BABEŞ-BOLYAI UNIVERSITY
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

# Multi-Agent Reinforcement Learning Approaches for Complex Game-Theoretic Scenarios

**PhD thesis summary**

PhD student: Mali Imre Gergely
Scientific supervisor: Prof. dr. Czibula Gabriela

2026

# Contents

# List of publications

The ranking of publications was performed according to the CNATDCU (National Council for the Recognition of University Degrees, Diplomas and Certificates) standards applicable for doctoral students enrolled after October 1, 2018. All rankings are listed according to the classification of journals[1] and conferences[2] in Computer Science.

## Publications in conferences proceedings

[Mal25] **Mali Imre Gergely** *A sequence-modeling approach to cooperation in Public Goods Games via Multi-Agent Transformers*. 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025) pp. 3708–3717 (**indexed Web of Science**)

**Rank B, 4 points.**

[Mal24] **Mali Imre Gergely**. *Multi-Agent Deep Reinforcement Learning for Collaborative Task Scheduling*. The 16th International Conference on Agents and Artifcial Intelligence (ICAART 2024), pp. 1076–1083 (**indexed Web of Science**)

**Rank B, 4 points.**

[MC23] **Mali Imre Gergely**, Czibula Gabriela. *Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game*. The 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2023), pp. 345–350 (**indexed Scopus**)

**Rank B, 4 points.**

[Mal22] **Mali Imre Gergely** *Finding Cooperation in the N-Player Iterated Prisoner's Dilemma with Deep Reinforcement Learning over Dynamic Complex Networks*. 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022) pp. 465–474 (**indexed Web of Science**)

**Rank B, 4 points.**

## Publications in international journals

[Mal26] **Mali Imre Gergely** *Self-Play Meta-Reinforcement Learning in Multi-Agent Games*, Acta Universitatis Sapientiae, 18 (5), 2026, pp. 1–17 (**indexed Web of Science**, Q4)

**Rank C, 2 points.**

## Publications score: 18 points.

[1] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole
[2] http://portal.core.edu.au/conf-ranks/

# Introduction

The aim of this thesis is to investigate the problem of engineering socially intelligent multi-agent systems by employing the tools of reinforcement learning and game theory.

Along with supervised and unsupervised learning, the paradigm of reinforcement learning (RL) is one of the core machine learning paradigms. While the latter two are designed to uncover patterns in labeled and unlabeled data that is generally readily produced, RL has its basis in the idea of learning from interaction by interacting with an environment, where an agent is deciding how to best act in order to maximize a numerical signal called reward [SB18]. RL is an intersection of many fields of science. Inspired by neuroscience, psychology and control theory, developed through the tools of mathematics and engineering, RL has seen wide influence from economics and computer science, successful real-world applications ranging from game playing, operations research, resource management, robotics and many others [ADBB17]. Recently, RL has gained tremendous attention in the field of Multi-Agent Systems (MAS) too, where multiple interacting agents are pursuing individual or common goals simultaneously. MAS offers powerful tools to model real-world situations with multiple interacting entities, whereas the instruments of RL offer a mathematically rigorous framework for learning. The combination of the two is labeled Multi-Agent Reinforcement Learning (MARL) which has grown into one of the most popular areas of agent-based research recently [NX24]. Indeed, the toolset of MARL is suitable to serve the goals of this thesis: engineering socially intelligent agents that can pursue common goals with keeping individual objectives in mind, but without falling victim to the selfish shortsightedness of exclusively pursuing individual goals.

To this end, the field of game theory is immensely useful. Initially developed in the 1950's, game theory offers mathematical models to study strategic interactions between rational decision makers [Smi79]. These models are labeled games, while the decision makers are often referred to as players. With key concepts such as strategies, equilibria and optimality, game theory offers rigorous mathematical tools to analyze social situations of interest.

A central issue of building socially intelligent systems is the problem of cooperation and conflict, which has been a long-standing puzzle in many fields of science. Biologists investigate the presence of structured conflict resolution, reciprocal cooperation and altruism in the animal kingdom through the tools of evolutionary game theory [Daw16, SP73, Smi79]. Economists study price wars, auction design and tariff strategies. Politics is interested in embargos, coalition building, negotiating treaties and sometimes, actual wars [AH81]. Game theory can provide valuable insights and help draw useful conclusions in all these fields, while RL provides a sound theoretical basis for developing learning agents.

The previously enumerated fields are mostly interested in the interactions, strategies, intents and actions of existing entities such as animals and humans (or groups of them). Most importantly however, studying strategic interactions becomes more and more relevant nowadays, when the epoch of autonomous agents is on the rise in various fields such as personal assistants, manufacturing robots, trading agents and autonomous driving [LZG+17]. As autonomous agents are ever more embedded into society, understanding the emergence of interaction patterns between learning agents grows more and more relevant. Indeed, designing learning agents that can internalize how to pursue individual goals while maintaining cooperation on the larger scale is one of the key issues studied by this thesis.

MARL with simulated game-theoretic environments is a useful method to study the aforementioned issues for a few key reasons. Mathematically computing the outcomes of the situations studied would be intractable due to the curse of dimensionality. The MARL methods involved are agent-based simulations of adaptation. This scientific method is similar in some key aspects to the two main mathematical methods of induction and deduction. Agent-based methods are conceived on a set of assumptions that serve as guiding principles for designing and running simulations, which then produce data that can be inductively scrutinized (and also used inside the agents' adaptation loops during experiments) [BBD18].

## Research directions

There are three main research directions followed in our research:

- a more theoretical direction in which we aim to understand and foster cooperation in social dilemmas;
- a performance-oriented direction embedded in real-world problems and engineering motivations of game-theoretic MARL systems;
- an all-encompassing general approach in which meta-learning is employed to produce agents that are robust and adaptable not only to cooperative but also competitive situations.

A recurrent topic met in all mentioned directions is the problem of opponent-induced non-stationarity. The multitude of interacting agents introduces an inherent non-stationarity [LWT+17, ZYB21] which yields to either higher variance in the value-function approximators, which in turn leads to worse performance or non-convergence due to the ever-changing nature of the value functions [FFA+17]. Tackling this kind of non-stationarity is recurrent in all our work, with a significant impact on the emergence of cooperation in social dilemmas, on real-world applications and on the performance and convergence of meta-learning agents as well.

The three main research directions are detailed in the following.

**Learning to Cooperate in Social Dilemmas with MARL.** As a first line of research, we developed multi-agent deep reinforcement learning approaches for *improving existing RL-based algorithms* with the goal of finding cooperation in social dilemmas.

While RL algorithms have long been the subject of academic scrutiny, the field still holds gaps in several theoretical aspects. More specifically, traditional RL algorithms deployed in environments where agents are met with dilemmatic situations of cooperation and conflict (of self-interest or collective benefit) are known to fall into the short-sighted local optima of selfishness and defection.

We target the pursuit of joint cooperative behavior in social dilemmas with MARL. Social dilemmas are situations that create conflict between the interests of the individual and the interests of the collective. These situations generally expose reward structures where selfish individuals are rewarded, yet collective selfishness leads to unfavorable outcomes on the long run for the collective and thus, the individual. While such situations mainly arise in society, individual agents in MARL settings often face similar situations (especially decentralized MARL, where agents learn and act independently). Most often these situations boil down to cooperation-competition type dilemmas in multi-agent systems such as peer to peer systems, file sharing, packet routing, wireless sensor networks etc., problems which are often intended to be tackled with reinforcement learning [SRAA+11]. While the exact details of the specific problems have numerous requirements like equipment and setup, the sheer problem of dilemmas can be abstracted away and be studied separately. This is usually done by employing game theory. The most famous social dilemma modeled by game theory is the Prisoner's Dilemma, however many other models have been proposed to capture different social dilemmas, such as the snowdrift game, the hawk-dove game, etc. A large portion of these games are worth investigating via MARL. Of these, our research focuses on the Prisoner's Dilemma and the Public Goods Game.

Furthermore, while RL algorithms have reached or surpassed the level of performance reached by humans in perfect information games [SSS+17, SHS+17], it is still challenging to find approximate Nash equilibria in imperfect-information games of large scale. Methods attempting to tackle this problem most frequently include some form of Neural Fictitious Self-Play [HS16], which however still often suffers of the curse of dimensionality, when approached with value-based RL methods [ZWLP19].

The problem becomes even more emphasized in fully decentralized systems where agents have no means of sharing information during training and the only source of signal is the reward.

This sub-direction materializes in our research in addressing the problem of decentralized MARL with networked agents (networks being graphs of agents with connections and relationships between them as edges). Decentralized MARL employs single-agent RL algorithms separately and independently. Another challenge from real-world problems included in networked agents settings is the limited communication among agents [FFA+17]. While most approaches focus explicitly on the RL methods employed, the nature of the interconnection graphs of the agents also has significant influence on the overall behavior and thus, performance of the system [BPB11]. Moreover, dynamic interconnection models (with time, agents change their relative positions to one another and thus, the neighbors who they communicate with) can also have significant influence [PBARA14]. It is particularly interesting to analyze the influence of the agent interconnection on the overall performance and indeed, our study [Mal22] of the N-Player Iterated Prisoner's Dilemma (NIPD) studies exactly that: the emergence of cooperation in dynamic networks of RL agents where interconnections change stochastically according to agents' cooperation rates. However, the introduction of cooperation-encouraging dynamicity is arguably an external mechanism beyond the reward structure. Our next work on the Public Goods Game [Mal25] in turn uses no such additional mechanism and employs modern Transformer-based multi-agent methods instead, which using agent histories similarly to [Mal22] obtains considerable amounts of cooperation.

**Applications and benchmarks of game theoretic MARL.** While MARL has witnessed wide success in a plethora of domains, testing and benchmarking different MARL algorithms have been a pain point of the field for long [TDM+18]. Conducting experiments on standard environments was made difficult by custom implementations which are not only wasted engineering effort but also hindering comparability between different studies. To combat this, several standards, API-s, environments and benchmarks have been proposed, the most remarkable being OpenAI Gym [BCP+16], RLLab [DCH+16], and AxelrodPy [pd16] for game theory. Apart from the Axelrod Project however, game-theoretic benchmarks are scarce, and the existing ones mostly target infamous social dilemmas like the Prisoner's Dilemma, and games with mixed-strategy equilibria are left out of focus. One of the proposals of this thesis is aiming to bridge exactly this gap: we generalize the notorious cycling game called Rock-Paper-Scissors to higher numbers of players and actions to create an environment that is tunable in strategic complexity but applicable to arbitrarily large numbers of agents, providing a useful benchmark for MARL agents that aim to learn mixed strategy equilibria [MC23].

Another line of work under this direction is the application of game-theory-based MARL to real-world problems in order to showcase how the problems of cooperation and conflict on the one hand, and the need for disambiguation by game-theoretical models such as Agent-Environment Cycle games appear in real scenarios. To pursue this, we apply MARL in our original work [Mal24] to the problem of task scheduling in cloud computing, where multiple machines consume from a shared job queue.

**Meta-Learning Applied to Game Theory.** While a significant part of research is focused on solving single environments, solving particular real-world issues or concentrate on solving singular social dilemmas, there is a need for MARL systems that can adapt to new situations as quickly as possible. For example, provided that circumstances are adequate, sports players can successfully apply their skills on any playing field, a cook can prepare meals in any kitchen, a musician can perform on any stage and a surgeon operate in any operating room. In contrast, in RL the test and validation environments are the same, thus agents overfit to one particular environment by design. However, the development of agents that are not only capable to account for opponent-induced non-stationarity but are also able to deal with changing environment landscapes are paramount in economics, politics and many other fields that are subject to constant change.

To overcome this challenge, meta-learning frameworks have been proposed that aim not to merely solve a specific task but instead adapt to a distribution of tasks with the goal to equip the trained agents with a good prior that makes adaptation to a specific task quicker, which in the field of RL change the traditional paradigm and introduce inner and outer loop trainings. Inner loops are traditional RL-adaptation steps, while the additional outer loop samples different environments from some distribution, while knowledge is retained via hidden states. The result is meta-learning agents which are capable of adapting to different environments by minimal amounts of interactions and surpass the limitations of traditional RL by avoiding catastrophic forgetting.

Thus, another promising direction we pursue in this thesis is the study of multi-agent meta reinforcement learning in multi-agent games, where we seek to find algorithms that produce agents able to co-adapt in different games where the payoffs and strategic landscapes are subject to perpetual change.

## Original Contributions

A general aim of this thesis is to move towards adaptive, cooperative and socially aware intelligence that manages to productively navigate multi-agent problems. To showcase the potential of different MARL methods in this pursuit, social dilemmas, cycling games, real-world applications and meta-learning setups were considered, which are detailed in Chapters 2, 3 and 4, and summarized below:

**MARL-based cooperation in Social Dilemmas** Traditional RL methods are known to converge to defective Nash equilibria without additional mechanisms that encourage cooperation. Our approaches aim to find cooperation in the N-Player Iterated Prisoner's Dilemma and the Public Goods Game:

- The fully decentralized value-based RL method mixed with dynamic complex networks presented in Section 2.1 based on our original work published in [Mal22] is conducted on the NIPD, and it is the first study that investigates this particular social dilemma under dynamic complex networks, where agents are initially arranged into different kinds of complex graphs, but edges evolve over time probabilistically according to cooperation rates. Agents take their actions of cooperation or defection based on the past actions of their current neighbors in the agent graph. Under this scenario, we are the first to demonstrate that cooperation can emerge and offer a detailed analysis of its extent based on past actions and initial neighborhood structures. Another contribution of this paper is the novel non-rigid state representation for deep RL methods that is naturally adaptable to a variable number of opponents.

- Our results obtained on the Public Goods Game, presented in paper [Mal25] and discussed in Section 2.2 is according to our best knowledge the first to successfully apply Transformer-based architectures such as Multi-Agent-Transformers (MAT) [WKL+22] to social dilemmas. MAT models consider multi-agent action selection as a sequence modeling problem and uses attention mechanisms to track how agents' choices affect each other. Since MAT uses the advantage decomposition theorem to decompose the joint objective into a sequence of individual objectives without losing scope, our use of the method in social dilemmas is well motivated. Similarly to our work on the Prisoner's Dilemma [Mal22], agents' observations are constructed from opponents' histories. Under these conditions, we empirically show that first, cooperation can be achieved by using attention-based mechanisms, and second, that feeding longer opponent histories into observations helps foster cooperation considerably more.

**Benchmarks and applications of Game-Theoretic MARL** MARL has become immensely popular, although the wealth of topics on which otherwise generic algorithms are applied make it hard to reliable compare results. To this end, several benchmarks such as SMAC [SRDW+19], OpenSpiel [LLL+19] and game-theoretic benchmarks [LSB+23] have been proposed. However, while benchmarks and theoretical testbeds offer a skeleton for strategic interaction that are reliable for proofs of concept, the reliability and viability of MARL methods need to be verified in more realistic scenarios that depart from the idealistic assumptions that game theory often presumes. In the pursuit of these lines of research, our results are as follows:

- We propose a strategically diverse, versatile, tunable and mathematically grounded testbed for MARL research in [MC23] and present it in Section 3.2. A crucial contribution of this work is that it successfully generalizes the notorious Rock-Paper-Scissors game to arbitrary numbers of agents and arbitrarily high odd number of actions, and mathematically proves the games' Nash equilibrium. This offers a testbed that is more general as the simultaneous interaction of multiple agents (which is core to MARL and was previously unaddressed in similar works such as [LSB+23]) naturally incorporable. Another noteworthy contribution of this work is our empirical results which demonstrate that the standard PPO [SWD+17] algorithm exhibits behavior analogous to economic price cycling when applied to this benchmark independently in each agent.

- In [Mal24] we demonstrate that game-theoretic MARL can be successfully applied to real-world applications and detail our contributions in Section 3.1. Specifically, we generalize the task scheduling simulator proposed in [MAMK16] to multiple machines and apply an Agent-Environment-Cycle-games [TBG+21] approach that is able to handle multiple agents with multiple machines per agent. Our results show that the joint scheduling policies found by agents are at comparable levels with the ones found by heuristics, offering an approach that can collaboratively scale to multiple agents simultaneously.

**Meta-MARL Approaches in Game-Theoretic Scenarios** A frequent criticism of RL is that the test and evaluation environments are the same [Wen19] and thus standard RL training is inherently overfitting. To combat this, Meta-Learning approaches have been proposed, however this problem is even more relevant in MARL, where besides the moving target of the environment, the non-stationarity inherent to multi-agent approaches makes learning and convergence enven more difficult. To address these issues, we proposed a meta-learning approach as follows:

- In our original work [Mal26] we propose a fully decentralized meta-MARL method that showcases adaptive behavior in both cooperative and adversarial scenarios and detail our results in Section 4.1. Specifically, we are the first to adapt a standard Meta-RL method called $RL^2$ [DSC+16] to a fully decentralized self-play meta-reinforcement learning setup, and offer two classes of game-theoretic environments. We use distributions of normal form games to create diverse sets of strategically alike but concretely different situations of cooperation (random coordination games) and conflict (random zero-sum games). Our results empirically show the ability of our proposed approach to produce agents that are sample-efficient, robust to changes and capable of strategic generalization across diverse sets of game-theoretic scenarios, as measured by the Nash-regret or Nikaido-Isoda metric [NI55].

## Thesis Structure

The rest of this thesis is organized as follows: Chapter 1 presents the necessary theoretical background, offering a general introduction to the fields of Reinforcement Learning, Multi-Agent Systems, Meta-Learning, Game Theory and Complex Networks.

Then, Chapter 2 introduces the detailed motivation for studying cooperation and the pursuit of joint cooperative behavior via RL agents in social dilemmas. Section 2.1 presents our first, value-based RL approach in a generalized and iterated form of the Prisoner's Dilemma [Mal22]. The section details our experiments over several network types, agent history lengths and agent rewiring behavior, along with comparison to related work. Next, Section 2.2 presents our next work on seeking cooperation in social dilemmas. This section is based on our original contributions according to our paper [Mal25], detailing the motivation of our model choice, discussing the obtained results and comparing to related work.

Further, Chapter 3 discusses our benchmark proposal and results in real-world applications. Section 3.1 discusses our multi-agent collaborative task-scheduling proposal presented in [Mal24]: we discuss motivations, present our general framework and discuss our results in comparison with existing task-scheduling heuristics. Then, Section 3.2 details our benchmark proposal published in [MC23], provides a mathematical proof of its equilibria and presents the results of our experimental evaluation.

Chapter 4 presents our original contributions in the field of meta-learning, embedded into MARL and game-theory as presented in [Mal26]. This chapter discusses the relevance of meta-learning, presents our environment distribution proposal based on normal-form games and evaluates performance based on Nash-regret.

Finally, the conclusions of the thesis and possible lines of future work are outlined.

# Chapter 1

# Background

The purpose of this chapter is to provide an introduction to the theoretical background for the topics and problems covered in this thesis, and to present the current state of recent solution methods and open problems respectively. Namely, the fields of reinforcement learning along with multi-agent systems will be presented as the core theoretical framework for multi-agent reinforcement learning, the toolsets of which encompass the entire thesis. Another core concepts of the present work is game theory, the descriptive and intuitive power of which is employed to shed light on different aspects of those multi-agent interactions the understanding of which constitute some of the main contributions of this thesis. Further, an introduction will be given to the field of complex networks, meta-learning as auxiliary fields employed in some of the covered research directions, while an introduction to the field of task scheduling is given as context for the applicative part of our work.

## 1.1 Reinforcement Learning

Reinforcement Learning (RL) is one of the three fundamental paradigms of machine learning. While the others (unsupervised and supervised learning) consider already collected and curated data, reinforcement learning gains data and experience from direct interaction. Learning from interaction is one of the fundamental ideas embracing the majority of theories about learning and intelligence [SB18]. Also, it is the most straightforward and concise definition of what we call *reinforcement learning*. Due to the lack of fixed human-designed and labeled outputs, the performance of RL is generally not bounded by the quality of existing expert data, but by the ability of the RL algorithm to successfully identify and exploit the patterns underlying the environment that can be used to make decisions which finally maximize reward. RL has seen tremendous progress and has attracted enormous attention in the past years. A multitude of real-world sequential decision making problems have successfully been solved by reinforcement learning approaches. Such problems include autonomous driving, software engineering, drone flying, robotics, playing games like Go, Chess and Poker, etc. [Li19].

While in other machine learning paradigms the core concepts are the learning algorithm and datasets, RL generally conceives a worldview consisted of agents and environments. The learning components of this setup lie within the agent, which acting inside the environment receives observations and rewards, then decides on the next action, thus creating a closed loop (Figure 1.1).

The main objective of the agent is to choose actions in a way that maximizes cumulative reward. From what framework an agent follows to make decisions, one can distinguish value-based, policy-based and actor-critic agents. A value-based agent's policy is implicit, as it can directly read out the best action from its value function. In contrast, a policy-based agent stores a policy directly and selects actions without ever explicitly storing the value-function. Actor-critic agents combine the

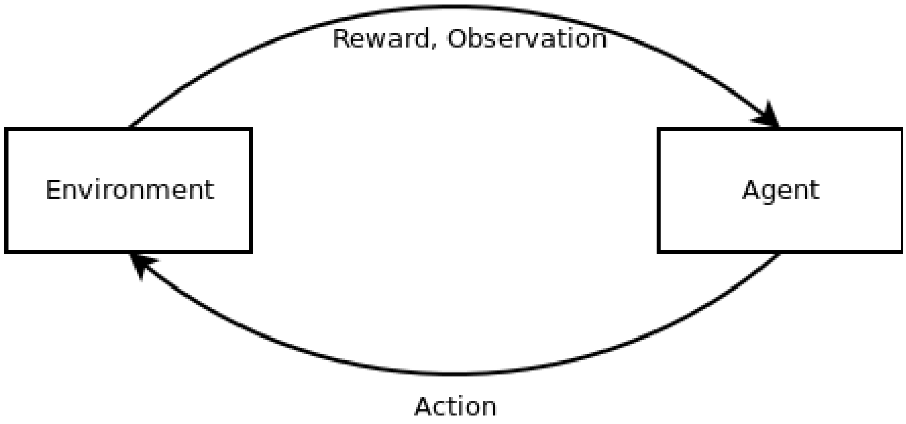


Figure 1.1: Agent-Environment loop

previous two (policies and Q/V-functions) and try to get the best out of both.

Another key classification of RL agents can be done based on knowledge of the environment: model-free and model-based agents. A model-based agent has full knowledge about the environment's dynamics and states, its model of the environment is complete and has only to learn the proper behavior in it, whereas model-free agents have an incomplete model of the environment(or none at all) and have to approximate the environment while learning to maximize reward. Most real-world problems addressed by RL employ model-based agents since the most problems of interest are infeasible to fully and tractably describe and/or too large to explicitly represent.

Everything inside the system, outside the agent is said to be part of the environment. There are multiple types of environments, depending on which the accuracy of the agent's model of it is also defined. Russel and Norvig [RN16] propose the following environment properties:

- deterministic or non-deterministic;
- discrete or continuous;
- episodic or non-episodic;
- static or dynamic;
- fully or partially observable;
- single or multi-agent.

## 1.2 Multi-Agent Systems and Multi-Agent Reinforcement Learning

Multi-Agent Systems (MAS) are setups where multiple agents act in a shared environment. Agents interact with the shared environment but also act among themselves and pursue some goal, which can be shared across all, some or none. Generally, agent-based systems and MAS too are described by agents that are autonomous, meaning no direct human intervention. This presupposition organically aligns with the nature of RL agents. Over time, MARL has shown immense potential and applicability in many applications in several domains such as robotics, medicine, agriculture, economy, etc. [ZYB21, HY20, HPHKPI20]. The application of MARL to these and other problems is justified by its effectiveness in breaking the curse of dimensionality, in leveraging the natural fault tolerance of MARL and sometimes by adherence to environment-specific constraints. In spite of these advantages, MARL is not short of its own set of challenges, the most prevalent of these being non-stationarity, scalability, partial observability, credit-assignment and coordination/competition.

First, non-stationarity is a core issue that comes directly from the multiplicity of agents. For each agent, the set of peers is considered as part of the environment. Since these agents in their turn are also simultaneously learning with them, the learning problem in MARL becomes a constantly moving target which makes convergence guarantees considerably more difficult to offer.

Second, the more agents a MARL system has, the bigger the joint action and observation spaces become. While a sensible number of agents does help in breaking the curse of dimensionality in some cases, this problem can come back at the other end when the number of agents grows too high and convergence can be hindered.

Third, most MARL systems consider agents that are able to observe only a local slice of the environment. This makes the application of RL algorithms particularly difficult when it comes to cooperation and/or competition.

Fourth, credit assignment is a central issue of MARL. When rewards are given in a shared manner, it gets difficult to decide which agents' actions lead to the particular reward. Improper credit assignment can make some agents learn while others not, mistakenly attribute reward to mistakes or vice-versa.

Lastly, the problem of coordination and competition is perhaps the one that grew most notorious. This notoriety is mostly attributable to the direct intuitions and parallels with social situations and dilemmas. In many cases, agents need to cooperate in order to achieve a common goal. Since all agents are learning at the same time (in other words, the problem of non-stationarity), learning cooperative joint policies gets increasingly difficult. In other situations, the objectives of some agents might be in conflict with each other, meaning they find themselves competing with each other: this often ends in non-converging cycles or agents overfit to objectively underperforming opponents.

To address the main intricacies of dealing with systemic bottlenecks in MARL, the learning process of each agent needs to account for the behavior of others, which in terms of training and execution leads to several solution architectures [HM23]:

- **Centralized Training, Centralized Execution**: this paradigm practically reduces the multi-agent setting to a single-agent one, where the multiplicity of agents, actions and observations is hidden under a single agent whose behavior is searched in the joint policy space.

- **Centralized Training, Decentralized Execution**: agents learn using a centralized brain but each acts on its own in view of their individual observations. This paradigm has the advantage of globally sharing experience during training but yet benefiting of agent separation through decentralized actions. This paradigm is prevalent in cooperative models such as MAPPO [YVV+22]

- **Decentralized Training, Decentralized Execution** or fully decentralized paradigm considers each agent with its own separate and independent learning and execution unit. Under this paradigm agents can't directly share experience and from the perspective of each agent, the others are truly independent parts of the environment. This separation often comes with the disadvantage of instability and divergence, however the independence of agents makes this model more scalable than the others.

To formally capture the intertwinement of a setting where multiple RL agents are acting, one can use an MDP generalization, called Markov Games (sometimes referred to as stochastic games) [Sha53]. Markov games have been introduced in [Lit94] as a general framework for MARL.

Formally, a Markov game [Lit94] is defined as a tuple $< N, S\{A^i\}_{i\in\mathbb{N}}, P, \{R^i\}_{i\in\mathbb{N}}, \gamma >$. Here $N$ denotes the set of agents($\geq 2$), $S$ denotes the set of states observed by all agents. $A^i$ and $R^i$ denote the action and reward functions for agent $i$, whereas $P$ and $\gamma$ bear their usual meanings from Markov Decision Processes: the transition probability function and the discount factor respectively [ZYB19].

The main objective of each particular agent is still the optimization of their individual cumulative reward. At each timestep indexed as $t$, from state $s_t$ every agent $i$ chooses and performs an action $a_t^i$, transitions to $s_{t+1}$ and receives their reward according to $R^i(s_t, a_t, s_{t+1})$. The value function $v^i : S \rightarrow \mathbb{R}$ is dependent of the joing policy defined as $v : S \rightarrow \Delta(A)$, written as $\pi(a, s) := \Pi_{i\in\mathbb{N}}\pi^i(a^i|s)$

Similarly to the single-agent setting, but taking the multitude of actors into account, one can define the value function for a joint policy $\pi$ and state $s$ as follows:

$$V^i_{\pi^i,\pi^{-i}}(s) := \mathbb{E}\left[\sum_{t\geq 0}\gamma^t R^i(s_t, a_t, s_{t+1})\middle| a_t^i \sim \pi^i(\cdot|s_t), s_0 = s\right] \tag{1.1}$$

Here -$i$ marks the indices of all agents with the exception of $i$. The fact that makes Markov Games exponentially more complex than simple MDPS is that the optimal behavior depends on more than the behavior of the current agent's own, and the behavior of all other agents plays a crucial role too. Due to this fact the concept of solution is also different from conventional MDP solutions. The most common solution is closely related to the concept of Nash equilibrium [BO98]. The Nash equilibrium of a Markov game is defined as a joint policy $\pi^* = (\pi^{1,*}, \pi^{2,*}, ..., \pi^{n,*})$ which for any state $s \in S$ and $i \in N$ respects the following condition:

$$V^i_{\pi^{i,*},\pi^{-i,*}} \geq V^i_{\pi^i,\pi^{-i,*}}, \forall\pi^i \tag{1.2}$$

This equilibrium is a point in the joint policy space, with the property that no agent has deviation incentives from it. The Markov Game abstraction of MARL tasks is general enough to capture multiple different scenarios, such as cooperative, competitive or mixed settings, as follows [ZYB19]:

- **Cooperative setting**: usually all agents share the same reward function, this setting is also known as *multi-agent MDP* [LR00]. In such settings, single-agent RL algorithms can be used, since the value of the Q-function is common to all agents, however, this also imposes the restriction of some level of homogeneity on the agents. However, cooperative Markov games can further be generalized to models that instead considers *team average reward*, which also allows for different agents to have private and distinct reward functions. This generalization, in contrast with the previous one allows more heterogeneity among agents, and facilitates the development of decentralized multi-agent reinforcement learning, a very ambitious goal pursued in the multi-agent reinforcement learning literature [KMP13, WYWH18].

- **Competitive setting**: usually modeled as zero-sum Markov games, most frequently with two players [SSS+17, Lit94] where wins and losses of the two players are symmetrical. Zero-sum games however, are also often used for robust learning, because the environments' uncertainty that slows down the learning or that can be seen as an impediment for the agent, can be attributed to an imaginary opponent in a game that plays against the agent.

- **Mixed setting**: is a generalization of the cooperative and competitive settings, modeled as *general-sum* games where each agent acts in a self-interested manner, and goals of different agents can conflict, or overlap. Having a set consisting of fully cooperative and fully competitive agents, for example teams of cooperative agents competing against each other [BKM+19, BBC+19] can be considered mixed-motive.

Another useful abstraction for MARL tasks is the one of **extensive form games**. These games are used to model games with no perfect information, and are defined as a tuple $< N \cup \{c\}, H, Z, A, \{R^i\}_{i \in N}, \tau, \pi^c, S >$. Here *N* marks the set of agents (c being a specially introduced chance of nature agent with a fixed stochastic policy [ZYB19]). *A* is the set of possible actions, while *H* is the set of histories, one history denoting a series of actions taken throughout one episode. *Z* is a subset of *H*, and denotes the set of terminal histories, according to which $R^i$ assigns utilities. $\tau$ is the identification function specifying which agent took what action. The elements of *S* are called information states.

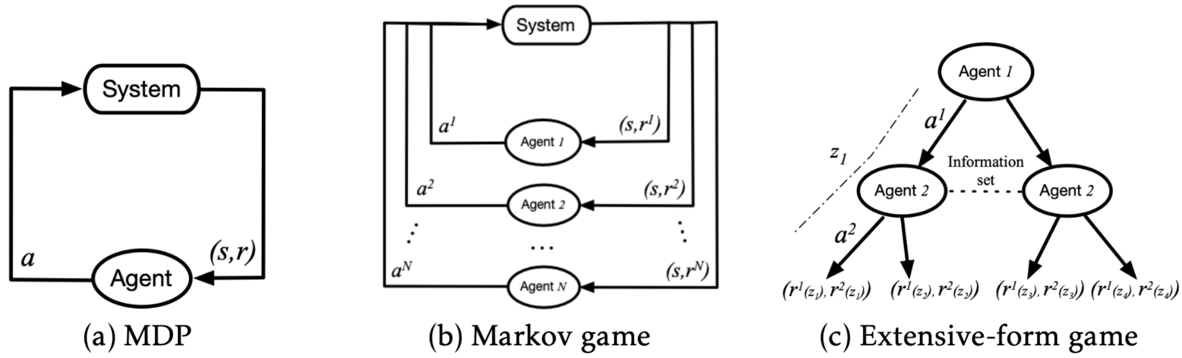


Figure 1.2: Markov process generalizations for reinforcement learning [ZYB19]

The goal of an extensive form game can be formalized as an $\epsilon$-*Nash equilibrium*, an approximation of the true Nash equilibrium, which is reached when $\epsilon = 0$. The $\epsilon$-*Nash-equilibrium* is defined as follows:

$$R^i(\pi^{i,*}, \pi^{-i,*}) \geq R^i(\pi^i, \pi^{-i,*}), \forall \pi^i \; of \; i \tag{1.3}$$

Figure 1.2 summarizes the types of Markov-process abstractions detailed above used for reinforcement learning, capturing from the simplest, single-agent settings to the more complex multi-agent settings.

## 1.3 Game Theory

The study of situations involving strategic interactions between rational decision making entities [Mye91] by means of mathematical models is called game theory. It is the formal study of conflict and/or cooperation among several players (people, groups, animals, companies, people, computers, etc.). From a theoretical perspective games can be thought of as models of interactive situations among rational players. The modeling of games provides the appropriate mathematical tooling and framework to understand the behavior of the players, when pursuing some goal in the game. While this domain has applications in many fields [Byu14, Pic19, FS03, JSB02, Cow12, Har09] and helps to formalize a large number of behavioral relations, it is of particular interest in artificial intelligence.

Games can be classified in several ways, among which the most important are:

- cooperative or non-cooperative;
- zero-sum vs general-sum;
- simultaneous or sequential;
- symmetric vs asymmetric;

- perfect or imperfect information.

In 1950, John Nash proved[N+50] that every finite game admins at least one equilibrium point, where every player (knowing the opponents' choices) the best action for themselves and have no incentive to deviate from their strategy. Nash's discovery made game theory a very active area of research for a long time.

the best response is the concept denoting the strategy of a given player, assuming the opponent(s)'s strategy, that makes the most favorable outcome for the player the most likely. A strategy is called dominant if it always outperforms the other(s) in terms of utility. Best responses were already introduced in the previous sections in RL-terms, here we revisit them in a purely game-theoretical perspective.

A set of strategies among players is said to be Pareto optimal or Pareto efficient if a better utility/payoff for any player cannot be achieved by resulting in a worse utility for some other player [BBD18]. This means that if rational players find themselves in a non-Pareto optimal situation, they do have rational incentives to move to a better one, as it can be done without hurting the payoff of any other player. Pareto efficiency, however, is a weak notion of optimality. The concept of efficiency is a much stronger notion, instead, and it denotes the set of strategies that maximizes the total utility of outcomes across all players.

A **Nash equilibrium** is said to be the set of strategies that for any agent would not be worth deviating from. Since agents are considered self-interested, it is very important to note that a **Nash equilibrium** might not be efficient and an efficient set of strategies might not be a Nash equilibrium.

While games can be formally represented in several ways (Normal form, Extensive form, etc.), the most straightforward formalization and the one most frequently used within our line of work is the Normal Form.

A game in Normal form is a tuple $< N, A, u >$. Here $N = \{1, 2, ...n\}$ is the finite set of players, $A = A_1 \times A_2 \times ... \times A_n$ is the aggregate action space of all agents, having each $A_i$ as the set of possible actions for player $i$. Then, $u$ is the set of utility functions $(u_1, u_2, ..., u_n)$. Each $u_i$ is defined as $u_i : A \mapsto \mathbb{R}$, mapping a joint action profile to some numerical payoff value.

A strategy profile $\sigma_i \in \Delta(A_i)$ defines how a player chooses actions, where $\Delta(A_i)$ is the set of every possible distribution of probabilities over $A_i$. If such a distribution contains only one element with probability 1, it is called a **pure strategy** whereas all other strategies are called **mixed strategies**. An action profile $a$ is a set of joint actions $(a_1, a_2, ..., a_n)$ for all players. Frequently, the notation $(a_i, a_{-i})$ or $(\sigma_i, \sigma_{-i})$ is used to denote the actions/strategies chosen by $i$ in contrast with the actions/strategies chosen by all other players such that $a_{-i} = (a_1, ...a_{i-1}, a_{i+1}, ...a_n)$ and $\sigma_{-i} = (\sigma_1, ...\sigma_{i-1}, \sigma_{i+1}, ...\sigma_n)$

Under these definitions, a strategy profile $\sigma_i^*$ is a best response to $\sigma_{-i}$, if:

$$u_i(\sigma_i^*, \sigma_{-i}) \geq u_i(\sigma_i, \sigma_{-i}), \forall \sigma_i \in \Delta(A_i) \tag{1.4}$$

If Equation 1.4 holds for all players, the strategy $\sigma$ is called a **Nash equilibrium**.

Game theory has been pivotal in uncovering valuable insights into various social situations that can be rigorously modeled through its toolset. The most well-known among these are called social dilemmas, where the collective good and individual self-interest of players are in a constant conflict. These situations are of particular interest in MARL too.

In the context of learning, equilibrium concepts serve both as benchmarks and guiding principles to evaluate and interpret agent behavior. While games can be represented in multiple ways (such as extensive form, graphical form, or state-based formulations) we focus on normal-form games, where each player selects a single action per interaction from a fixed action set represented by so called payoff matrices. This representation is not only compact and expressive for simultaneous-move games, but also lends itself well to parameterization and efficient sampling, which are essential for large-scale experimentation, MARL applications and meta-learning.

It is of utmost importance to distinguish the types of social dilemmas considered, and this is most clearly done through a comparison among social dilemmas that had received the most academic scrutiny, allowing a deeper understanding of the games in question. While numerous game-theoretic models aim to capture the conundrum of individual against collective good, each model does captures it distinctly. In order to shed some light on these nuances, we briefly consider the incentive structures in the most famous models, such as the Prisoner's Dilemma,the Public Goods Game (PGG), Stag Hunt and the game of Chicken [FJW24].

# Chapter 2

# Reinforcement Learning Based Approaches to Cooperation in Social Dilemmas

Cooperation as a social phenomenon is the animal kingdom and human societies is often viewed as an evolutionary puzzle. Why do individuals end up cooperating, often sacrificing their own good when taking advantage of their peers looks often far more rational? Why does cooperation in other cases fail spectacularly by collapsing into mindlessly selfish individuals that destroy common good? What are the prospects of cooperation in a society ever more embedded into systems with artificial agents?

The quest for understanding cooperation and possibly encourage it in certain situations has been ongoing for decades. The question is pressing from many directions: understanding cooperation has direct impact on political and economical systems, impacting stability, international relations, trade wars and the management of depletable public goods. Second, from the standpoint of biology and psychology, it is important to understand how non-rational beings like insects and animals cooperate, but also the motivations and conditions under which humans develop cooperative behavior. Most relevant to our thesis however is the perspective of artificial intelligence: as the era of automation is blooming, more and more autonomous agents are deployed which have to be successfully integrated into a system where agents need to collaborate with each other and with other humans.

This central focus of this chapter is to present two results in the direction of finding joint cooperative behavior in multi-agent systems powered by RL agents. Since the pioneering work of Axelrod [Axe81] in the Prisoner's Dilemma, the research field of agent-based simulations has become more and more popular. The addition of learning algorithms to this trend has brought many intriguing results. Most of these works explore how some learning agents can cooperate under the rules of certain social dilemmas like the Prisoner's Dilemma, the Public Goods Game, the Snowdrift Game or the Stag Hunt Game. However, most learning-agent based approaches fail to converge to cooperative equilibria. To this end, many additional mechanisms have been added to the setup, such as the rules of the game, the agents themselves or the reward structure with the goal of encouraging the emergence of cooperation.

Our first work, based on our original approach in [Mal22] on the Prisoner's Dilemma game is such an approach: we introduce a networked-agents approach where the agent interconnections are tweaked in such a way to encourage cooperation. Our second approach, applied to the Public Goods Game based on our original work in [Mal25] however, uses no additional mechanisms and yet achieves considerable cooperation rates.

# Chapter 3

# Proposed benchmarks and applications

Multi-Agent Reinforcement learning has witnessed a rapid development in the past years. The field gained significant insights not only from applications to concrete real-world problems but from benchmark domains as well. Successful and promising MARL systems have been developed in the fields of cellular networks, computer vision, autonomous driving, traffic control, energy management, cloud computing and many others. Further, proposed benchmarks such as Gym [BCP+16] environments and games such as Atari, Chess, Go, Poker have been immensely useful in assessing the performance of MARL while keeping experiments comparable, reproducible. Benchmarks also offer settings that are highly controllable, straightforward to reason about and that reduce engineering overhead.

The intent of this chapter is presenting two results of our work from the line of study of benchmarks and applications.

In our original approach [Mal24] we present a direct application of game theoretic MARL in the field of task scheduling. These results demonstrate that game theoretic models point above mere theoretical scrutiny and well-crafted game theory inspired MARL systems that explicitly prioritize aspects such as cooperation via the usage of Agent Environment Cycle Games can have a significant impact on the performance of real-world problems such as the field of task scheduling in cloud systems.

Second, we detail our benchmark proposal for MARL algorithms based on our original approach introduced in [MC23]. This proposal draws inspiration from the classical game of Rock-Paper-Scissors, and offers a flexible and controllable environment to test agents in settings where mixed strategies need to be learned. Further, two key original contributions of this work need to be noted: on the one hand, the successful and sound generalization of the game to arbitrary number of agents and actions while maintaining the motivational structure of the game and second, the mathematical demonstration of the existence and identification of the Nash equilibria of the generalized Rock-Paper Scissors game.

# Chapter 4

# Meta-Reinforcement Learning in Multi-Agent Games

Interactions in multi-agent systems are often framed through the tools of game theory; however, in real-world scenarios, the structure and parameters of the underlying game faced by agents are frequently unknown or non-stationary. This presents a critical challenge: agents must rapidly infer the nature of their environment and adapt their strategies accordingly, even in the presence of multiple other agents.

Meta-reinforcement learning (meta-RL) has demonstrated the ability to facilitate fast adaptation in tasks such as multi-armed bandits, Markov decision processes, and visual navigation. In this chapter, we present the results of our original work [Mal26] where we extend the application of meta-RL to multi-agent games. By training agents via self-play meta-reinforcement learning on diverse classes of normal-form games, parameterized by their payoff matrices and sampled from a distribution, we develop algorithms that are not only sample-efficient and robust to changes, but also capable of strategic generalization across distinct game-theoretic structures. Although it remains limited to a theoretical proof of concept, our approach bridges the gap between classical game-theoretic modeling and modern meta-learning techniques, with promising implications for adaptive behavior in dynamic multi-agent environments.

# Conclusions

Reinforcement learning is a ubiquitous area of research that has had a large influence on many other fields. Since RL is not constrained by domain knowledge which usually marks a margin plateau for supervised learning techniques, RL is said to have a very high potential for building systems that surpass human performance, especially in fields where human-like learning and behavior is desired, but where manual human intervention was previously infeasible.

Nonetheless, the field still holds plentiful open problems and is still awaiting solutions for various problems of real life interest. Similarly, for many problems already employing RL techniques there is still a wealth of areas where enhancements can be made, or where more investigation is necessary, not mentioning the fact that even RL theory itself carries many open problems, whose eventual solutions might significantly impact the evolution of the field.

Specifically, the application of reinforcement learning in game theoretic models in multi-agent systems has far-reaching implications in real-world problems that extend far beyond theoretical research. The ever increasing adoption of artificially intelligent agents render the importance of game-theoretic models larger and larger. RL's applicability and potential is substantial in the digital sphere, be it optimizing recommendations or dynamic resource management. Its capacity to navigate and optimize in complex and changing environments such as logistics and supply-chain efficiency make it an attractive tool to employ. Healthcare and finance are also heavily impacted: ptimising treatment plans and balancing patient needs, automatically handing budgets and designing investment plans that successfully adapt to changing market conditions are also of high impact. RL-tailored education plans, RL-driven research on protein folding and the rise of autonomous driving systems all prove society's increasing readiness to delegate more and more control to artificial intelligence, much of it driven by RL.

Much if not all the aforementioned areas however begin to employ possibly more decision-making agents whose coordination, cooperation and/or competition is a problem rising along with adoption. In this context, the issues rising from multiple interacting agents, like non-stationarity of MAS becomes crucial to handle. Insights into the intrinsic incentives of these agents also need to be gained, and agents' motivations needs to be understood in order to have granular control over them. Similarly, the transparency of these systems needs to be understood in order to ensure that their operation respects safety regulations. Game-theoretic models help immensely to address and investigate these problems by describing scenarios that arise in real-world situations, without necessitating to carry all details of specific problems like the ones enumerated previously. Issues like transparency, adaptability, cooperation, coordination, conflict and motivation are cross-cutting concerns across all multi-agent systems, that can successfully be captured by game-theoretic models.

The thesis aimed to present our original contributions in employing RL, with a specific focus on deep RL in game-theoretic models. These promising results are but the tip of the iceberg, as we believe that the full potential of DRL in game-theoretic models is yet to be fully harnessed.

We focused not only on how these models can be exploited to harness decision-making processes in multi-agent scenarios on an abstract and generic level, but also showed how incorporating them into real-world applications can lead to more clarity and performance improvements, such as cloud task scheduling, and finally showcased the potential of meta-learning models in these contexts.

The large variety of problems where MARL is applicable is a self-sustaining argument for its potential in building truly socially intelligent systems of agents, while the obtained results presented herewith show its incontrovertible significance.

# Bibliography


[ADBB17] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[AH81] Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[Axe81] Robert Axelrod. The emergence of cooperation among egoists. *The American Political Science Review*, pages 306–318, 1981.

[BBC+19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[BBD18] Juan C Burguillo, Burguillo, and Ditzinger. *Self-organizing coalitions for managing complexity*. Springer, 2018.

[BCP+16] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[BKM+19] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.

[BO98] Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

[BPB11] Ana LC Bazzan, Ana Peleteiro, and Juan C Burguillo. Learning to cooperate in the Iterated Prisoner's Dilemma by means of social attachments. *Journal of the Brazilian computer society*, 17(3):163–174, 2011.

[Byu14] Chong Hyun Christie Byun. The Prisoner's Dilemma and Economics 101: Do Active Learning Exercises Correlate with Student Performance? *Journal of the Scholarship of Teaching and Learning*, 14(5):79–91, 2014.

[Cow12] CC Cowden. Game theory, evolutionary stable strategies and the evolution of biological interactions. *Nature Education Knowledge*, 3(6), 2012.

[Daw16] Richard Dawkins. *The extended selfish gene*. Oxford University Press, 2016.

[DCH+16] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016.

[DSC+16] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[FFA+17] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. corr abs/1705.08926 (2017). *arXiv preprint arXiv:1705.08926*, 2017.

[FJW24] Shaheen Fatima, Nicholas R Jennings, and Michael Wooldridge. Learning to resolve social dilemmas: a survey. *Journal of Artificial Intelligence Research*, 79:895–969, 2024.

[FS03] Tobias Frick and Stefan Schuster. An example of the prisoner's dilemma in biochemistry. *Naturwissenschaften*, 90(7):327–331, 2003.

[Har09] Garrett Hardin. The tragedy of the commons. *Journal of Natural Resources Policy Research*, 1(3):243–253, 2009.

[HM23] Dom Huh and Prasant Mohapatra. Multi-agent reinforcement learning: A comprehensive survey. *arXiv preprint arXiv:2312.10256*, 2023.

[HPHKPI20] Manuel Herrera, Marco Pérez-Hernández, Ajith Kumar Parlikad, and Joaquín Izquierdo. Multi-agent systems and complex networks: Review and applications in systems engineering. *Processes*, 8(3):312, 2020.

[HS16] Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

[HY20] Ammar Haydari and Yasin Yılmaz. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):11–32, 2020.

[JSB02] Dominic DP Johnson, Pavel Stopka, and Josh Bell. Individual variation evades the Prisoner's Dilemma. *BMC Evolutionary Biology*, 2(1):1–8, 2002.

[KMP13] Soummya Kar, José MF Moura, and H Vincent Poor. QD-Learning: A Collaborative Distributed Strategy for Multi-Agent Reinforcement Learning Through Consensus. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

[Li19] Yuxi Li. Reinforcement learning applications. *arXiv preprint arXiv:1908.06973*, 2019.

[Lit94] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[LLL+19] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[LR00] Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

[LSB+23] Marc Lanctot, John Schultz, Neil Burch, Max Olan Smith, Daniel Hennes, Thomas Anthony, and Julien Perolat. Population-based Evaluation in Repeated RPS as a Benchmark for Multiagent Reinforcement Learning. *arXiv preprint arXiv:2303.03196*, 2023.

[LWT+17] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[LZG+17] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[Mal22] Imre Gergely Mali. Finding cooperation in the n-player iterated prisoner's dilemma with deep reinforcement learning over dynamic complex networks. In *26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES) - Procedia Computer Science*, volume 207, pages 465–474, 2022.

[Mal24] Imre-Gergely Mali. Multi-Agent Deep Reinforcement Learning for Collaborative Task Scheduling. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024)*, pages 1076–1083, 2024.

[Mal25] Imre Gergely Mali. A sequence-modeling approach to cooperation in public goods games via multi-agent transformers. *Procedia Computer Science*, 270:3708–3717, 2025.

[Mal26] Imre Gergely Mali. Self-play meta-reinforcement learning in multi-agent games. *Acta Universitatis Sapientiae, Informatica*, 18(1), feb 2026.

[MAMK16] Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*, pages 50–56, 2016.

[MC23] Imre-Gergely Mali and Gabriela Czibula. Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game. In *Proceedings of the 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2023)*, pages 345–350, 2023.

[Mye91] R Myerson. Game Theory: Analysis of Conflict Harvard Univ. *Press, Cambridge*, 1991.

[N+50] John F Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[NI55] Hukukane Nikaidô and Kazuo Isoda. Note on non-cooperative convex games. 1955.

[NX24] Zepeng Ning and Lihua Xie. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 3(2):73–91, 2024.

[PBARA14] Ana Peleteiro, Juan C Burguillo, Josep Ll Arcos, and Juan A Rodriguez-Aguilar. Fostering cooperation through dynamic coalition formation and partner switching. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 9(1):1–31, 2014.

[pd16] The Axelrod project developers. Axelrod: v4.10.0, April 2016.

[Pic19] Elvis Picardo. The Prisoner's Dilemma in business and the economy. *Retrieved May*, 25:2019, 2019.

[RN16] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Sha53] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[SHS+17] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[Smi79] John Maynard Smith. Game theory and the evolution of behaviour. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):475–488, 1979.

[SP73] J Maynard Smith and George R Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.

[SRAA+11] Norman Salazar, Juan A Rodriguez-Aguilar, Josep Ll Arcos, Ana Peleteiro, and Juan C Burguillo-Rial. Emerging cooperation on complex networks. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 669–676, 2011.

[SRDW+19] 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. *arXiv preprint arXiv:1902.04043*, 2019.

[SSS+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[SWD+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[TBG+21] J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

[TDM+18] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[Wen19] Lilian Weng. Meta reinforcement learning. *lilianweng.github.io*, 2019.

[WKL+22] Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521, 2022.

[WYWH18] Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. *arXiv preprint arXiv:1806.00877*, 2018.

[YVV+22] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

[ZWLP19] Li Zhang, Wei Wang, Shijian Li, and Gang Pan. Monte carlo neural fictitious self-play: approach to approximate nash equilibrium of imperfect-information games. *arXiv preprint arXiv:1903.09569*, 2019.

[ZYB19] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[ZYB21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.