher variance in initial states, to provide a more general and configurable Dec-POMDP benchmark regarding stochastic partial observability. We showed visually in Fig. 2 and experimentally in Section 6 that MessySMAC scenarios pose a greater challenge than their original SMAC counterparts due to observation and initialization stochasticity.

Compared to state-based CTDE, AERIAL offers a simple but effective approach to general Dec-POMDPs, being competitive in original SMAC and superior in Dec-Tiger and MessySMAC, which both exhibit observation and initialization stochasticity unlike original SMAC. Simply replacing the true state with memory representations can already improve performance in most scenarios, confirming the need for more accurate closed-loop information about decentralized agent decisions. Self-attention can correct for the naive independence assumption of agent-wise recurrence to further improve performance, especially when observation or initialization stochasticity is high.

We plan to further evaluate AERIAL in SMACv2 and mixed competitive-cooperative settings with multiple CTDE instances (Lowe et al., 2017; Phan et al., 2020).

Acknowledgements

This work was partially funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems.

References

Amato, C., Bernstein, D. S., and Zilberstein, S. Optimizing Memory-Bounded Controllers for Decentralized POMDPs. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pp. 1–8, 2007.

Bernstein, D. S., Hansen, E. A., and Zilberstein, S. Bounded Policy Iteration for Decentralized POMDPs. In *IJCAI*, pp. 52–57, 2005.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

Boutilier, C. Planning, Learning and Coordination in Multiagent Decision Processes. In *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*, pp. 195–210. Morgan Kaufmann Publishers Inc., 1996.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision Transformer: Reinforcement Learning via Sequence Modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, 2014.

Ellis, B., Moalla, S., Samvelyan, M., Sun, M., Mahajan, A., Foerster, J. N., and Whiteson, S. SMACv2: An Improved Benchmark for Cooperative Multi-Agent Reinforcement Learning. 2022.

Emery-Montemerlo, R., Gordon, G., Schneider, J., and Thrun, S. Approximate Solutions for Partially Observable Stochastic Games with Common Payoffs. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '04, pp. 136–143, USA, 2004. IEEE Computer Society. ISBN 1581138644.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual Multi-Agent Policy Gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

Gupta, J. K., Egorov, M., and Kochenderfer, M. Cooperative Multi-Agent Control using Deep Reinforcement Learning. In *Autonomous Agents and Multiagent Systems*, pp. 66–83. Springer, 2017.

Hausknecht, M. and Stone, P. Deep Recurrent Q-Learning for Partially Observable MDPs. In *2015 AAAI Fall Symposium Series*, 2015.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

Hu, H. and Foerster, J. N. Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning. In *International Conference on Learning Representations*, 2019.

Iqbal, S. and Sha, F. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2961–2970, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Iqbal, S., De Witt, C. A. S., Peng, B., Boehmer, W., Whiteson, S., and Sha, F. Randomized Entity-wise Factorization for Multi-Agent Reinforcement Learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4596–4606. PMLR, 18–24 Jul 2021.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and Acting in Partially Observable Stochastic Domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Khan, M. J., Ahmed, S. H., and Sukthankar, G. Transformer-Based Value Function Decomposition for Cooperative Multi-Agent Reinforcement Learning in StarCraft. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 18(1):113–119, Oct. 2022.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Lyu, X., Xiao, Y., Daley, B., and Amato, C. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, pp. 844–852, 2021.

Lyu, X., Baisero, A., Xiao, Y., and Amato, C. A Deeper Understanding of State-Based Critics in Multi-Agent Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9396–9404, Jun. 2022. doi: 10.1609/aaai.v36i9.21171.

Nair, R., Tambe, M., Yokoo, M., Pynadath, D., and Marsella, S. Taming Decentralized POMDPs: Towards Efficient Policy Computation for Multiagent Settings. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pp. 705–711, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

Oliehoek, F. A. and Amato, C. *A Concise Introduction to Decentralized POMDPs*, volume 1. Springer, 2016.

Oliehoek, F. A., Spaan, M. T., and Vlassis, N. Optimal and Approximate Q-Value Functions for Decentralized POMDPs. *Journal of Artificial Intelligence Research*, 32: 289–353, 2008.

Phan, T., Gabor, T., Sedlmeier, A., Ritz, F., Kempter, B., Klein, C., Sauer, H., Schmid, R., Wieghardt, J., Zeller, M., et al. Learning and Testing Resilience in Cooperative

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

Multi-Agent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '20, pp. 1055–1063. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

Phan, T., Ritz, F., Belzner, L., Altmann, P., Gabor, T., and Linnhoff-Popien, C. VAST: Value Function Factorization with Variable Agent Sub-Teams. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24018–24032. Curran Associates, Inc., 2021.

Phan, T., Ritz, F., Nüßlein, J., Kölle, M., Gabor, T., and Linnhoff-Popien, C. Attention-Based Recurrency for Multi-Agent Reinforcement Learning under State Uncertainty. In *Extended Abstracts of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pp. 2839–2841. International Foundation for Autonomous Agents and Multiagent Systems, 2023. ISBN 9781450394321.

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., and Whiteson, S. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304. PMLR, 10–15 Jul 2018.

Rashid, T., Farquhar, G., Peng, B., and Whiteson, S. Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10199–10210. Curran Associates, Inc., 2020.

Samvelyan, M., Rashid, T., Schroeder de Witt, C., Farquhar, G., Nardelli, N., Rudner, T. G., Hung, C.-M., Torr, P. H., Foerster, J., and Whiteson, S. The StarCraft Multi-Agent Challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '19, pp. 2186–2188, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., and Yi, Y. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896. PMLR, 09–15 Jun 2019.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., and Graepel, T. Value-Decomposition Networks for Cooperative Multi-Agent Learning based on Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '18, pp. 2085–2087, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

Szer, D., Charpillet, F., and Zilberstein, S. MAA*: A Heuristic Search Algorithm for Solving Decentralized POMDPs. UAI'05, pp. 576–583, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

Tan, M. Multi-Agent Reinforcement Learning: Independent versus Cooperative Agents. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, ICML'93, pp. 330–337, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc. ISBN 1558603077.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is All You Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster Level in StarCraft II using Multi-Agent Reinforcement Learning. *Nature*, pp. 1–5, 2019.

Wang, J., Ren, Z., Liu, T., Yu, Y., and Zhang, C. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *International Conference on Learning Representations*, 2021.

Watkins, C. J. and Dayan, P. Q-Learning. *Machine Learning*, 8(3-4):279–292, 1992.

Wen, M., Kuba, J. G., Lin, R., Zhang, W., Wen, Y., Wang, J., and Yang, Y. Multi-Agent Reinforcement Learning is a Sequence Modeling Problem. *arXiv preprint arXiv:2205.14953*, 2022.

Whittlestone, J., Arulkumaran, K., and Crosby, M. The Societal Implications of Deep Reinforcement Learning. *Journal of Artificial Intelligence Research*, 70:1003–1030, May 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12360.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

A. Limitations and Societal Impacts

A.1. Limitations

AERIAL does not significantly outperform state-of-the-art baselines in easier domains without stochastic partial observability as indicated by the original SMAC results in Table 1, implying that simplified Dec-POMDP settings might benefit from more specialized algorithms. The dependence on joint memory representations $\mathbf{h_t} = \langle h_{t,i} \rangle_{i \in \mathcal{D}}$ might induce some bias w.r.t. agent behavior policies which could limit performance in hard exploration domains therefore requiring additional mechanisms beyond the scope of this work. The full version of AERIAL requires additional compute[4] due to the transformer component in Fig. 1 which can be compensated by using a more (parameter) efficient value function factorization operator Ψ, e.g., using QMIX instead of QPLEX.

A.2. Potential Negative Societal Impacts

The goal of our work is to realize autonomous systems to solve complex tasks under stochastic partial observability as motivated in Section 1. We refer to (Whittlestone et al., 2021) for a general overview regarding societal implications of deep RL and completely focus on cooperative MARL settings in the following.

AERIAL is based on a centralized training regime to learn decentralized policies with a common objective. That objective might include bias of a central authority and could potentially harm opposing parties, e.g., via discrimination or misleading information. Since training is conducted in a laboratory or a simulation, the resulting system might exhibit unsafe or questionable behavior when being deployed in the real world due to poor generalization, e.g., leading to accidents or unfair decisions. The transformer component in Fig. 1 might require a significant amount of additional compute for tuning and training therefore increasing overall cost. The self-attention weights of Eq. 9 could be used to discriminate participating individuals in an unethical way, e.g., discarding less relevant groups of individuals according to the softmax output.

Similar to original SMAC, MessySMAC is based on team battles, indicating that any MARL algorithm mastering that challenge could be misused for real combat, e.g., in autonomous weapon systems to realize distributed and coordinated strategies. Since MessySMAC covers the aspect of stochastic partial observability, successfully evaluated algorithms could be potentially more effective and dangerous in real-world scenarios.

B. Dec-Tiger Example

Given the Dec-Tiger example from Section 4.1 with a horizon of $T = 2$, the tiger being behind the right door (s_R), and both agents having listened in the first step, where agent 1 heard z_L and agent 2 heard z_R: The final state-based values are defined by $Q^*_{MDP}(s_t, \mathbf{a_t}) = \mathcal{R}(s_t, \mathbf{a_t})$.

Due to both agents perceiving different observations, i.e., z_L and z_R respectively, the probability of being in state s_R is 50% according to the belief state, i.e., $b(s_R|\boldsymbol{\tau_t}) = b(s_L|\boldsymbol{\tau_t}) = \frac{1}{2}$. Thus, the true optimal Dec-POMDP values for the final time step are defined by:

$$\begin{aligned} Q^*(\boldsymbol{\tau_t}, \mathbf{a_t}) &= \sum_{s_t \in \mathcal{S}} b(s_t|\boldsymbol{\tau_t}) \mathcal{R}(s_t, \mathbf{a_t}) \\ &= \frac{1}{2}(Q^*_{MDP}(s_L, \mathbf{a_t}) + Q^*_{MDP}(s_R, \mathbf{a_t})) \end{aligned} \tag{10}$$

The values of Q^*_{MDP} and Q^* for the final time step $t = 2$ in the example are given in Table 2. Both agents can reduce the expected penalty when always performing the same action. Therefore, it is likely for MARL to converge to a joint policy that recommends the same actions for both agents, especially when synchronization techniques like parameter sharing are used (Tan, 1993; Gupta et al., 2017; Yu et al., 2022).

C. Full Algorithm of AERIAL

The complete formulation of AERIAL is given in Algorithm 1. Note that AERIAL does not depend on true states s_t at all, since the experience samples e_t (Line 23) used for training do not record any states.

[4]The additional amount regarding wall clock time was negligible in our experiments though.

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.

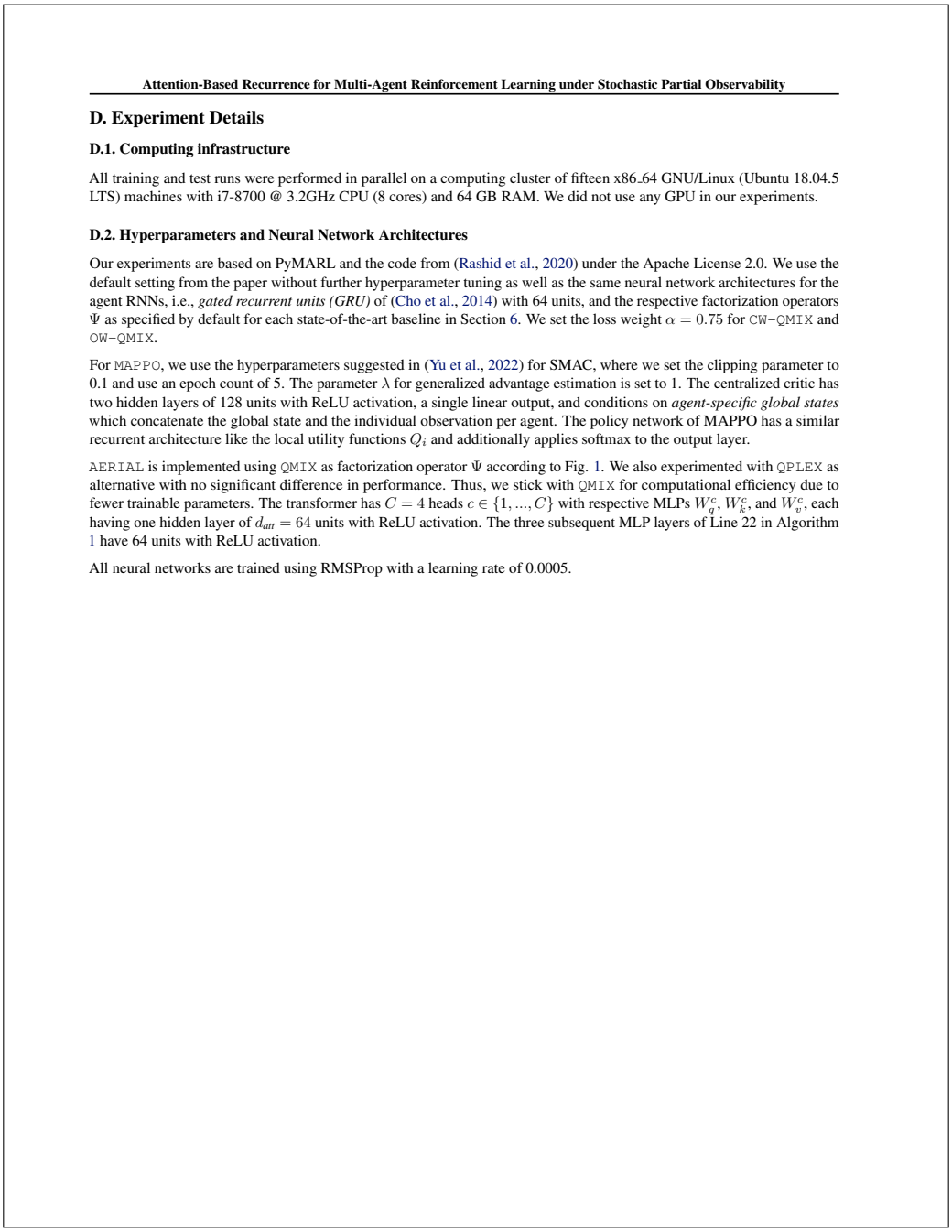
Table 2. The values of Q^*_{MDP} and Q^* for the final time step $t = 2$ in the Dec-Tiger example from Section 4.1.

$\mathbf{a_t}$	$Q^*_{MDP}(s_L, \mathbf{a_t})$	$Q^*_{MDP}(s_R, \mathbf{a_t})$	$Q^*(\boldsymbol{\tau_t}, \mathbf{a_t})$
$\langle li, li \rangle$	$\mathbf{-2}$	$\mathbf{-2}$	$\mathbf{-2}$
$\langle li, o_L \rangle$	-101	+9	-46
$\langle li, o_R \rangle$	+9	-101	-46
$\langle o_L, li \rangle$	-101	+9	-46
$\langle \mathbf{o_L}, \mathbf{o_L} \rangle$	$\mathbf{-50}$	$\mathbf{+20}$	$\mathbf{-15}$
$\langle o_L, o_R \rangle$	-100	-100	-100
$\langle o_R, li \rangle$	+9	-101	-46
$\langle o_R, o_L \rangle$	-100	-100	-100
$\langle \mathbf{o_R}, \mathbf{o_R} \rangle$	$\mathbf{+20}$	$\mathbf{-50}$	$\mathbf{-15}$

Algorithm 1 Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL)

Initialize parameters for $\langle Q_i \rangle_{i \in \mathcal{D}}$ and Ψ.
for episode $m \leftarrow 1, E$ **do**
  Sample s_0, $\mathbf{z_0}$, and $\boldsymbol{\tau_0}$ via b_0 and Ω
  for time step $t \leftarrow 0, T-1$ **do**
    for agent $i \in \mathcal{D}$ **do**
      $a_{t,i} \leftarrow \pi_i(\tau_{t,i})$ {Use $argmax_{a_{t,i} \in \mathcal{A}_i} Q_i(\tau_{t,i}, a_{t,i})$}
      $rand \sim U(0,1)$ {Sample from uniform distribution}
      if $rand \leq \epsilon$ **then**
        Select random action $a_{t,i} \in \mathcal{A}_i$ {Explore with ϵ-greedy}
      end if
    end for
    $\mathbf{a_t} \leftarrow \langle a_{t,i} \rangle_{i \in \mathcal{D}}$
    Execute joint action $\mathbf{a_t}$
    $s_{t+1} \sim \mathcal{T}(s_{t+1} | s_t, \mathbf{a_t})$
    $\mathbf{z_{t+1}} \sim \Omega(\mathbf{z_{t+1}} | \mathbf{a_t}, s_{t+1})$
    $\mathbf{h_t} \leftarrow \langle h_{t,i} \rangle_{i \in \mathcal{D}}$ {Query memory representations of all agents}
    Detach $\mathbf{h_t}$ from computation graph{Avoid additional differentiation through Ψ or Eq. 9}
    $\boldsymbol{\tau_{t+1}} \leftarrow \langle \boldsymbol{\tau_t}, \mathbf{a_t}, \mathbf{z_{t+1}} \rangle$ {Concatenate $\boldsymbol{\tau_t}$, $\mathbf{a_t}$, and $\mathbf{z_{t+1}}$}
    for attention head $c \leftarrow 1, C$ **do**
      $attention_c \leftarrow att_c(\mathbf{h_t})$ {Process individual recurrences according to Eq. 9}
    end for
    $rec_t \leftarrow MLP(\sum_{c=1}^{C} attention_c)$ {See Section 4.2}
    $e_t \leftarrow \langle \boldsymbol{\tau_t}, \mathbf{a_t}, r_t, \mathbf{z_{t+1}}, rec_t \rangle$
    Store experience sample e_t
  end for
  Train Ψ and $\langle Q_i \rangle_{i \in \mathcal{D}}$ using all e_t {See Fig. 1}
end for

Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. "Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability". In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. To appear.



D. Experiment Details

D.1. Computing infrastructure

All training and test runs were performed in parallel on a computing cluster of fifteen x86_64 GNU/Linux (Ubuntu 18.04.5 LTS) machines with i7-8700 @ 3.2GHz CPU (8 cores) and 64 GB RAM. We did not use any GPU in our experiments.

D.2. Hyperparameters and Neural Network Architectures

Our experiments are based on PyMARL and the code from (Rashid et al., 2020) under the Apache License 2.0. We use the default setting from the paper without further hyperparameter tuning as well as the same neural network architectures for the agent RNNs, i.e., *gated recurrent units (GRU)* of (Cho et al., 2014) with 64 units, and the respective factorization operators Ψ as specified by default for each state-of-the-art baseline in Section 6. We set the loss weight $\alpha = 0.75$ for `CW-QMIX` and `OW-QMIX`.

For `MAPPO`, we use the hyperparameters suggested in (Yu et al., 2022) for SMAC, where we set the clipping parameter to 0.1 and use an epoch count of 5. The parameter λ for generalized advantage estimation is set to 1. The centralized critic has two hidden layers of 128 units with ReLU activation, a single linear output, and conditions on *agent-specific global states* which concatenate the global state and the individual observation per agent. The policy network of MAPPO has a similar recurrent architecture like the local utility functions Q_i and additionally applies softmax to the output layer.

`AERIAL` is implemented using `QMIX` as factorization operator Ψ according to Fig. 1. We also experimented with `QPLEX` as alternative with no significant difference in performance. Thus, we stick with `QMIX` for computational efficiency due to fewer trainable parameters. The transformer has $C = 4$ heads $c \in \{1, ..., C\}$ with respective MLPs W_q^c, W_k^c, and W_v^c, each having one hidden layer of $d_{att} = 64$ units with ReLU activation. The three subsequent MLP layers of Line 22 in Algorithm 1 have 64 units with ReLU activation.

All neural networks are trained using RMSProp with a learning rate of 0.0005.

A.10. Emergent Escape-based Flocking Behavior using Multi-Agent Reinforcement Learning

Publication

Carsten Hahn, Thomy Phan, Thomas Gabor, Lenz Belzner, and Claudia Linnhoff-Popien. "Emergent Escape-Based Flocking Behavior using Multi-Agent Reinforcement Learning". In *The 2021 Conference on Artificial Life (ALIFE)*, pages 598–605, 2019.

DOI: https://doi.org/10.1162/isal_a_00226
URL: https://direct.mit.edu/isal/proceedings/isal2019/31/598/99238

Contributions

1. Emergent swarm behavior through self-interested MARL.
2. Analysis of swarm behavior compared to alternative strategies.

Credit

Hahn conceived the concepts and conducted the empirical analysis. Phan initiated the work by stating the hypothesis and discussed the concepts and results. Gabor and Belzner discussed and reviewed the results. Linnhoff-Popien consulted the process and reviewed the results.

Purpose

Main focus of Chapter 6 regarding the research question **Q6** from Section 1.2.