Artificial Intelligence Review (2022) 55:895–943
https://doi.org/10.1007/s10462-021-09996-w

Multi-agent deep reinforcement learning: a survey

Sven Gronauer[1] · **Klaus Diepold**[1]

Published online: 15 April 2021

Abstract

The advances in reinforcement learning have recorded sublime success in various domains. Although the multi-agent domain has been overshadowed by its single-agent counterpart during this progress, multi-agent reinforcement learning gains rapid traction, and the latest accomplishments address problems with real-world complexity. This article provides an overview of the current developments in the field of multi-agent deep reinforcement learning. We focus primarily on literature from recent years that combines deep reinforcement learning methods with a multi-agent scenario. To survey the works that constitute the contemporary landscape, the main contents are divided into three parts. First, we analyze the structure of training schemes that are applied to train multiple agents. Second, we consider the emergent patterns of agent behavior in cooperative, competitive and mixed scenarios. Third, we systematically enumerate challenges that exclusively arise in the multi-agent domain and review methods that are leveraged to cope with these challenges. To conclude this survey, we discuss advances, identify trends, and outline possible directions for future work in this research area.

Keywords Multi-agent systems · Multi-agent learning · Machine learning · Reinforcement learning · Deep learning · Survey

1 Introduction

A multi-agent system describes multiple distributed entities—so-called agents—which take decisions autonomously and interact within a shared environment (Weiss 1999). Each agent seeks to accomplish an assigned goal for which a broad set of skills might be required to build intelligent behavior. Depending on the task, an intricate interplay between agents can occur such that agents start to collaborate or act competitively to excel opponents. Specifying intelligent behavior a-priori through programming is a tough, if not impossible, task for complex systems. Therefore, agents require the ability to adapt and learn over time

✉ Sven Gronauer
sven.gronauer@tum.de

Klaus Diepold
kldi@tum.de

[1] Department of Electrical and Computer Engineering, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany

by themselves. The most common framework to address learning in an interactive environment is reinforcement learning (RL), which describes the change of behavior through a trial-and-error approach.

The field of reinforcement learning is currently thriving. Since the breakthrough of deep learning methods, works have been successful at mastering complex control tasks, e.g. in robotics (Levine et al. 2016; Lillicrap et al. 2016) and game playing (Mnih et al. 2015; Silver et al. 2016). The key to these results is based on learning techniques that employ neural networks as function approximators (Arulkumaran et al. 2017). Despite these achievements, the majority of works investigated single-agent settings only, although many real-world applications naturally comprise multiple decision-makers that interact at the same time. The areas of application encompass the coordination of distributed systems (Cao et al. 2013; Wang et al. 2016b) such as autonomous vehicles (Shalev-Shwartz et al. 2016) and multi-robot control (Matignon et al. 2012a), the networking of communication packages (Luong et al. 2019), or the trading on financial markets (Lux and Marchesi 1999). In these systems, each agent discovers a strategy alongside other entities in a common environment and adapts its policy in response to the behavioral changes of others. Carried by the advances of single-agent deep RL, the multi-agent reinforcement learning (MARL) community has been surged with new interest and a plethora of literature has emerged lately (Hernandez-Leal et al. 2019; Nguyen et al. 2020). The use of deep learning methods enabled the community to exceed the historically investigated tabular problems to challenging problems with real-world complexity (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019; Vinyals et al. 2019).

In this paper, we provide an extensive review of the recent advances in the area of multi-agent deep reinforcement learning (MADRL). Although multi-agent systems enjoy a rich history (Busoniu et al. 2008; Shoham et al. 2003; Stone and Veloso 2000; Tuyls and Weiss 2012), this survey aims to shed light on the contemporary landscape of the literature in MADRL.

1.1 Related work

The intersection of multi-agent systems and reinforcement learning holds a long record of active research. As one of the first surveys in the field, Stone and Veloso (2000) analyzed multi-agent systems from a machine learning perspective and classified the reviewed literature according to heterogeneous and homogeneous agent structures as well as communication skills. The authors discussed issues associated with each classification. Shoham et al. (2003) criticized the ill-posed problem statement of MARL which is in the authors' opinion unclear and called for more grounded research. They proposed a coherent research agenda which includes four directions for future research. Yang and Gu (2004) reviewed algorithms and pointed out that the main difficulty lies in the generalization to continuous action and state spaces and in the scaling to many agents. Similarly, Busoniu et al. (2008) presented selected algorithms and discussed benefits as well as challenges of MARL. Benefits include computational speed-ups and the possibility of experience sharing between agents. In contrast, drawbacks are the specification of meaningful goals, the non-stationarity of the environment, and the need for coherent coordination in cooperative games. In addition to that, they posed challenges such as the exponential increase of computational complexity with the number of agents and the alter-exploration problem where agents must gauge between the acquisition of new knowledge and the exploitation of current knowledge. More specifically, Matignon et al. (2012b) identified challenges for the coordination of independent learners that arise in fully cooperative Markov Games such as non-stationarity, stochasticity, and shadowed equilibria. Further, they analyzed conditions under which algorithms can address such coordination issues. Another work by Tuyls and Weiss (2012) accounted for the historical developments of MARL and evoked non-technical challenges. They criticized that the intersection of RL techniques and game theory dominates multi-agent learning, which may render the scope of the field too narrow and investigations are limited to simplistic problems such as grid worlds. They claimed that the scalability to high numbers of agents and large and continuous spaces are the holy grail of this research domain.

Since the advent of deep learning methods and the breakthrough of deep RL, the field of MARL has attained new interest and a plethora of literature has emerged during the last years. Nguyen et al. (2020) presented five technical challenges including nonstationarity, partial observability, continuous spaces, training schemes, and transfer learning. They discussed possible solution approaches alongside their practical applications. Hernandez-Leal et al. (2019) concentrated on four categories including the analysis of emergent behaviors, learning communication, learning cooperation, and agent modeling. Further survey literature focuses on one particular sub-field of MADRL. Oroojlooyjadid and Hajinezhad (2019) reviewed recent works in the cooperative setting while Da Silva and Costa (2019) and Da Silva et al. (2019) focused on knowledge reuse. Lazaridou and Baroni (2020) reviewed the emergence of language and connected two perspectives, which comprise the conditions under which language evolves in communities and the ability to solve problems through dynamic communication. Based on theoretical analysis, Zhang et al. (2019) focused on MARL algorithms and presented challenges from a mathematical perspective.

1.2 Contribution and survey structure

The contribution of this paper is to present a comprehensive survey of the recent research directions pursued in the field of MADRL. We depict a holistic overview of current challenges that arise exclusively in the multi-agent domain of deep RL and discuss state-of-the-art solutions that were proposed to address these challenges. In contrast to the surveys of Hernandez-Leal et al. (2019) and Nguyen et al. (2020), which focus on a subset of topics, we aim to provide a widened and more comprehensive overview of the current investigations conducted in the field of MADRL while recapitulating what has already been accomplished. We identify contemporary challenges and discuss literature that addresses such. We see our work complementary to the theoretical survey of Zhang et al. (2019).

We dedicate this paper to an audience who wants an excursion to the realm of MADRL. Readers shall gain insights about the historical roots of this still young field and its current developments, but also understand the open problems to be faced by future research. The contents of this paper are organized as follows. We begin with a formal introduction to both single-agent and multi-agent RL and reveal pathologies that are present in MARL in Sect. 2. We then continue with the main contents, which are categorized according to the three-fold taxonomy as illustrated in Fig. 1.

We analyze training architectures in Sect. 3, where we categorize approaches according to a centralized or distributed training paradigm and additionally differentiate into execution schemes. Thereafter, we review literature that investigates emergent patterns of agent behavior in Sect. 4. We classify works in terms of the reward structure (Sect. 4.1), the language between multiple agents (Sect. 4.2), and the social context (Sect. 4.3). In Sect. 5, we enumerate current challenges of the multi-agent domain,

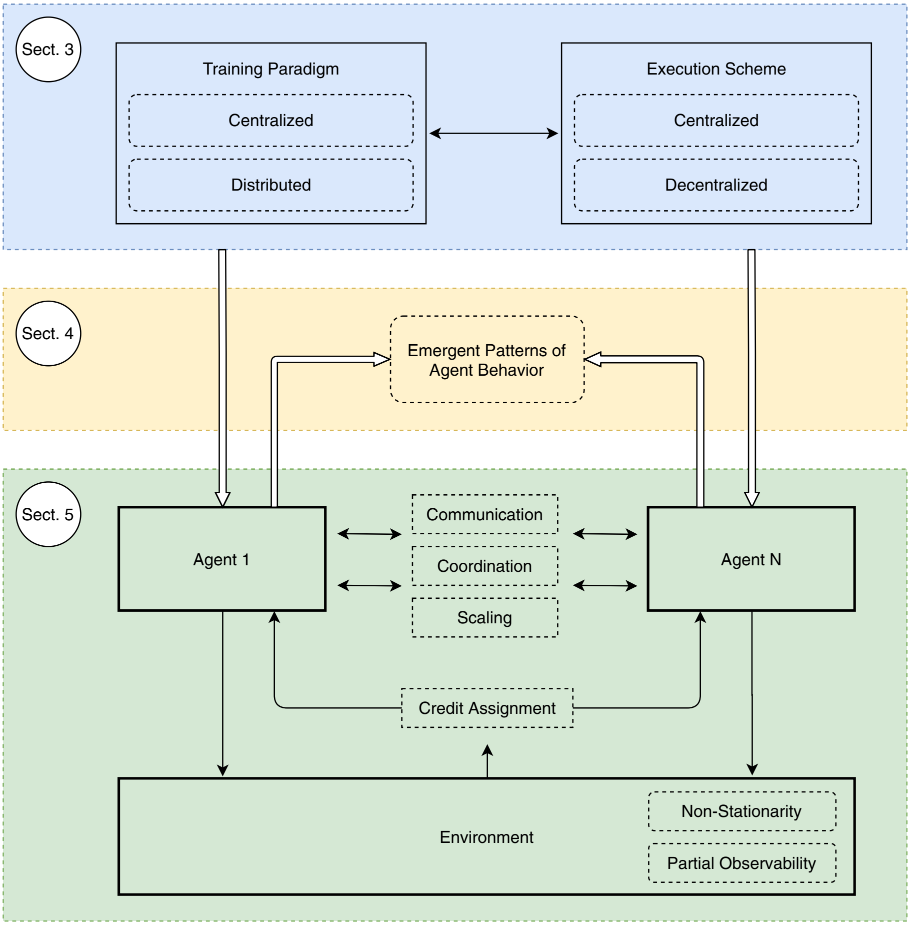

Fig. 1 Schematic structure of the main contents in this survey. In Sect. 3, we review schemes that are applied to train agent behavior in the multi-agent setting. The training of agents can be divided into two paradigms which are namely distributed (Sect. 3.1) and centralized (Sect. 3.2). In Sect. 4, we consider the emergent patterns of agent behavior with respect to the reward structure (Sect. 4.1), the language (Sect. 4.2) and the social context (Sect. 4.3). In Sect. 5, we enumerate current challenges of MADRL which include the non-stationarity of the environment due to co-adapting agents (Sect. 5.1), the learning of communication (Sect. 5.2), the need for a coherent coordination of actions (Sect. 5.3), the credit assignment problem (Sect. 5.4), the ability to scale to an arbitrary number of decision-makers (Sect. 5.5), and non-Markovian environments due to partial observations (Sect. 5.6)

which include the non-stationarity of the environment due to simultaneously adapting learners (Sect. 5.1), the learning of meaningful communication protocols in cooperative tasks (Sect. 5.2), the need for coherent coordination of agent actions (Sect. 5.3), the credit assignment problem (Sect. 5.4), the ability to scale to an arbitrary number of decision-makers (Sect. 5.5), and non-Markovian environments due to partial observations (Sect. 5.6). We discuss the matter of MADRL, pose trends that we identified in recent literature, and outline possible future work in Sect. 6. Finally, this survey concludes in Sect. 7.

2 Background

In this section, we provide a formal introduction into the concepts of RL. We start with the Markov decision process as a framework for single-agent learning in Sect. 2.1. We continue with the multi-agent case and introduce the Markov Game in Sect. 2.2. Finally, we pose pathologies that arise in the multi-agent domain such as the non-stationarity of the environment from the perspective of a single learner, relative over-generalization, and the credit assignment problem in Sect. 2.3. We provide the formal concepts behind these MARL pathologies in order to drive our discussion about the state-of-the-art approaches in Sect. 5. The scope of this background section is deliberately focusing on classical MARL works to reveal the roots of the domain and to give the reader insights into the early works on which modern MADRL approaches rest.

2.1 Single-agent reinforcement learning

The traditional reinforcement learning problem (Sutton and Barto 1998) is concerned with learning a control policy that optimizes a numerical performance by making decisions in stages. The decision-maker called agent interacts with an environment of unknown dynamics in a trial-and-error fashion and occasionally receives feedback upon which the agent wants to improve. The standard formulation for such sequential decision-making is the Markov decision process, which is defined as follows (Bellman 1957; Bertsekas 2012, 2017; Kaelbling et al. 1996).

Definition 1 *Markov decision process (MDP)* A Markov decision process is formalized by the tuple $(\mathscr{X}, \mathscr{U}, \mathscr{P}, R, \gamma)$ where $\mathscr{X}$ and $\mathscr{U}$ are the state and action space, respectively, $\mathscr{P} : \mathscr{X} \times \mathscr{U} \rightarrow P(\mathscr{X})$ is the transition function describing the probability of a state transition, $R : \mathscr{X} \times \mathscr{U} \times \mathscr{X} \rightarrow \mathbb{R}$ is the reward function providing an immediate feedback to the agent, and $\gamma \in [0, 1)$ describes the discount factor.

The agent's goal is to act in such a way as to maximize the expected performance on a long-term perspective with regard to an unknown transition function $\mathscr{P}$. Therefore, the agent learns a behavior policy $\pi : \mathscr{X} \rightarrow P(\mathscr{U})$ that optimizes the expected performance J throughout learning. The performance is defined as the expected value of discounted rewards

$$J = \mathbb{E}_{x_0 \sim \rho_0,\, x_{t+1} \sim \mathscr{P},\, u_t \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(x_t, u_t, x_{t+1}) \right] \tag{1}$$

over the initial state distribution ρ_0 while selected actions are governed by the policy π. Here, we regard the infinite-horizon problem where the interaction between agent and environment does not terminate after a countable number of steps. Note that the learning objective can also be formalized for finite-horizon problems (Bertsekas 2012, 2017). As an alternative to the policy performance, which describes the expected performance as a function of the policy, one can define the utility of being in a particular state in terms of a *value function*. The state-value function $V_\pi : \mathscr{X} \rightarrow \mathbb{R}$ describes the utility under policy π when starting from state x, i.e.

$$V_\pi(x) = \mathbb{E}_{x_{t+1}\sim\mathscr{P},\, u_t\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(x_t, u_t, x_{t+1}) \mid x_0 = x\right]. \tag{2}$$

In a similar manner, the action-value function $Q_\pi : \mathscr{X} \times \mathscr{U} \rightarrow \mathbb{R}$ describes the utility of being in state x, performing action u, and following the policy π thereafter, that is

$$Q_\pi(x, u) = \mathbb{E}_{x_{t+1}\sim\mathscr{P},\, u_{t>0}\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R(x_t, u_t, x_{t+1}) \mid x_0 = x, u_0 = u\Big]. \tag{3}$$

In the context of deep reinforcement learning, either the policy, a value function or both are represented by neural networks.

2.2 Multi-agent reinforcement learning

When the sequential decision-making is extended to multiple agents, Markov Games[1] are commonly applied as framework. The Markov Game was originally introduced by Littman (1994) to generalize MDPs to multiple agents that simultaneously interact within a shared environment and possibly with each other. The definition is formalized in a discrete-time setting and is denoted as follows (Littman 1994).

Definition 2 *Markov Games (MG)* The Markov Game is an extension to the MDP and is formalized by the tuple $\left(\mathscr{N}, \mathscr{X}, \{\mathscr{U}^i\}, \mathscr{P}, \{R^i\}, \gamma\right)$, where $\mathscr{N} = \{1, \ldots, N\}$ denotes the set of $N > 1$ interacting agents and $\mathscr{X}$ is the set of states observed by all agents. The joint action space is denoted by $\mathscr{U} = \mathscr{U}^1 \times \cdots \times \mathscr{U}^N$ which is the collection of individual action spaces from agents $i \in \mathscr{N}$. The transition probability function $\mathscr{P} : \mathscr{X} \times \mathscr{U} \rightarrow P(\mathscr{X})$ describes the chance of a state transition. Each agent owns an associated reward function $R^i : \mathscr{X} \times \mathscr{U} \times \mathscr{X} \rightarrow \mathbb{R}$ that provides an immediate feedback signal. Finally, $\gamma \in [0, 1)$ describes the discount factor.

At stage t, each agent $i \in \mathscr{N}$ selects and executes an action depending on the individual policy $\pi^i : \mathscr{X} \rightarrow P(\mathscr{U}^i)$. The system evolves from state x_t under the joint action u_t with respect to the transition probability function $\mathscr{P}$ to the next state x_{t+1} while each agent receives R^i as immediate feedback to the state transition. Akin to the single-agent problem, the aim of each agent is to change its policy in such a way as to optimize the received rewards on a long-term perspective.

A special case of the MG is the stateless setting $\mathscr{X} = \emptyset$ called strategic-form game[2]. Strategic-form games describe one-shot interactions where all agents simultaneously execute an action and receive a reward based on the joint action after which the game ends. Significant progress within the MARL community has been accomplished by studying this simplified stateless setting, which is still under active research to cope with several pathologies as discussed later in this section. These games are also known

[1] Markov games are also known as Stochastic Games (Shapley 1953), but we continue to use the term Markov Game to draw a clear distinction between deterministic Markov Games and stochastic Markov Games.

[2] The strategic-form game is also known as matrix game or normal-form game. The most commonly studied strategic-form game is the one with $N = 2$ players, the so-called bi-matrix game.

as matrix games because the reward function is represented by an $N \times N$ matrix. The formalism which extends to multi-step sequential stages is called extensive-form game.

In contrast to the single-agent case, the value function $V^i : \mathscr{X} \rightarrow \mathbb{R}$ does not only depend on the individual policy of agent i but also on the policies of other agents, i.e. the value function for agent i is the expected sum

$$V^i_{\pi^i, \boldsymbol{\pi}^{-i}}(x) = \mathbb{E}_{x_{t+1} \sim \mathscr{P}, u_t \sim \boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t R^i(x_t, u_t, x_{t+1}) \mid x_0 = x \right] \tag{4}$$

when the agents behave according to the joint policy $\boldsymbol{\pi}$. We denote the joint policy $\boldsymbol{\pi} : \mathscr{X} \rightarrow P(\mathscr{U})$ as the collection of all individual policies, i.e. $\boldsymbol{\pi} = \{\pi^1, \ldots, \pi^N\}$. Further, we make use of the convention that $-i$ denotes all agents except i, meaning for policies that $\boldsymbol{\pi}^{-i} = \{\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^N\}$.

The optimal policy is determined by the individual policy and the other agents' strategies. However, when other agents' policies are fixed, the agent i can maximize its own utility by finding the best response π^i_* with respect to the other agents' strategies.

Definition 3 *Best response* The agent's i best response $\pi^i_* \in \Pi^i$ to the joint policy $\boldsymbol{\pi}^{-i}$ of other agents is

$$V^i_{\pi^i_*, \boldsymbol{\pi}^{-i}}(x) \geq V^i_{\pi^i, \boldsymbol{\pi}^{-i}}(x)$$

for all states $x \in \mathscr{X}$ and policies $\pi^i \in \Pi^i$.

In general, when all agents learn simultaneously, the found best response may not be unique (Shoham and Leyton-Brown 2008). The concept of best response can be leveraged to describe the most influential solution concept from game theory: the *Nash equilibrium*.

Definition 4 *Nash equilibrium* A solution where each agent's policy π^*_i is the best response to the other agents' policy $\boldsymbol{\pi}^{-i}_*$ such that the following inequality

$$V^i_{\pi^i_*, \boldsymbol{\pi}^{-i}_*}(x) \ \geq \ V^i_{\pi^i, \boldsymbol{\pi}^{-i}_*}(x)$$

holds true for all states $x \in \mathscr{X}$ and all policies $\pi^i \in \Pi^i \ \forall i$ is called Nash equilibrium.

Intuitively spoken, a Nash equilibrium is a solution where one agent cannot improve when the policies of other agents are fixed, that is no agent can improve by unilaterally deviating from π^*. However, a Nash equilibrium may not be unique. Thus, the concept of Pareto-optimality might be useful (Matignon et al. 2012b).

Definition 5 *Pareto-optimality* A joint policy $\boldsymbol{\pi}$ Pareto-dominates a second joint policy $\hat{\boldsymbol{\pi}}$ if and only if

$$V^i_{\boldsymbol{\pi}}(x) \ \geq \ V^i_{\hat{\boldsymbol{\pi}}}(x) \quad \forall i, \forall x \in \mathscr{X} \quad \text{and} \quad V^j_{\boldsymbol{\pi}}(x) \ > \ V^j_{\hat{\boldsymbol{\pi}}}(x) \quad \exists j, \exists\, x \in \mathscr{X}.$$

A Nash equilibrium is regarded to be Pareto-optimal if no other has greater value and, thus, is not Pareto-dominated.

Classical MARL literature can be categorized according to different features, such as the type of task and the information available to agents. In the remainder of this section, we introduce MARL concepts based on the taxonomy proposed in Busoniu et al. (2008). For one, the primary factor that influences the learned agent behavior is the type of task. Whether agents compete or cooperate is promoted by the designed reward structure.

(1) *Fully cooperative setting* All agents receive the same reward $R = R^i = \cdots = R^N$ for state transitions. In such an equally-shared reward setting, agents are motivated to collaborate and try to avoid the failure of an individual to maximize the performance of the team. More generally, we talk about *cooperative settings* when agents are encouraged to collaborate but do not own an equally-shared reward.

(2) *Fully competitive setting* Such problem is described as a *zero-sum* Markov Game where the sum of rewards equals zero for any state transition, i.e. $R = \sum_{i=1}^{N} R^i(x, u, x') = 0$. Agents are prudent to maximize their own individual reward while minimizing the reward of the others. In a loose sense, we refer to competitive games when agents are encouraged to excel against opponents, but the sum of rewards does not equal zero.

(3) *Mixed setting* Also known as *general-sum* game, the mixed setting is neither fully cooperative nor fully competitive and, thus, does not incorporate restrictions on agent goals.

Beside the reward structure, other taxonomy may be used to differentiate between the information available to the agents. Claus and Boutilier (1998) distinguished between two types of learning, namely independent learners and joint-action learners. The former ignores the existence of other agents and cannot observe the rewards and selected actions of others as considered in Bowling and Veloso (2002) and Lauer and Riedmiller (2000). Joint-action learners, however, observe the taken actions of all other actions a-posteriori as shown in Hu and Wellman (2003) and Littman (2001).

2.3 Formal introduction to multi-agent challenges

In the single-agent formalism, the agent is the only decision-instance that influences the state of the environment. State transitions can be clearly attributed to the agent, whereas everything outside the agent's field of impact is regarded as part of the underlying system dynamics. Even though the environment may be stochastic, the learning problem remains stationary.

On the contrary, one of the fundamental problems in the multi-agent domain is that agents update their policies during the learning process simultaneously, such that the environment appears non-stationary from the perspective of a single agent. Hence, the Markov assumption of an MDP no longer holds, and agents face—without further treatment—a moving target problem (Busoniu et al. 2008; Yang and Gu 2004).

Definition 6 *Non-stationarity* A single agent faces a moving target problem when the transition probability function changes

$$\mathscr{P}(x' \mid x, u, \pi^1, \ldots, \pi^N) \neq \mathscr{P}(x' \mid x, u, \bar{\pi}^1, \ldots, \bar{\pi}^N),$$

due to the co-adaption $\pi^i \neq \bar{\pi}^i \; \exists \; i \in \mathscr{N}$ of agents.

Above, we have introduced the Nash equilibrium as a solution concept where each agent's policy is the best response to the others. However, it has been shown that agents can converge, despite a high degree of randomness in action selection, to sub-optimal solutions or can get stuck between different solutions (Wiegand 2004). Fulda and Ventura (2007) investigated such convergence to solutions and described a Pareto-selection problem called *shadowed equilibrium*.

Definition 7 *Shadowed equilibrium* A joint policy $\bar{\boldsymbol{\pi}}$ is shadowed by another joint policy $\hat{\boldsymbol{\pi}}$ in a state x if and only if

$$V_{\pi^i,\bar{\boldsymbol{\pi}}^{-i}}(x) < \min_{j,\pi_j} V_{\pi^j,\hat{\boldsymbol{\pi}}^{-j}}(x) \quad \exists\, i, \pi_i. \tag{5}$$

An equilibrium is shadowed by another when at least one agent exists who, when unilaterally deviating from $\bar{\boldsymbol{\pi}}$, will see no better improvement than for deviating from $\hat{\boldsymbol{\pi}}$ (Matignon et al. 2012b). As a form of shadowed equilibrium, the pathology of *relative overgeneralization* describes that a sub-optimal Nash equilibrium in the joint action space is preferred over an optimal solution. This phenomenon arises since each agent's policy performs relatively well when paired with arbitrary actions from other agents (Panait et al. 2006; Wei and Luke 2016; Wiegand 2004).

In a Markov Game, we assumed that each agent observes a state x, which encodes all necessary information about the world. However for complex systems, complete information might not be perceivable. In such partially observable settings, the agents do not observe the whole state space but merely a subset $\mathcal{O}^i \subset \mathcal{X}$. Hence, the agents are confronted to deal with sequential decision-making under uncertainty. The partially observable Markov Game (Hansen et al. 2004) is the generalization of both MG and MDP.

Definition 8 *Partially observable Markov Games (POMG)* The POMG is mathematically denoted by the tuple $\left(\mathcal{N}, \mathcal{X}, \{\mathcal{U}^i\}, \{\mathcal{O}^i\}, \mathcal{P}, \{R^i\}, \gamma\right)$, where $\mathcal{N} = \{1, \dots, N\}$ denotes the set of $N > 1$ interacting agents, $\mathcal{X}$ is the set of global but unobserved system states, and $\mathcal{U}$ is the set of individual action spaces $\mathcal{U}_i$. The observation space $\mathcal{O}$ denotes the collection of individual observation spaces $\mathcal{O}^i$. The transition probability function is denoted by $\mathcal{P}$, the reward function associated with agent i by R^i, and the discount factor is γ.

When agents face a cooperative task with a shared reward function, the POMG is then known as *decentralized Partially Observable Markov decision process* (dec-POMDP) (Bernstein et al. 2002; Oliehoek and Amato 2016). In partially observable domains, the inference of good policies is extended in complexity since the history of interactions becomes meaningful. Hence, the agents usually incorporate history-dependent policies $\pi^i_t : \{\mathcal{O}^i\}_{t>0} \rightarrow P(\mathcal{U}^i)$, which map from a history of observations to a distribution over actions.

Definition 9 *Credit assignment problem* In the fully-cooperative setting with joint reward signals, an individual agent cannot conclude the impact of its own action towards the team's success and, thus, faces a credit assignment problem.

In cooperative games, agents are encouraged to maximize a common goal through a joint reward signal. However, agents cannot ascertain their contribution to the eventual reward when they do not experience the taken joint action or deal with partial observations. Associating rewards to agents is known as the *credit assignment problem* (Chang et al. 2004; Weiß 1995; Wolpert and Tumer 1999).

Some of the above-introduced pathologies occur in all cooperative, competitive, and mixed tasks, whereas some pathologies like relative over-generalization, credit assignment, and miss-coordination are predominant issues in cooperative settings. To cope with these pathologies, still commonly studied settings are tabular worlds such as variations of the climbing game where solutions are not yet found, e.g. when the environment exhibits reward stochasticity (Claus and Boutilier 1998). Thus, simple worlds remain a fertile ground for further research, especially for problems like shadowed equilibria, non-stationarity or alter-exploration problems[3] and continue to matter for modern deep learning approaches.

3 Analysis of training schemes

The training of multiple agents has long been a computational challenge (Becker et al. 2004; Nair et al. 2003). Since the complexity in the state and action space grows exponentially with the number of agents, even modern deep learning approaches may reach their limits. In this section, we describe training schemes that are used in practice for learning agent policies in the multi-agent setting similar to the ones described in Bono et al. (2019). We denote training as the process during which agents acquire data to build up experience and optimize their behavior with respect to the received reward signals. In contrast, we refer test time[4] to the step after the training when the learned policy is evaluated but is no further refined. The training of agents can be broadly divided into two paradigms, namely centralized and distributed (Weiß 1995). If the training of agents is applied in a centralized manner, policies are updated based on the mutual exchange of information during the training. This additional information is then usually removed at test time. In contrast to the centralized scheme, the training can also be handled in a distributed fashion where each agent performs updates on its own and develops an individual policy without utilizing foreign information.

In addition to the training paradigm, agents may deviate in the way of how they select actions. We recognize two execution schemes. Centralized execution describes that agents are guided from a centralized unit, which computes the joint actions for all agents. On the contrary, agents determine actions according to their individual policy for decentralized execution. An overview of the training schemes is depicted in Fig. 2 while Table 1 lists the reviewed literature of this section.

3.1 Distributed training

In distributed training schemes, agents learn independently of other agents and do not rely on explicit information exchange.

[3] The alter-exploration dilemma, also known as the exploration-exploitation problem, describes the tradeoff an agent faces to decide whether to choose actions that extend experience or take decisions that are already optimal according to the current knowledge.

[4] Note that test and execution time are often used interchangeably in recent literature. For clarity, we use the term test for the post-training evaluation and the term execution for the action selection with respect to some policy.

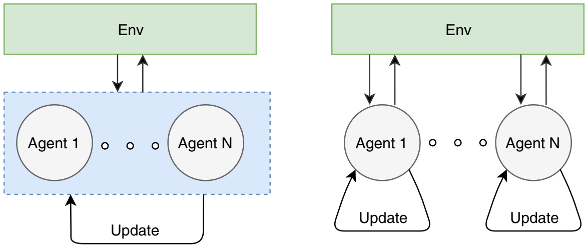

Fig. 2 Training schemes in the multi-agent setting. (Left) CTCE holds a joint policy for all agents. (Middle) Each agent updates its own individual policy in DTDE. (Right) CTDE enables agents to exchange additional information during training which is then discarded at test time

Definition 10 *Distributed training decentralized execution (DTDE)* Each agent i has an associated policy $\pi^i : \mathcal{O}^i \rightarrow P(\mathcal{U}^i)$ which maps local observations to a distribution over individual actions. No information is shared between agents such that each agent learns independently.

The fundamental drawback of the DTDE paradigm is that the environment appears non-stationary from a single agent's viewpoint because agents neither have access to the knowledge of others, nor do they perceive the joint action. The first approaches in this training scheme were studied in tabular worlds. The work by Tan (1993) investigated the question if independently learning agents can match with cooperating agents. The results showed that independent learners learn slower in tabular and deterministic worlds. Based on that, Claus and Boutilier (1998) examined both independent and joint-action learners in cooperative stochastic-form games and empirically showed that both types of learning can converge to an equilibrium in deterministic games. Subsequent works elaborated on the DTDE scheme in discretized worlds (Hu and Wellman 1998; Lauer and Riedmiller 2000).

More recent works report that distributed training schemes scale poorly with the number of agents due to the extra sample complexity, which is added to the learning problem. Gupta et al. (2017) showed that distributed methods have inferior performance compared to policies that are trained with a centralized training paradigm. Similarly, Foerster et al. (2018b) showed that the speed of independently learning actor-critic methods is slower than using centralized training. In further works, DTDE has been applied to cooperative navigation tasks (Chen et al. 2016; Strouse et al. 2018), to partially observable domains (Dobbe et al. 2017; Nguyen et al. 2017b; Srinivasan et al. 2018), and to social dilemmas (Leibo et al. 2017).

Due to limited information in the distributed setting, independent learners are confronted with several pathologies (Matignon et al. 2012b). Besides non-stationarity, environments may exhibit stochastic transitions or stochastic rewards, which further complicates learning. In addition to that, the search for an optimal policy influences the other agents' decision-making, which may lead to action shadowing and impacts the balance between exploration and knowledge exploitation.

A line of recent works expands independent learners with techniques to cope with the aforementioned MARL pathologies in cooperative domains. First, Omidshafiei et al. (2017) introduced a decentralized experience replay extension called Concurrent Experience Replay Trajectories (CERT) that enables independent learners to face a cooperative

Table 1 Overview of training schemes applied in recent MADRL works

Scheme	Approach	Literature
CTCE		Gupta et al. (2017) and Sunehag et al. (2018)
CTDE	Parameter sharing	Ahilan and Dayan (2019), Chu and Ye (2017), Gupta et al. (2017), Peng et al. (2017), Sukhbaatar et al. (2016) and Sunehag et al. (2018)
	Value-based	Castellini et al. (2019), Foerster et al. (2016), Jorge et al. (2016), Rashid et al. (2018), Son et al. (2019) and Sunehag et al. (2018)
	Centralized critic	Bono et al. (2019), Das et al. (2019), Foerster et al. (2018b), Lowe et al. (2017), Iqbal and Sha (2019), Wei et al. (2018) and Wu et al. (2018)
	Master-slave	Kong et al. (2017) and Kumar et al. (2017)
DTDE		Chen et al. (2016), Dobbe et al. (2017), Foerster et al. (2018b), Gupta et al. (2017), Jaderberg et al. (2019), Jaques et al. (2019), Leibo et al. (2017), Liu et al. (2019), Lyu and Amato (2020), Nguyen et al. (2017b), Omidshafiei et al. (2017), Palmer et al. (2018), Palmer et al. (2019), Srinivasan et al. (2018), Strouse et al. (2018) and Zheng et al. (2018a)

and partially observable setting by rendering samples more stable and efficient. Similarly, Palmer et al. (2018) extended the experience replay of Deep Q-Networks with leniency, which associates stored state-action pairs with decaying temperature values that govern the amount of applied leniency. They showed that this induces optimism in value function updates and can overcome relative over-generalization. Another work by Palmer et al. (2019) proposed negative update intervals double-DQN as an mechanism that identifies and removes generated data from the replay buffer that leads to mis-coordination. Alike, Lyu and Amato 2020 proposed decentralized quantile estimators which identify non-stationary transition samples based on the likelihood of returns. Another work that aims to improve upon independent learners can be found in Zheng et al. (2018a) who used two auxiliary mechanisms, including a lenient reward approximation and a prioritized replay strategy.

A different research direction can be seen in distributed population-based training schemes where agents are optimized through an online evolutionary process such that under-performing agents are substituted by mutated versions of better agents (Jaderberg et al. 2019; Liu et al. 2019).

3.2 Centralized training

The centralized training paradigm describes agent policies that are updated based on mutual information. While the sharing of mutual information between agents is enabled during the training, this additional information is then discarded at test time. The centralized training can be further differentiated into the centralized and decentralized execution scheme.

Definition 11 *Centralized training centralized execution (CTCE)* The CTCE scheme describes a centralized executor $\pi : \mathcal{O} \rightarrow P(\mathcal{U})$ modeling the joint policy that maps the collection of distributed observations to a set of distributions over individual actions.

Some applications assume an unconstrained and instantaneous information exchange between agents. In such a setting, a centralized executor can be leveraged to learn the joint policy for all agents. The CTCE paradigm allows the straightforward employment of single-agent training methods such as actor-critics (Mnih et al. 2016) or policy gradient algorithms (Schulman et al. 2017) to multi-agent problems. An obvious flaw is that state-action spaces grow exponentially by the number of agents. To address the so-called *curse of dimensionality*, the joint model can be factored into individual policies for each agent. Gupta et al. (2017) represented the centralized executor as a set of independent sub-policies such that agents' individual action distributions are captured rather than the joint action distribution of all agents, i.e. the joint action distribution $P(\mathcal{U}) = \prod_i P(\mathcal{U}^i)$ is factored into independent action distributions. Next to the policy, the value function can be factored so that the joint value is decomposed into a sum of local value functions, e.g. the joint action-value function can be expressed by $Q_\pi(o^1, \dots, o^N, u^1, \dots, u^n) = \sum_i Q^i_\pi(o^i, u^i)$ as shown in Russell and Zimdars (2003). A recent approach for the value function factorization is investigated in Sunehag et al. (2018). However, a phenomenon called *lazy agents* may occur in the CTCE setting when one agent learns a good policy but a second agent has less incentive to learn a good policy, as his actions may hinder the first agent, resulting in a lower reward (Sunehag et al. 2018).

Although CTCE regards the learning problem as a single-agent case, we include the paradigm in this paper because the training schemes presented in the subsequent sections occasionally use CTCE as performance baseline and conduct comparisons.

Definition 12 *Centralized training decentralized execution (CTDE)* Each agent i holds an individual policy $\pi^i : \mathcal{O}^i \rightarrow P(\mathcal{U}^i)$ which maps local observations to a distribution over individual actions. During training, agents are endowed with additional information, which is then discarded at test time.

The CTDE paradigm presents the state-of-the-art practice for learning with multiple agents (Kraemer and Banerjee 2016; Oliehoek et al. 2008). In classical MARL, such setting was utilized as *joint action learners* which has the advantage that perceiving joint actions a-posteriori discards the non-stationarity in the environment (Claus and Boutilier 1998). As of late, CTDE has been successful in MADRL approaches (Foerster et al. 2016; Jorge et al. 2016). Agents utilize shared computational facilities or other forms of communication to exchange information during training. By sharing mutual information, the training process can be eased and the learning speed can become superior when matched against independently trained agents (Foerster et al. 2018b). Moreover, agents can bypass non-stationarity when extra information about the selected actions is available to all agents during training such that the consequences of actions can be attributed to the respective agents. In what follows, we classify the CTDE literature according to the agent structure.

Homogeneous agents exhibit a common structure or the same set of skills, e.g. the same learning model or share common goals. Owning the same structure, agents can share parts of their learning model or experience with other agents. These approaches can scale well with the number of agents and may allow an efficient learning of behaviors. Gupta et al. (2017) showed that policies based on *parameter sharing* can be trained more efficiently and, thus, can outperform independently learned ones. Although agents own the same policy network, different agent behaviors can emerge because each agent perceives different observations at test time. It has been thoroughly demonstrated that parameter sharing can help to accelerate the learning progress (Ahilan and Dayan 2019; Chu and Ye 2017; Peng et al. 2017; Sukhbaatar et al. 2016; Sunehag et al. 2018). Next to parameter sharing, homogeneous agents can employ *value-based* methods where an approximation of the value function is learned based on mutual information. Agents profit from the joint actions and other agents' policies that are available during training and incorporate this extra information into centralized value functions (Foerster et al. 2016; Jorge et al. 2016). Such information is then discarded at test time. Many approaches consider the decomposition of a joint value function into combinations of individual value functions (Castellini et al. 2019; Rashid et al. 2018; Son et al. 2019; Sunehag et al. 2018). Through decomposition, each agent faces a simplified sub-problem of the original problem. Sunehag et al. (2018) showed that agents learning on local sub-problems scale better with the number of agents than CTCE or independent learners. We elaborate on value function-based factorization more detailed in Sect. 5.4 as an effective approach to tackle credit assignment problems.

Heterogeneous agents, on the contrary, differ in structure and skill. An instance for heterogeneous policies can be seen in the extension of an actor-critic approach with a *centralized critic*, which allows information sharing to amplify the performance of individual agent policies. These methods can be distinguished from each other based on the representation of the critic. Lowe et al. (2017) utilized one centralized critic for each agent that is augmented with additional information during training. The critics are provided with information about every agent's policy, whereas the actors perceive only local observations. As a result, the agents do not depend on explicit communication and can overcome the non-stationarity in the environment. Likewise, Bono et al. (2019) trained multiple agents with individual policies that share information with a centralized critic and demonstrated that such setup might improve results on standard benchmarks. Besides the utilization of one critic for each agent, Foerster et al. (2018b) applied one centralized critic for all agents to estimate a counterfactual baseline function that marginalizes out a single agent's action. The critic is conditioned on the history of all agents' observations or, if available, on the true global state. Typically, actor-critic methods underlie a variance in the critic estimation that is further exacerbated by the number of agents. Therefore, Wu et al. (2018) proposed an action-dependent baseline which includes information from other agents to reduce the variance in the critic estimation function. Further works that incorporate one centralized critic for distributed policies can be found in Das et al. (2019), Iqbal and Sha (2019) and Wei et al. (2018).

Another way to perform decentralized execution is by employing a *master-slave* architecture, which can resolve coordination conflicts between multiple agents. Kong et al. (2017) applied a centralized master executor which shares information with decentralized slaves. In each time step, the master receives local information from the slaves and shares its internal state in return. The slaves compute actions conditioned on their local observation and the master's internal state. Similar approaches that make use of different levels of abstraction are hierarchical methods (Kumar et al. 2017) that operate at different time scales or levels of abstraction. We elaborate on hierarchical methods in more detail in Sect. 5.3.

4 Emergent patterns of agent behavior

Agents adjust their policy to maximize the task success and react to the behavioral changes of other agents. The dynamic interaction between multiple decision-makers, which simultaneously affects the state of the environment, can cause the emergence of specific behavioral patterns. An obvious way to influence the development of agent behavior is through the designed reward structure. By promoting incentives for cooperation, agents can learn team strategies where they try to collaborate and optimize upon a mutual goal. Agents support other agents since the cumulative reward for cooperation is greater than acting selfishly. On the contrary, if the appeals for maximizing the individual performance are larger than being cooperative, agents can learn greedy strategies and maximize their individual reward. Such competitive attitudes can yield high-level strategies like manipulating adversaries to gain an advantage. However, the boundaries between competition and cooperation can be blurred in the multi-agent setting. For instance, if one agent competes with other agents, it is sometimes useful to cooperate temporarily in order to receive a higher reward in the long run.

In this section, we review the literature that is interested in developed agent behaviors. We differentiate occurring behaviors according to the reward structure (Sect. 4.1), the language between agents (Sect. 4.2), and the social context (Sect. 4.3). Table 2 summarizes the reviewed literature based on this classification. Note that we focus in this section not on works that introduce new methodologies but on literature that analyzes the emergent behavioral patterns.

Table 2 Overview of MADRL papers that investigate emergent patterns of agent behavior

Emergence	Setting	Literature
Reward structure	Cooperative	Diallo et al. (2017), Leibo et al. (2017) and Tampuu et al. (2017)
	Competitive	Bansal et al. (2018), Leibo et al. (2017), Liu et al. (2019) and Tampuu et al. (2017)
	Intrinsic rewards	Baker et al. (2020), Hughes et al. (2018), Jaderberg et al. (2019), Jaques et al. (2019), Jaques et al. (2018), Peysakhovich and Lerer (2018), Sukhbaatar et al. (2017), Wang et al. (2019) and Wang et al. (2020b)
Language	Referential games	Choi et al. (2018), Evtimova et al. (2018), Havrylov and Titov (2017), Jorge et al. (2016), Lazaridou et al. (2017), Lazaridou et al. (2018), Lee et al. (2017) and Mordatch and Abbeel (2018)
	Dialogues	Cao et al. (2018), Das et al. (2017) and Lewis et al. (2017)
Social context	Commons dilemmas	Foerster et al. (2018a), Jaques et al. (2018), Jaques et al. (2019), Leibo et al. (2017) and Lerer and Peysakhovich (2017)
	Public good dilemmas	Pérolat et al. (2017), Hughes et al. (2018) and Zhu and Kirley (2019)

4.1 Reward structure

The primary factor that influences the emergence of agent behavior is the reward structure. If the reward for mutual *cooperation* is larger than individual reward maximization, agents tend to learn policies that seek to collaboratively solve the task. In particular, Leibo et al. (2017) compared the magnitude of the team reward in relation to the individual agent reward. They showed that the higher the numerical team reward is compared to the individual reward, the greater is the willingness to collaborate with other agents. The work by Tampuu et al. (2017) demonstrated that punishing the whole team of agents for the failure of a single agent can also cause cooperation. Agents learn policies to avoid the malfunction of an individual, support other agents to prevent failure, and improve the performance of the whole team. Similarly, Diallo et al. (2017) used the Pong video game to investigate the coordination between agents and examined how developed behaviors change regarding the reward function. For a comprehensive review of learning in cooperative settings, one can consider the article by Panait and Luke (2005) for classical MARL and Oroojlooyjadid and Hajinezhad (2019) for recent MADRL.

In contrast to the cooperative scenario, one can value individual performance greater than the collaboration among agents. A *competitive* setting motivates agents to outperform their adversary counterparts. Tampuu et al. (2017) used the video game Pong and manipulated the rewarding structure to examine the emergence of agent behavior. They showed that the higher the reward for competition, the more likely an agent tries to outplay its opponents by using techniques such as wall bouncing or faster ball speed. Employing such high-level strategies to overwhelm the adversary maximizes the individual reward. Similarly, Bansal et al. (2018) investigated competitive scenarios, where agents competed in a 3D world with simulated physics to learn locomotion skills such as running, blocking, or tackling other agents with arms and legs. They argued that adversarial training could help to learn more complex agent behaviors than the environment can exhibit. Likewise, the works of Leibo et al. (2017) and Liu et al. (2019) investigated the emergence of behaviors due to the reward structure in competitive scenarios.

If the rewards appear in sparse frequency, agents can be equipped with *intrinsic reward functions* that provide denser feedback signals and, thus, can overcome the sparsity or even the absence of external rewards. One way to realize this is with intrinsic motivation, which is based on the concept of maximizing an internal reinforcement signal by actively discovering novel or surprising patterns (Chentanez et al. 2005; Oudeyer and Kaplan 2007; Schmidhuber 2010). Intrinsic motivation encourages agents to explore states that have been scarcely or never visited and to perform novel actions in those states. Most approaches of intrinsic motivation can be broadly divided into two categories (Pathak et al. 2017). First, agents are encouraged to explore unknown states where the novelty of states is measured by a model that captures the distribution of visited environment states (Bellemare et al. 2016). Second, agents can be motivated to reduce the uncertainty about the consequences of their own actions. The agent builds a model that learns the dynamics of the environment by lowering the prediction error of the follow-up states with respect to the taken actions. The uncertainty indicates the novelty of new experience since the model can only be accurate in states which it has already encountered or can generalize from previous knowledge (Houthooft et al. 2016; Pathak et al. 2017). For a recent survey on intrinsic motivation in RL, one can regard the paper by Aubret et al. (2019). The concept of intrinsic motivation was transferred to the multi-agent domain by Sequeira et al. (2011), who studied the motivational impact on multiple agents. Investigations on the emergence of agent behavior based on intrinsic rewards have been abundantly conducted in Baker et al. (2020), Hughes et al. (2018), Jaderberg et al. (2019), Jaques et al. (2018), Jaques et al. (2019), Peysakhovich and Lerer (2018), Sukhbaatar et al. (2017), Wang et al. (2019) and Wang et al. (2020b).

4.2 Language

The development of language corpora and communication skills of autonomous agents attracts great attention within the community. For one, the behavior that emerges during the deployment of abstract language as well as the learned composition of multiple words to form meaningful contexts is of interest (Kirby 2002). Deep learning methods have widened the scope of computational methodologies for investigating the development of language between dynamic agents (Lazaridou and Baroni 2020). For building rich behaviors and complex reasoning, communication based on high-dimensional data like visual perception is a widespread practice (Antol et al. 2015). In the following, we focus on works that investigate the emergence of language and analyze behavior. Papers that propose new methodologies for developing communication protocols are discussed in Sect. 5.2. We classify the learning of language according to the performed task and the type of interaction the agents pursue. In particular, we differentiate between referential games and dialogues.

The former, *referential games*, describe cooperative games where the speaking agent communicates an objective via messages to another listening agent. Lazaridou et al. (2017) showed that agents could learn communication protocols solely through interaction. For a meaningful information exchange, agents evolved semantic properties in their language. A key element of the study was to analyze if the agents' interactions are interpretable for humans, showing limited yet encouraging results. Likewise, Mordatch and Abbeel (2018) investigated the emergence of abstract language that arises through the interaction between agents in a physical environment. In their experiments, the agents should learn a discrete set of vocabulary by solving navigation tasks through communication. By involving more than three agents in the conversation and by penalizing an arbitrary size of vocabulary, agents agreed on a coherent set of vocabulary and discouraged ambiguous words. They also observed that agents learned a syntax structure in the communication protocol that is consistent in vocabulary usage. Another work by Li and Bowling (2019) found out that compositional languages are easier to communicate with other agents than languages with less structure. In addition, changing listening agents during the learning can promote the emergence of language grounded on a higher degree of structure. Many studies are concerned with the development of communication in referential games grounded on visual perception as it can be found in Choi et al. (2018), Evtimova et al. (2018), Havrylov and Titov (2017), Jorge et al. (2016), Lazaridou et al. (2018) and Lee et al. (2017). Further works consider the development of communication in social dilemmas (Jaques et al. 2018, 2019).

As the second category, we describe the emergence of behavioral patterns in communication while conducting *dialogues*. One type of dialogue are negotiations in which agents pursue to agree on decisions. In a study about negotiations with natural language, Lewis et al. (2017) showed that agents could master linguistic and reasoning problems. Two agents were both shown a collection of items and were instructed to negotiate about how to divide the objects among both agents. Each agent was expected to maximize the value of the bargained objects. Eventually, the agents learned to use high-level strategies such as deception to accomplish higher rewards over their opponents. Similar studies concerned with negotiations are covered in Cao et al. (2018) and He et al. (2018). Another type of dialogue are scenarios where the emergence of communication is investigated in a question-answering style as shown by Das et al. (2017). One agent received an image as input and was instructed to ask questions about the shown image while the second agent responded, both in natural language.

Many of the above-mentioned papers report that utilizing a communication channel can increase task performance in terms of the cumulative reward. However, numerical performance measurements provide evidence but do not give insights about the communication abilities learned by the agents. Therefore, Lowe et al. (2019) surveyed metrics which are applied to assess the quality of learned communication protocols and provided recommendations about the usage of such metrics. Based on that, Eccles et al. (2019) proposed to incorporate inductive bias into the learning objective of agents, which could promote the emergence of a meaningful communication. They showed that inductive bias could lead to improved results in terms of interpretability.

4.3 Social context

Next to the reward structure and language, the research community actively investigates the emerging agent behaviors in social contexts. Akin to humans, artificial agents can develop strategies that exploit patterns in complex problems and adapt behaviors in response to others (Baker et al. 2020; Jaderberg et al. 2019). We differentiate the following literature along different dimensions, such as the type of social dilemma and the examined psychological variables.

Social dilemmas have long been studied as conflict scenario in which agents gauge between individualistic and collective profits (Crandall and Goodrich 2011; De Cote et al. 2006). The tension between cooperation and defection is evaluated as an atomic decision according to the numerical values of a pay-off matrix. This pay-off matrix satisfies inequalities in the reward function such that agents must decide between cooperation, to benefit as a whole team, or defection, to maximize selfish performance. To temporally extend matrix games, sequential social dilemmas have been introduced to investigate long-term strategic decisions of agent policies rather than short-term actions (Leibo et al. 2017). The arising behaviors in these dilemmas can be classified along psychological variables known from human interaction (Lange et al. 2013) such as the gain of individual benefits (Lerer and Peysakhovich 2017), the fear of future consequences (Pérolat et al. 2017), the assessment of the impact on another agent's behavior (Jaques et al. 2018, 2019), the trust between agents (Pinyol and Sabater-Mir 2013; Ramchurn et al. 2004; Yu et al. 2013), and the impact of emotions on the decision-making (Moerland et al. 2018; Yu et al. 2013).

Kollock (1998) divided social dilemmas into commons dilemmas and public goods dilemmas. The former, *commons dilemmas* describe the trade-off between individualistic short-term benefits and long-term common interests on a task that is shared by all agents. Recent works on the commons dilemma can be found in Foerster et al. (2018a), Leibo et al. (2017) and Lerer and Peysakhovich (2017). In *public goods dilemmas*, agents face a scenario where common-pool resources are constrained and oblige a sustainable use of resources. The phenomenon called the tragedy of commons predicts that self-interested agents fail to find socially positive equilibria, which eventually results in the over-exploitation of the common resources (Hardin 1968). Investigations on the trial-and-error learning in common-pool resource scenarios with multiple decision-makers are covered in Hughes et al. (2018), Pérolat et al. (2017) and Zhu and Kirley (2019).

5 Current challenges

In this section, we depict several challenges that arise in the multi-agent RL domain and, thus, are currently under active research. We approach the problem of non-stationarity (Sect. 5.1) due to the presence of multiple learners in a shared environment and review literature regarding the development of communication skills (Sect. 5.2). We further investigate the challenge of learning coordination (Sect. 5.3). Then, we survey the difficulty of attributing rewards to specific agents as the credit assignment problem (Sect. 5.4) and examine scalability issues (Sect. 5.5), which increase with the number of agents. Finally, we consider environments where states are only partially observable (Sect. 5.6). While some challenges are omnipresent in the MARL domain, such as non-stationarity or scalability, others like the credit assignment problem or the learning of coordination and communication are prevailing in the cooperative setting.

We aim to provide a holistic overview of the contemporary challenges that constitute the landscape in reinforcement learning with multiple agents and survey treatments that were suggested in recent works. In particular, we focus on those challenges which are currently under active research and where progress has been accomplished recently. There are still open problems that have not been or partially addressed so far. Such problems are discussed in Sect. 6. Deliberately, we do not regard challenges that also persist in the single-agent domain, such as sparse rewards or the exploration-exploitation dilemma. We refer the interested reader for an overview of those topics to the articles of Arulkumaran et al. (2017) and Li (2018). Much of the surveyed literature cannot be assigned to one particular but rather to several of the proposed challenges. Hence, we associate the subsequent literature to the one challenge which we believe best addresses it (Table 3).

5.1 Non-stationarity

One major problem resides in the presence of multiple agents that interact within a shared environment and learn simultaneously. Due to the co-adaption, the environment dynamics appear non-stationary from the perspective of a single agent. Thus, agents face a moving target problem if they are not provided with additional knowledge about other agents. As a result, the Markov assumption is violated, and the learning constitutes an inherently difficult problem (Hernandez-Leal et al. 2017; Laurent et al. 2011). The naïve approach is to neglect the adaptive behavior of agents. One can either ignore the existence of other agents (Matignon et al. 2012b) or discount the adaptive behavior by assuming the others' behavior to be static or optimal (Lauer and Riedmiller 2000). By making such assumptions, the agents are considered as independent learners, and traditional single-agent reinforcement algorithms can be applied. First attempts have been studied in Claus and Boutilier (1998) and Tan (1993), which showed that independent learners could perform well in simple deterministic environments. However, in complex or stochastic environments, independent learners often result in poor performance (Lowe et al. 2017; Matignon et al. 2012b). Moreover, Lanctot et al. (2017) argued that independent learners could over-fit to other agents' policies during the training and, thus, may fail to generalize at test time.

In the following, we review literature, which addresses the non-stationarity in a multi-agent environment, and categorize the approaches into those with experience replay, centralized units, and meta-learning. A similar categorization proposed Papoudakis et al. (2019). We identify further approaches which cope with non-stationarity by establishing

Table 3 Overview of MADRL challenges and approaches proposed in recent literature

Challenge	Approach	Literature
Non-stationarity	Experience replay	Foerster et al. (2017), Palmer et al. (2018), Tang et al. (2018), and Zheng et al. (2018a)
	Centralized training	Bono et al. (2019), Foerster et al. (2016), Foerster et al. (2018b), Iqbal and Sha (2019), Jorge et al. (2016), Lowe et al. (2017), Rashid et al. (2018), and Wei et al. (2018)
	Meta-learning	Al-Shedivat et al. (2018), and Rabinowitz et al. (2018)
Communication	Broadcasting	Foerster et al. (2016), Peng et al. (2017), and Sukhbaatar et al. (2016)
	Targeted	Das et al. (2019), Hoshen (2017), Jain et al. (2019), Jiang and Lu (2018), and Singh et al. (2019)
	Networked	Chu et al. (2020), Chu et al. (2020), Qu et al. (2020), Zhang et al. (2018), and Zhang et al. (2019)
	Extensions	Celikyilmaz et al. (2018), Jaques et al. (2018), Jaques et al. (2019), Kim et al. (2019), Li et al. (2019b), Singh et al. (2019), and Wang et al. (2020c)
Coordination	Independent learners	Foerster et al. (2018b), Lyu and Amato (2020), Omidshafiei et al. (2017), Palmer et al. (2018), Palmer et al. (2019), Sunehag et al. (2018), and Zheng et al. (2018a)
	Constructing models	Barde et al. (2019), Everett and Roberts (2018), Foerster et al. (2018a), Foerster et al. (2019), Grover et al. (2018), He et al. (2016), Hong et al. (2017), Hoshen (2017), Jaques et al. (2019), Le et al. (2017), Letcher et al. (2019), Raileanu et al. (2018), Tacchetti et al. (2019), Yang et al. (2018a), and Zheng et al. (2018b)
	Hierarchical methods	Ahilan and Dayan (2019), Cai et al. (2013), Han et al. (2019), Jaderberg et al. (2019), Kumar et al. (2017), Lee et al. (2020), Ma and Wu (2020), Tang et al. (2018), and Vezhnevets et al. (2019)
Credit assignment	Decomposition	Castellini et al. (2019), Chen et al. (2018), Nguyen et al. (2017b), Rashid et al. (2018), Son et al. (2019), Sunehag et al. (2018), Wang et al. (2020a), Wang et al. (2020c), Yang et al. (2018b)
	Marginalization	Foerster et al. (2018b), Nguyen et al. (2018), and Wu et al. (2018)
	Inverse RL	Barrett et al. (2017), Le et al. (2017), Lin et al. (2018), Song et al. (2018), and Yu et al. (2019)
Scalability	Knowledge Reuse	Baker et al. (2020), Da Silva et al. (2017), Da Silva and Costa (2017), Gupta et al. (2017), Hernandez-Leal et al. (2019), Jiang and Lu (2018), Long et al. (2020), Luketina et al. (2019), Narvekar et al. (2016), Omidshafiei et al. (2019), Peng et al. (2017), Sukhbaatar et al. (2016), Sukhbaatar et al. (2017), Sunehag et al. (2018), and Svetlik et al. (2017)
	Complexity reduction	Chen et al. (2018), Lin et al. (2018), Nguyen et al. (2017a), Nguyen et al. (2017b), and Yang et al. (2018b)
	Robustness	Baker et al. (2020), Bansal et al. (2018), Berner et al. (2019), Gleave et al. (2020), Heinrich and Silver (2016), Lanctot et al. (2017), Li et al. (2019a), Liu et al. (2020), Lowe et al. (2017), Pinto et al. (2017), Raghu et al. (2018), Silver et al. (2016), Silver et al. (2018), Spooner and Savani (2020), and Sukhbaatar et al. (2017)

Table 3 (continued)

Challenge	Approach	Literature
Partial observability	Memory mechanism	Dibangoye and Buffet (2018), Foerster et al. (2018b), Foerster et al. (2019), Gupta et al. (2017), and Omidshafiei et al. (2017)

communication between agents (Sect. 5.2) or building models (Sect. 5.3). However, we discuss these topics separately in the respective sections.

Experience replay mechanism Recent successes with reinforcement learning methods such as deep Q-networks (Mnih et al. 2015) rest upon an experience replay mechanism. However, it is not straightforward to employ experience replays to the multi-agent setting because past experience becomes obsolete with the adaption of agent policies over time. To encounter this, Foerster et al. (2017) proposed two approaches. First, they decay outdated transition samples from the replay memory to stabilize targets and then use importance sampling to incorporate off-policy samples. Since the agents' policies are known during the training, off-policy updates can be corrected with importance-weighted policy likelihoods. Second, the state space of each agent is enhanced with estimates of the other agents' policies, so-called fingerprints[5], to prevent non-stationarity. The value functions can then be conditioned on a fingerprint, which clears the age of data sampled from the replay memory. Another extension for experience replays was proposed by Palmer et al. (2018) who applied leniency to every stored transition sample. Leniency associates each sample of the experience memory with a temperature value, which gradually decays by the number of state-action pair visits. Further utilization of the experience replay mechanism to cope with non-stationarity can be found in Tang et al. (2018) and Zheng et al. (2018a). Nevertheless, if the contemporary dynamics of the learners are neglected, algorithms can utilize short-term buffers as applied in Baker et al. (2020) and Leibo et al. (2017).

Centralized Training Scheme As already discussed in Sect. 3.2, the CTDE paradigm can be leveraged to share mutual information between learners to ease training. The availability of information during the training can loosen the non-stationarity of the environment since agents are augmented with information about others. One approach is to enhance *actor-critic* methods with centralized critics over which mutual information is shared between agents during the training (Bono et al. 2019; Iqbal and Sha 2019; Wei et al. 2018). Lowe et al. (2017) embedded each agent with one centralized critic that is augmented with all agents' observations and actions. Based on this additional information, agents face a stationary environment during the training while acting decentralized on local observations at test time. Next to the equipment of one critic per agent, all agents can share one global centralized critic. Foerster et al. (2018b) applied one centralized critic conditioned on the joint action and observations of all agents. The critic computes an agent's individual advantage through estimating the value of the joint action based on a counterfactual baseline, which marginalizes out single agents' influence. Another approach to the CTDE scheme can be seen in *value-based* methods. Rashid et al. (2018) learned a joint action-value function conditioned on the joint observation-action history. The joint action-value function is then divided into agent individual value functions based on monotonic non-linear composition. Foerster et al. (2016) used action-value functions that share information through a communication channel during the training but then discarded it at test time. Similarly, Jorge et al. (2016) employed communication during training to promote information exchange for optimizing action-value functions.

Meta-Learning Sometimes, it can be useful to learn how to adapt to the behavioral changes of others. This learning-to-learn approach is known as meta-learning (Finn and Levine 2018; Schmidhuber et al. 1996). Recent works in the single-agent domain have shown promising results (Duan et al. 2016; Wang et al. 2016a). Al-Shedivat et al. (2018)

[5] Fingerprints draw their inspiration from Tesauro (2004) who eluded non-stationarity by conditioning each agent's policy on estimates of other agents' policies.

transferred this approach to the multi-agent domain and developed a meta-learning based method to tackle the consecutive adaptation of agents in non-stationary environments. Regarding non-stationarity as a sequence of stationary tasks, agents learn to exploit dependencies between successive tasks and generalize over co-adapting agents at test time. They evaluated the resulting behaviors in a competitive multi-agent setting where agents fight in a simulated physics environment. Meta-learning can also be utilized to construct agent models (Rabinowitz et al. 2018). By learning how to model other agents and make inferences on them, agents learn to predict the other agent's future action sequences. They embedded this principle into how one agent learns to capture the behavioral patterns of other agents efficiently.

5.2 Learning communication

Agents capable of developing communication and language corpora pose one of the vital challenges in machine intelligence (Kirby 2002). Intelligent agents must not only decide on what to communicate but also when and with whom. It is indispensable that the developed language is grounded on a common consensus such that all agents understand the spoken language, including its semantics. The research efforts in learning to communicate have intensified because many pathologies can be overcome by incorporating communication skills into agents, including non-stationarity, coherent coordination among agents, and partial observability. For instance, when an agent knows the actions taken by others, the learning problem becomes stationary again from a single agent's perspective in a fully observable environment. Even partial observability can be loosened by messaging local observations to other participants through communication, which helps compensate for limited knowledge (Goldman and Zilberstein 2004).

The common framework to investigate communication is the dec-POMDP (Oliehoek and Amato 2016) which is a fully cooperative setting where agents perceive partial observations of the environment and try to improve upon an equally-shared reward. In such distributed systems, agents must not only learn how to cooperate but also how to communicate in order to optimize the mutual objective. Early MARL works investigated communication rooted in tabular worlds with limited observability (Kasai et al. 2008). Since the spring of deep learning methods, the research of learning communication has witnessed great attention because advanced computational methods provide new opportunities to study highly complex data.

In the following, we categorize the surveyed literature according to the message addressing. First, we describe the broadcasting scenario where sent messages are received by all agents. Second, we look into works that use targeted messages to decide on the recipients by using an attention mechanism. Third and last, we review communication in networked settings where agents communicate only with their local neighborhood instead of the whole population. Figure 3 shows a schematic illustration of this categorization. Another taxonomy may be based on the discrete or continuous nature of messages and the frequency of passed messages.

Broadcasting Messages are addressed to all participants of the communication channel. Foerster et al. (2016) studied how agents learn *discrete* communication protocols in dec-POMDPs in order to accomplish a fully-cooperative task. Being in a CTDE setting, the communication is not restricted during the training but bandwidth-limited at test time. To discover meaningful communication protocols, they proposed two methods. The first, reinforced inter-agent learning (RIAL), is based on deep recurrent Q-networks combined with

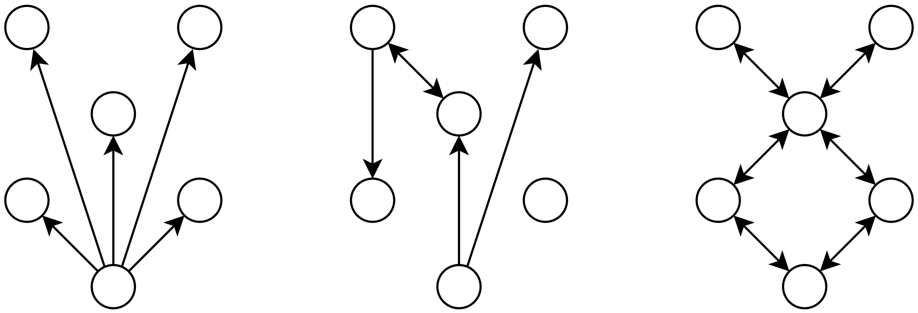

Fig. 3 Schematic illustration of communication types. Unilateral arrows represent unidirectional messages, while bilateral arrows symbolize bidirectional message passing. (Left) In broadcasting, messages are sent to all participants of the communication channel. For better visualization, the broadcasting of only one agent is illustrated but each agent can broadcast messages to all other agents. (Middle) Agents can target the communication through an attention mechanism that determines when, what and with whom to communicate. (Right) Networked communication describes the local connection to neighborhood agents

independent Q-learning where each agent learns an action-value function conditioned on the observation history as well as messages from other agents. Additionally, they applied parameter sharing so that all agents share and update common features from only one Q-network. The second method, differentiable inter-agent learning (DIAL), combines the centralized learning paradigm with deep Q-networks. Messages are delivered over discrete connections, which are based on a relaxation to become differentiable. In contrast, Sukhbaatar et al. (2016) proposed CommNet as an architecture that allows the learning of communication between agents purely based on *continuous* protocols. They showed that each agent learns the joint-action and a sparse communication protocol that encodes meaningful information. The authors emphasized that the decreased observability of vicious states encourages the importance of communication between agents. To foster scalable communication protocols that also facilitate heterogeneous agents, Peng et al. (2017) introduced the bidirectionally-coordinated network (BiCNet) where agents learn in a vectorized actor-critic framework to communicate. Through communication, they were able to coordinate heterogeneous agents in a combat game of StarCraft.

Targeted communication When agents are endowed with targeted communication protocols, they utilize an attention mechanism to determine when, what and with whom to communicate. Jiang and Lu (2018) introduced ATOC as an attentional communication model that enables agents to send messages dynamically and selectively so that communication takes place among a group of agents only when required. They argued that attention is essential for large-scale settings because agents learn to decide which information is most useful for decision-making. Selective communication is the reason why ATOC outperforms CommNet and BiCNet on the conducted navigation tasks. A similar conclusion was drawn by Hoshen (2017) who introduced the vertex attention interaction network (VAIN) as an extension to the CommNet. The baseline approach is extended with an attention mechanism that increases performance due to the focus on only relevant agents. The work by Das et al. (2019) introduced targeted multi-agent communication (TarMAC) that uses attention to decide with whom and what to communicate by actively addressing other agents for message passing. Jain et al. (2019) proposed TBONE for visual navigation in cooperative tasks. In contrast to former works, which are limited to the fully-cooperative setting, Singh et al. (2019) considered mixed settings where each agent owns an individual reward function. They proposed the individualized controlled continuous communication model (IC3Net), where agents learn when to exchange information using a gating mechanism that blocks incoming communication requests if necessary.

Networked communication Another form of communication is a networked communication protocol where agents can exchange information with their neighborhood (Nedic and Ozdaglar 2009; Zhang et al. 2018). Agents act decentralized based on local observations and received messages from network neighbors. Zhang et al. (2018) used an actor-critic framework where agents share their critic information with their network neighbors to promote global optimality. Chu et al. (2020) introduced the neural communication protocol (NeurComm) to enhance communication efficiency by reducing queue length and intersection delay. Further, they showed that a spatial discount factor could stabilize training when only the local vicinity is regarded to perform policy updates. For theoretical contributions, one may consider the works of Qu et al. (2020), Zhang et al. (2018) and Zhang et al. (2019) whereas the paper of Chu et al. (2020) provides an application perspective in the domain of traffic light control.

Extensions Further methods approach the improvement of coordination skills by applying intrinsic motivation (Jaques et al. 2018, 2019), by making the communication protocol more robust or scalable (Kim et al. 2019; Singh et al. 2019), and maximizing the utility of the communication through efficient encoding (Celikyilmaz et al. 2018; Li et al. 2019b; Wang et al. 2020c).

The above-reviewed papers focus on new methodologies about communication protocols. Besides that, a bulk of literature considers the analysis of emergent language and the occurrence of agent behavior, which we discuss in Sect. 4.2.

5.3 Coordination

Successful coordination in multi-agent systems requires agents to agree on a consensus (Wei Ren et al. 2005). In particular, accomplishing a joint goal in cooperative settings demands a coherent action selection such that the joint action optimizes the mutual task performance. Cooperation among agents is complicated when stochasticity is present in system transitions and rewards or when agents observe only partial information of the environment's state. Mis-coordination may arise in the form of action shadowing when exploratory behavior influences the other agents' search space during learning and, as a result, sub-optimal solutions are found.

Therefore, the agreement upon a mutual consensus necessitates the sharing and collection of information about other agents to derive optimal decisions. Finding such a consensus in the decision-making may happen explicitly through communication or implicitly by constructing models of other agents. The former requires skills to communicate with others so that agents can express their purpose and align their coordination. For the latter, agents need the ability to observe other agents' behavior and reason about their strategies to build a model. If the prediction model is accurate, an agent can learn the other agents' behavioral patterns and direct actions towards a consensus, leading to coordinated behavior. Besides explicit communication and constructing agent models, the CTDE scheme can be leveraged to build different levels of abstraction, which are applied to learn high-level coordination while independent skills are trained at low-level.

In the remainder of this section, we focus on methods that solve coordination issues without establishing communication protocols between agents. Although communication may ease coordination, we discuss this topic separately in Sect. 5.2.

Independent learners The naïve approach to handle multi-agent problems is to regard each agent individually such that other agents are perceived as part of the environment and, thus, are neglected during learning. Opposed to joint action learners, where agents experience the selected actions of others a-posteriori, independently learning agents face the main difficulty of coherently choosing actions such that the joint action becomes optimal concerning the mutual goal (Matignon et al. 2012b). During the learning of good policies, agents influence each other's search space, which can lead to *action shadowing*. The notion of coordination among several autonomously and independently acting agents enjoys a long record, and a bulk of research was conducted in settings with non-communicative agents (Fulda and Ventura 2007; Matignon et al. 2012b). Early works investigated the convergence of independent learners and showed that the convergence to solutions is feasible under certain conditions in deterministic games but fails in stochastic environments (Claus and Boutilier 1998; Lauer and Riedmiller 2000). Stochasticity, relative over-generalization, and other pathologies such as non-stationarity and the alter-exploration problem led to new branches of research including hysteretic learning (Matignon et al. 2007) and leniency (Potter and De Jong 1994). *Hysteretic* Q-learning was introduced to encounter the over-estimation of the value function evoked by stochasticity. Two learning rates are used to increase and decrease the value function updates while relying on an optimistic form of learning. A modern approach to hysteretic learning can be seen in Palmer et al. (2018) and Omidshafiei et al. (2017). An alternative method to adjust the degree of applied optimism during learning is *leniency* (Panait et al. 2006; Wei and Luke 2016). Leniency associates selected actions with decaying temperature values that govern the amount of applied leniency. Agents are optimistic during the early phase when exploration is still high but become less lenient for frequently visited state-action pairs over the training so that value estimations become more accurate towards the end of learning.

Further works expanded independent learners with enhanced techniques to cope with the MARL pathologies mentioned above. Extensions to the deep Q-network can be seen in additional mechanisms used for the experience replay (Palmer et al. 2019), the utilization of specialized estimators (Zheng et al. 2018a) and the use of implicit quantile networks (Lyu and Amato 2020). Further literature investigated independent learners as benchmark reference but reported limited success in cooperative tasks of various domains when no other techniques are applied to alleviate the issue of independent learners (Foerster et al. 2018b; Sunehag et al. 2018).

Constructing models An implicit way to achieve coordination among agents is to capture the behavior of others by constructing models. Models are functions that take past interaction data as input and output predictions about the agents of interest. This can be very important to render the learning process robust against the decision-making of other agents in the environment (Hu and Wellman 1998). The constructed models and the predicted behavior vary widely depending on the approaches and the assumptions being made (Albrecht and Stone 2018).

One of the first works based on deep learning methods was conducted by He et al. (2016) in an adversarial setting. They proposed an architecture that utilizes two neural networks. One neural network captures the opponents' strategies, and the second network estimates the opponents' Q-values. These networks jointly learn models of opponents by encoding observations into a deep Q-network. Another work by Foerster et al. (2018a) introduced a learning method where the policy updates rely on the impact on other agents. The opponent's policy parameters can be inferred from the observed trajectory by using a maximum likelihood technique. The arising non-stationarity is tackled by accounting only recent data. An additional possibility is to address the information gain about other agents through Bayesian methods. Raileanu et al. (2018) employed a model where agents estimate the other agents' hidden states and embed these estimations into their own policy. Inferring other agents' hidden states from their behavior allows them to choose appropriate actions and promotes eventual coordination. Foerster et al. (2019) used all publicly available observations in the environment to calculate a public belief over agents' local information. Another work by Yang et al. (2018a) used Bayesian techniques to detect opponent strategies in competitive games. A particular challenge is to learn agent models in the presence of fast adapting agents, which amplifies the problem of non-stationarity. As a countermeasure, Everett and Roberts (2018) proposed the switching agent model (SAM), which learns a set of opponent models and a switching mechanism between models. By tracking and detecting the behavioral adaption of other agents, the switching mechanism learns to select the best response from the learned set of opponent models and, thus, showed superior performance over single model learners.

Further works on constructing models can be found in cooperative tasks (Barde et al. 2019; Tacchetti et al. 2019; Zheng et al. 2018b) with imitation learning (Grover et al. 2018; Le et al. 2017), in social dilemmas (Jaques et al. 2019; Letcher et al. 2019), and by predicting behaviors from observations (Hong et al. 2017; Hoshen 2017). For a comprehensive survey on constructing models in multi-agent systems, one may consider the work of Albrecht and Stone (2018).

Besides resolving the coordination problem, building models of other agents can cope with the non-stationarity in the environment. As soon as one agent has knowledge about others' behavior, previously unexplainable transition dynamics can be attributed to the responsible agents, and the environment becomes stationary again from the viewpoint of an individual agent.

Hierarchical methods Learning to coordinate can be challenging if multiple decision-makers are involved due to the increasing complexity (Bernstein et al. 2002). An approach to deal with the coordination problem is by abstracting low-level coordination to higher levels. The idea originated in the single-agent domain where hierarchies for temporal abstraction are employed to ease long-term reward assignments (Dayan and Hinton 1993; Sutton et al. 1999). Lower levels entail only partial information of the higher levels so that the learning task becomes simpler the lower the level of abstraction. First attempts for hierarchical multi-agent RL can be found in the tabular case (Ghavamzadeh et al. 2006; Makar et al. 2001). A deep approach was proposed by Kumar et al. (2017), where a higher-level controller guides the information exchange between decentralized agents. Grounded on the high-level controller, the agents communicate with only one other agent at each time step, which allows the exploration of distributed policies. Another work by Han et al. (2019) is built upon the options framework (Sutton et al. 1999) where they embedded a dynamic termination criterion for Q-learning. By adding a termination criterion, agents could flexibly quit the option execution and react to the behavioral changes of other agents. Related to the idea of feudal networks (Dayan and Hinton 1993), Ahilan and Dayan (2019) applied a two-level abstraction of agents to a cooperative multi-agent setting where, in contrast to other methods, the hierarchy relied on rewards instead of state goals. They showed that this approach could be well suited for decentralized control problems. Jaderberg et al. (2019) used hierarchical representations that allowed agents to reason at different time scales. The authors demonstrated that agents are capable of solving mixed cooperative and competitive tasks in simulated physics environments. Another work by Lee et al. (2020) proposed a hierarchical method to coordinate two agents on robotic manipulation and locomotion tasks to accomplish collaboration such as object pick and placement. They learned primitive skills on the low-level, which are guided by a higher-level policy. Further works cover hierarchical methods in cooperation tasks (Cai et al. 2013; Ma and Wu 2020; Tang et al. 2018) or social dilemmas (Vezhnevets et al. 2019). An open challenge for hierarchical methods is the autonomous creation and discovery of abstract goals from data (Schaul et al. 2015; Vezhnevets et al. 2017).

5.4 Credit assignment problem

In the fully-cooperative setting, agents are encouraged to maximize an equally-shared reward signal. Even in a fully-observable state space, it is difficult to determine which agents and actions contributed to the eventual reward outcome when agents do not have access to the joint action. Claus and Boutilier (1998) showed that independent learners could not differentiate between the teammate's exploration and the stochasticity in the environment even in a simple bi-matrix game. This can render the learning problem difficult because agents should be ideally provided with feedback corresponding to the task performance to enable sufficient learning. Associating rewards to agents is known as the *credit assignment problem* (Weiß 1995; Wolpert and Tumer 1999). This problem is intensified by the sequential nature of reinforcement learning where agents must understand not only the impact of single actions but also the entire action sequences that eventually lead to the reward outcome (Sen and Weiss 1999). An additional challenge arises when agents have only access to local observations of the environment, which we discuss in Sect. 5.6. In the remainder of this section, we consider three actively investigated approaches that deal with how to determine the contribution of agents jointly-shared reward settings.

Decomposition Early works approached the credit assignment problem by applying filters (Chang et al. 2004) or modifying the reward function such as reward shaping (Ng et al. 1999). Recent approaches focus on exploiting dependencies between agents to decompose the reward among the agents with respect to their actual contribution towards the global reward (Kok and Vlassis 2006). The learning problem is simplified by dividing the task into smaller and, hence, easier sub-problems through decomposition. Sunehag et al. (2018) introduced the value decomposition network (VDN) which factorizes the joint action-value function into a linear combination of individual action-value functions. The VDN learns how to optimally assign an individual reward according to the agent's performance. The neural network helps to disambiguate the joint reward signal concerning the impact of the agent. Rashid et al. (2018) proposed QMIX as an improvement over VDN. QMIX learns a centralized action-value function that is decomposed into agent individual action-value functions through non-linear combinations. Under the assumption of monotonic relationships between the centralized Q-function and the individual Q-functions, decentralized policies can be extracted by individual argmax operations. As an advancement over both VDN and QMIX, Son et al. (2019) proposed QTRAN, which discards the assumption of linearity and monotonicity in the factorization and allows any non-linear combination of value functions. Further approaches about the factorization of value functions can be found in Castellini et al. (2019), Chen et al. (2018), Nguyen et al. (2017b), Wang et al. (2020a), Wang et al. (2020c) and Yang et al. (2018b).

Marginalization Next to the decomposition into simpler sub-problems, one can apply an extra function that marginalizes out the effect of agent individual actions. Nguyen et al. (2018) introduced a mean collective actor-critic framework which marginalizes out the actions of agents by using an approximation of the critic and reduces the variance of the gradient estimation. Similarly, Foerster et al. (2018b) marginalized out the individual actions of agents by applying a counterfactual baseline function. The counterfactual baseline function uses a centralized critic, which calculates the advantage of a single agent by comparing the estimated return of the current joint-action to the counterfactual baseline. The impact of a single agent's action is determined and can be attributed to the agent itself. Another work by Wu et al. (2018) used a marginalized action-value function as a baseline to reduce the variance of critic estimates. The marginalization approaches are closely related to the *difference rewards* proposed by Tumer and Wolpert (2004) who determine the impact of an agent's individual action compared to the average reward of all agents.

Inverse reinforcement learning Credit assignment problems can be evoked by a bad design of the reinforcement learning problem. Misinterpretations of the agents can lead to failure because unintentional strategies are explored, e.g. if the reward function does not capture all important aspects of the underlying task (Amodei et al. 2016). Therefore, an important step in the problem design is the reward function. However, designing a reward function can be challenging for complex problems (Hadfield-Menell et al. 2017) and becomes even more complicated for multi-agent systems since different agents may accomplish different goals. Another approach to address the credit assignment problem is by *inverse reinforcement learning* (Ng and Russell 2000) that describes how an agent learns a reward function that explains the demonstrated behavior of an expert without having access to the reward signal. The learned reward function can then be used to build strategies. The work of Lin et al. (2018) applied the principle of inverse reinforcement learning to the multi-agent setting. They showed that multiple agents could recover reward functions that are correlated with the ground truths. Related to inverse RL, imitation learning can be used to learn from expert knowledge. Yu et al. (2019) imitated expert behaviors to learn high-dimensional policies in both cooperative and competitive environments. They were able to recover the expert policies for each individual agent from the provided expert demonstrations. Further works on imitation learning consider the fully cooperative setting (Barrett et al. 2017; Le et al. 2017) and Markov Games with mixed settings (Song et al. 2018).

5.5 Scalability

Training a large number of agents is inherently difficult. Every agent involved in the environment adds extra complexity to the learning problem such that the computational effort grows exponentially by the number of agents. Besides complexity concerns, sufficient scaling also demands agents to be robust towards the behavioral adaption of other agents. However, agents can leverage the benefit of distributed knowledge shared and reused between agents to accelerate the learning process. In the following, we review approaches that address the handling of many agents and discuss possible solutions. We broadly classify the surveyed works into those that apply some form of knowledge reuse, reduce the complexity of the learning problem, and develop robustness against the policy adaptions of other agents.

Knowledge reuse The training of individual learning models does scale poorly with the increasing number of agents because the computational effort increases due to the combinatorial possibilities. Knowledge reuse strategies are employed to ease the learning process and scale RL to complex problems by reutilizing previous knowledge into new tasks. Knowledge reuse can be applied in many facets (Silva et al. 2018).

First, agents can make use of a *parameter sharing* technique if they exhibit homogeneous structures, e.g. the weights in a neural network for sharing parts or the whole learning model with others. Sharing the parameters of a policy enables an efficient training process that can scale up to an arbitrary number of agents and, thus, can boost the learning process (Gupta et al. 2017). Parameter sharing has proven to be useful in various applications such as learning to communicate (Foerster et al. 2016; Jiang and Lu 2018; Peng et al. 2017; Sukhbaatar et al. 2016), modeling agents (Hernandez-Leal et al. 2019), and in partially observable cooperative games (Sunehag et al. 2018). For a discussion on different parameter sharing strategies, one may consider the paper by Chu and Ye (2017).

As the second approach, knowledge reuse can be applied in form of *transfer learning* (Da Silva et al. 2019; Da Silva and Costa 2019). Experience obtained in learning to perform one task may also improve the performance in a related but different task (Taylor and Stone 2009). Da Silva and Costa (2017) used a knowledge database from which an agent can extract previous solutions of related tasks and embed such information into the current task's training. Likewise, Da Silva et al. (2017) applied expert demonstrations where the agents take the role of students that ask a teacher for advice. They demonstrated that simultaneously learning agents could advise each other through knowledge transfer. Further works on transfer learning can be found in the cooperative multi-agent setting (Omidshafiei et al. 2019) and in natural language applications (Luketina et al. 2019). In general multi-agent systems, the works of (Boutsioukis et al. 2012; Taylor et al. 2013) substantiate that transfer learning can speed up the learning process.

Besides parameter sharing and transfer learning, *curriculum learning* may be applied for the scaling to many agents. Since tasks become more challenging to master and more time consuming to train as the number of agents increases, it is often challenging to learn from scratch. Curriculum learning starts with a small number of agents and then gradually enlarges the number of agents over the training course. Through the steady increase within the curriculum, trained policies can perform better than without a curriculum (Gupta et al. 2017; Long et al. 2020; Narvekar et al. 2016). Curriculum learning schemes can also cause improved generalization and faster convergence of agent policies (Bengio et al. 2009). Further works show that agents can generate learning curricula automatically (Sukhbaatar et al. 2017; Svetlik et al. 2017) or can create arms races in competitive settings (Baker et al. 2020).

Complexity reduction Many real-world applications naturally encompass large numbers of simultaneously interacting agents (Nguyen et al. 2017a, b). As the quantity of agents increases, the requirement to contain the curse of dimensionality becomes inevitable. Yang et al. (2018b) addressed the issue of scalability with a mean-field method. The interactions between large numbers of agents are estimated by the impact of a single agent compared to the mean impact of the whole or local agent population. The complexity reduces as the problem is broken down into pairwise interactions between an agent and its neighborhood. Regarding the average effect to its neighbors, each agent learns the best response towards its proximity. Another approach to constrain the explosion in complexity is by factorizing the problem into smaller sub-problems (Guestrin et al. 2002). Chen et al. (2018) decomposed the joint action-value function into independent components and used pairwise interactions between agents to render large-scale problems computationally tractable. Further works studied large-scale MADRL problems with graphical models (Nguyen et al. 2017a) and the CTDE paradigm (Lin et al. 2018).

Robustness Another desired property is the *robustness* of learned policies to perturbations in the environment caused by other agents. Perturbations are fortified by the number of agents and the resulting growth of the state-action space. In supervised learning, a common problem is that models can over-fit to the data set. Similarly, over-fitting can occur in RL frameworks if environments provide little or no deviation (Bansal et al. 2018). To maintain robustness over the training process and to the other agents' adaption, several methods have been proposed.

First, *regularization* techniques can be used to prevent over-fitting to other agents' behavior. Examples can be seen in policies ensembles (Lowe et al. 2017), where a collection of different sub-policies is trained for each agent, or can be found in best responses to policy mixtures (Lanctot et al. 2017).

Second, *adversarial training* can be applied to mitigate the vulnerability of polices towards perturbations. Pinto et al. (2017) added an adversarial agent to the environment that applied targeted disturbances to the learning process. By hampering the training, the agents were compelled to encounter these disturbances and develop robust policies. Similarly, Li et al. (2019a) used an adversarial setting to reduce the sensitivity of agents towards the environment. Bansal et al. (2018) demonstrated that policies, which are trained in a competitive setting, could yield behaviors that are far more complex than the environment itself. From an application perspective, Spooner and Savani (2020) studied robust decision-making in market making.

The observations from above are in accordance with the findings of related studies about the impact of *self-play* (Raghu et al. 2018; Sukhbaatar et al. 2017). Heinrich and Silver (2016) used self-play to learn approximate Nash equilibria of imperfect-information games and showed that self-play could be used to obtain better robustness in the learned policies. Similarly, self-play was used to compete with older versions of policies to render the learned behaviors more robust (Baker et al. 2020; Berner et al. 2019; Silver et al. 2018). Silver et al. (2016) adapted self-play as a regularization technique to prevent the policy network from over-fitting by playing against older versions of itself. However, Gleave et al. (2020) studied the existence of adversarial policies in competitive games and showed that complex policies could be fooled by comparably easy strategies. Although agents trained through self-play proved to be more robust, allegedly random and uncoordinated strategies caused agents to fail at the task. They argued that the vulnerability towards adversarial attacks increases with the dimensionality of the observation space. A further research direction for addressing robustness is to render the learning representation invariant towards permutations, as shown in Liu et al. (2020).

5.6 Partial observability

Outside an idealized setting, agents neither can observe the global state of the environment, nor do they have access to the internal knowledge of other agents. By perceiving only partial observations, a single observation does not capture all relevant information about the environment and its history. Hence, the Markov property is not fulfilled, and the environment appears non-Markovian. An additional difficulty elicited by partial observability is the *lazy agent problem* which can occur in cooperative settings (Sunehag et al. 2018). As introduced in Sect. 2.2, the common frameworks that deal with partial observability are POMPDs for general settings and dec-POMDPs for cooperative settings with a shared reward function. Dec-POMDPs are computationally challenging (Bernstein et al. 2002) and still intractable when solving problems with real-world complexity (Amato et al. 2015). However, recent work accomplished promising results in video games with imperfect information (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019; Vinyals et al. 2019).

A natural way to deal with non-Markovian environments is through information exchange between the decision-makers (Goldman and Zilberstein 2004). Agents that are able to communicate can compensate for their limited knowledge by propagating information and fill the lack of knowledge about other agents or the environment (Foerster et al.

2016). As we already discussed in Sect. 5.2, there are several ways to incorporate communication capabilities into agents. A primary example is Jiang and Lu (2018) who used an attention mechanism to establish communication under partial observations. Rather than having a fixed frequency for the information exchange, they learned to communicate on-demand. Further approaches under partial observability have been investigated in cooperative tasks (Das et al. 2019; Sukhbaatar et al. 2016) or mixed settings (Singh et al. 2019).

In the following, we review papers that cope with partial observability by incorporating a memory mechanism. Agents, which have the capability of memorizing past experiences, can compensate for the lack of information.

Memory mechanism A common way to tackle partial observability is the usage of deep recurrent neural networks, which equip agents with a memory mechanism to store information that can be relevant in the future (Hausknecht and Stone 2015). However, long-term dependencies render the decision-making difficult since experiences that were observed in the further past may have been forgotten (Hochreiter and Schmidhuber 1997). Approaches involving recurrent neural networks to deal with partial observability can be realized with value-based approaches (Omidshafiei et al. 2017) or actor-critic methods (Dibangoye and Buffet 2018; Foerster et al. 2018b; Gupta et al. 2017). Foerster et al. (2019) used a Bayesian method to tackle partial observability in cooperative settings. They used all publicly available features of the environment and agents to determine a public belief over the agents' internal states. A severe concern in MADRL is that the memorization of past information is exacerbated by the number of agents involved during the learning process.

6 Discussion

In this section, we discuss findings from previous sections. We enumerate trends that we have identified in recent literature. Since these trends are useful for addressing current challenges, they may also be an avenue for upcoming research. To the end of our discussion, we point out possible future work. We elaborate on problems where only a minority of research has been conducted and pose two problems which we find the toughest ones to overcome.

Despite the recent advances in many directions, many pathologies such as relative overgeneralization combined with reward stochasticity are not yet solved, even in allegedly simple tabular worlds. MADRL has taken profit from the history of MARL by scaling up the insights to more complex problems. Approaches where strong solutions exist in simplified MARL settings may be transferable to the MADRL domain. Thus by enhancing older methods with new deep learning approaches, unsolved problems and concepts from MARL continue to matter in MADRL. An essential point for MADRL is that reproducibility is taken conscientiously. Well-known papers from the single-agent domain underline the significance of hyper-parameters, the number of independent random seeds, and chosen code-base towards the eventual task performance (Henderson et al. 2018; Islam et al. 2017). To maintain steady progress, the reporting of all used hyper-parameters and a transparent conduction of experiments is crucial. We want to make the community aware that these findings may also be valid for the multi-agent domain. Therefore, it is inevitable that standardized frameworks are created in which different algorithms can be compared along with their merits and demerits. Many individual environments have been proposed which exhibit intricate structure and real-world complexity (Baker et al. 2020; Beattie et al. 2016; Johnson et al. 2016; Juliani et al. 2018; Song et al. 2019; Vinyals et al. 2017). However,

Table 4 Our identified trends in MADRL and the addressed challenges

Trend	Addressed challenge(s)
Curriculum learning	Scalability
Memory	Non-stationarity, partial observability
Communication	Non-stationarity, coordination, partial observability
CTDE	Non-stationarity, coordination, partial observability, credit assignment, scalability

no consistent benchmark yet exists that provides a unified interface and allows a fair comparison between different kinds of algorithms grounded on a great variety of tasks like the *OpenAI Gym* (Brockman et al. 2016) for single-agent problems.

6.1 Trends

Over the last years, approaches in the multi-agent domain achieved successes based on recurring patterns of good practice. We have identified four trends in state-of-the-art literature that have been frequently applied to address current challenges (Table 4).

As the first trend, we observe curriculum learning as an approach to divide the learning process into stages to deal with scalability issues. By starting with a small quantity, the number of agents is gradually enlarged over the learning course so that large-scale training becomes feasible (Gupta et al. 2017; Long et al. 2020; Narvekar et al. 2016). Alternatively, curricula can also be employed to create different stages of difficulty, where agents face relatively easy tasks at the beginning and gradually more complex tasks as their skills increase (Vinyals et al. 2019). Besides that, curriculum training is used to investigate the emergence of agent behavior. Curricula describe engineered changes in the dynamics of the environment. Agents adapt their behaviors over time in response to the strategic changes of others, which can yield arms races between agents. This process of continual adaption is referred to *autocurricula* (Leibo et al. 2019), which have been reported in several works (Baker et al. 2020; Sukhbaatar et al. 2017; Svetlik et al. 2017).

Second, we recognize a trend towards deep neural networks embedded with recurrent units to memorize experience. By having the ability to track the history of state transitions and the decisions of other agents, the non-stationarity of the environment due to multiple decision-makers and partially observable states can be addressed in small problems (Omidshafiei et al. 2017), and can be managed sufficiently well in complex problems (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019).

Third, an active line of research is exploring the development of communication skills. Due to the rise of deep learning methods, new computational approaches are available to investigate the emergence of language between interactive agents (Lazaridou and Baroni 2020). Despite the plethora of works that analyze emergent behaviors and semantics, many works propose methods that endow agents with communication skills. By expressing their intension, agents can align their coordination and find a consensus (Foerster et al. 2016). The non-stationarity from the perspective of a single learner can be eluded when agents disclose their history. Moreover, agents can share their local information with others to alleviate partial observability (Foerster et al. 2018b; Omidshafiei et al. 2017).

Fourth and last, we note a clear trend towards the CTDE paradigm that enables the sharing of information during the training. Local information such as the observation-action history, function values, or policies can be made available to all agents during the training, which renders the environment stationary from the viewpoint of an individual agent and may diminish partial observability (Lowe et al. 2017). Further, the credit assignment problem can be addressed when information is available about all agents, and a centralized mechanism can attribute the individual contribution to the respective agent (Foerster et al. 2018b). Further challenges that can be loosened are coordination and scalability when the lack of information of an individual agent is compensated, and the learning process is accelerated (Gupta et al. 2017).

6.2 Future work

Next to our identified trends, which are already under active research, we recognize areas that have not been sufficiently explored yet. One such area is *multi-goal learning* where each agent has an individually associated goal that needs to be optimized. However, global optimality can only be accomplished if agents also allow others to be successful in their task (Yang et al. 2020). Typical scenarios are cooperative tasks such as public good dilemmas, where agents are obliged to the sustainable use of limited resources, or autonomous driving, where agents have individual destinations and are supposed to coordinate the path-finding to avoid crashes. A similar direction is *multi-task* learning where agents are expected to perform well not only on one single but also on related other tasks (Omidshafiei et al. 2017; Taylor and Stone 2009). Besides multi-goal and multi-task learning, another avenue for future work is present in *safe* MADRL. Safety is a highly desired property because autonomously acting agents are expected to ensure system performance while holding to safety guarantees during learning and employment (García et al. 2015). Several works in single-agent RL are concerned with safety concepts, but its applicability to multiple agents is limited and still in its infancy (Zhang and Bastani 2019; Zhu et al. 2020). Akin to the growing interest in learning to communicate, a similar effect may happen in the multi-agent domain, where deep learning methods open new paths. For an application perspective on safe autonomous driving, one can consider the article by Shalev-Shwartz et al. (2016). Another possible direction for future research offers the intersection between MADRL and *evolutionary* methodologies. Evolutionary algorithms have been used in versatile contexts of multi-agent RL, e.g. for building intrinsic motivation (Wang et al. 2019), shaping rewards (Jaderberg et al. 2019), generating curricula (Long et al. 2020) and analyzing dynamics (Bloembergen et al. 2015). Since evolution requires many entities to adapt, multi-agent RL is a natural playground for such algorithms.

Beyond the current challenges and reviewed literature of Sect. 5, we identify two problems that we regard as the most challenging problems to overcome by future work. We primarily choose these two problems since they are the ones that matter the most when it comes to the applicability of algorithms to real-world scenarios. Most research focuses on learning within homogeneous settings where agents share common interests and optimize a mutual goal. For instance, the learning of communication is mainly studied in dec-POMDPs, where agents are expected to optimize upon a joint reward signal. When agents share common interests, the CTDE paradigm is usually a beneficial choice to exchange information between agents, and problems like non-stationarity, partial observability, and coordination can be diminished. However, *heterogeneity* implies that agents may have their own interests and goals, individual experience and knowledge, or different skills and capabilities. Limited research has been conducted in heterogeneous scenarios, although many real-world problems naturally comprise a mixture of different entities. Under real-world conditions, agents have only access to local and heterogeneous information on which decisions must be taken. The fundamental problem in the multi-agent domain is and ever has been the *curse of dimensionality* (Busoniu et al. 2008; Hernandez-Leal et al. 2019). The state-action space and the combinatorial possibilities of agent interactions increase exponentially by the number of agents, which renders sufficient exploration itself a difficult problem. This is intensified when agents have only access to partial observations of the environment or when the environment is of continuous nature. Although powerful function approximators like neural networks can cope with continuous spaces and generalize well over large spaces, open questions remain like how to explore large and complex spaces sufficiently well and how to solve large combinatorial optimization problems.

7 Conclusion

Even though multi-agent reinforcement learning enjoys a long record, historical approaches hardly exceeded the complexity of discretized environments with a limited amount of states and actions (Busoniu et al. 2008; Tuyls and Weiss 2012). Since the breakthrough of deep learning methods, the field is undergoing a rapid transformation, and many previously unsolved problems have become step by step tractable. Latest advances showed that tasks with real-world complexity could be mastered (Baker et al. 2020; Berner et al. 2019; Jaderberg et al. 2019; Vinyals et al. 2019). Still, MADRL is a young field which attracts growing interest, and the amount of published literature rises swiftly. In this article, we surveyed recent works that combine deep learning methods with multi-agent reinforcement learning. We analyzed training schemes that are used to learn policies, and we reviewed patterns of agent behavior that emerge when multiple entities interact simultaneously. In addition, we systematically investigated challenges that are present in the multi-agent context and studied recent approaches that are under active research. Finally, we outlined trends which we have identified in state-of-the-art literature and proposed possible avenues for future work. With this contribution, we want to equip interested readers with the necessary tools to understand the contemporary challenges in MADRL by providing a more holistic overview of the recent approaches. We want to emphasize its potential and reveal opportunities as well as its limitations. In the foreseeable future, we expect an abundance of new literature to emanate and, hence, we want to encourage the community for further developments in this interesting and young field of research.

Acknowledgements We would like to thank the editor and the three anonymous reviewers for providing their comprehensive feedback. Without their suggestions, this manuscript would not look as it does in this final version. We want to thank our colleagues and friends who read through earlier versions of this manuscript. In particular, we appreciate the help of Matthias Kissel, Patrick Krämer, Anke Müller and Martin Gottwald.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Ahilan S, Dayan P (2019) Feudal multi-agent hierarchies for cooperative reinforcement learning. CoRR arxiv: abs/1901.08492

Al-Shedivat M, Bansal T, Burda Y, Sutskever I, Mordatch I, Abbeel P (2018) Continuous adaptation via meta-learning in nonstationary and competitive environments. In: International conference on learning representations. https://openreview.net/forum?id=Sk2u1g-0-

Albrecht SV, Stone P (2018) Autonomous agents modelling other agents: a comprehensive survey and open problems. Artif Intell 258:66–95. https://doi.org/10.1016/j.artint.2018.01.002. http://www.sciencedirect.com/science/article/pii/S0004370218300249

Amato C, Konidaris G, Cruz G, Maynor CA, How JP, Kaelbling LP (2015) Planning for decentralized control of multiple robots under uncertainty. In: 2015 IEEE international conference on robotics and automation (ICRA), pp 1241–1248. https://doi.org/10.1109/ICRA.2015.7139350

Amodei D, Olah C, Steinhardt J, Christiano PF, Schulman J, Mané D (2016) Concrete problems in AI safety. CoRR. arxiv: abs/1606.06565,

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D (2015) Vqa: Visual question answering. In: The IEEE international conference on computer vision (ICCV)

Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) Deep reinforcement learning: a brief survey. IEEE Signal Process Mag 34(6):26–38. https://doi.org/10.1109/MSP.2017.2743240

Aubret A, Matignon L, Hassas S (2019) A survey on intrinsic motivation in reinforcement learning. arXiv e-prints arXiv:1908.06976,

Baker B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, Mordatch I (2020) Emergent tool use from multi-agent autocurricula. In: International conference on learning representations. https://openreview.net/forum?id=SkxpxJBKwS

Bansal T, Pachocki J, Sidor S, Sutskever I, Mordatch I (2018) Emergent complexity via multi-agent competition. In: International conference on learning representations. https://openreview.net/forum?id=Sy0GnUxCb

Barde P, Roy J, Harvey FG, Nowrouzezahrai D, Pal C (2019) Promoting coordination through policy regularization in multi-agent reinforcement learning. arXiv e-prints arXiv:1908.02269,

Barrett S, Rosenfeld A, Kraus S, Stone P (2017) Making friends on the fly: cooperating with new teammates. Artif Intell 242:132–171

Beattie C, Leibo JZ, Teplyashin D, Ward T, Wainwright M, Küttler H, Lefrancq A, Green S, Valdés V, Sadik A, Schrittwieser J, Anderson K, York S, Cant M, Cain A, Bolton A, Gaffney S, King H, Hassabis D, Legg S, Petersen S (2016) Deepmind lab. CoRR. arxiv: abs/1612.03801

Becker R, Zilberstein S, Lesser V, Goldman CV (2004) Solving transition independent decentralized Markov decision processes. J Artif Intell Res 22:423–455

Bellemare M, Srinivasan S, Ostrovski G, Schaul T, Saxton D, Munos R (2016) Unifying count-based exploration and intrinsic motivation. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29, Curran Associates, Inc., pp 1471–1479. http://papers.nips.cc/paper/6383-unifying-count-based-exploration-and-intrinsic-motivation.pdf

Bellman R (1957) A Markovian decision process. J Math Mechanics 6(5):679–684. http://www.jstor.org/stable/24900506

Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning, ACM, New York, NY, USA, ICML '09, pp 41–48. https://doi.org/10.1145/1553374.1553380,

Berner C, Brockman G, Chan B, Cheung V, Debiak P, Dennison C, Farhi D, Fischer Q, Hashme S, Hesse C, Józefowicz R, Gray S, Olsson C, Pachocki JW, Petrov M, de Oliveira Pinto HP, Raiman J, Salimans T, Schlatter J, Schneider J, Sidor S, Sutskever I, Tang J, Wolski F, Zhang S (2019) Dota 2 with large scale deep reinforcement learning. ArXiv arxiv: abs/1912.06680

Bernstein DS, Givan R, Immerman N, Zilberstein S (2002) The complexity of decentralized control of Markov decision processes. Math Oper Res 27(4):819–840. https://doi.org/10.1287/moor.27.4.819.297

Bertsekas DP (2012) Dynamic programming and optimal control, vol 2, 4th edn. Athena Scientific, Belmont

Bertsekas DP (2017) Dynamic programming and optimal control, vol 1, 4th edn. Athena Scientific, Belmont

Bloembergen D, Tuyls K, Hennes D, Kaisers M (2015) Evolutionary dynamics of multi-agent learning: a survey. J Artif Intell Res 53:659–697

Bono G, Dibangoye JS, Matignon L, Pereyron F, Simonin O (2019) Cooperative multi-agent policy gradient. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G (eds) Machine learning and knowledge discovery in databases. Springer International Publishing, Cham, pp 459–476

Boutsioukis G, Partalas I, Vlahavas I (2012) Transfer learning in multi-agent reinforcement learning domains. In: Sanner S, Hutter M (eds) Recent advances in reinforcement learning. Springer, Berlin, pp 249–260

Bowling M, Veloso M (2002) Multiagent learning using a variable learning rate. Artif Intell 136(2):215–250

Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W (2016) Openai gym. arXiv:1606.01540

Busoniu L, Babuska R, De Schutter B (2008) A comprehensive survey of multiagent reinforcement learning. IEEE Trans Syst Man Cybern Part C (Appl Rev) 38(2):156–172. https://doi.org/10.1109/TSMCC.2007.913919

Cai Y, Yang SX, Xu X (2013) A combined hierarchical reinforcement learning based approach for multi-robot cooperative target searching in complex unknown environments. In: 2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL), pp 52–59. https://doi.org/10.1109/ADPRL.2013.6614989

Cao K, Lazaridou A, Lanctot M, Leibo JZ, Tuyls K, Clark S (2018) Emergent communication through negotiation. In: International conference on learning representations. https://openreview.net/forum?id=Hk6WhagRW

Cao Y, Yu W, Ren W, Chen G (2013) An overview of recent progress in the study of distributed multi-agent coordination. IEEE Trans Industr Inf 9(1):427–438. https://doi.org/10.1109/TII.2012.2219061

Castellini J, Oliehoek FA, Savani R, Whiteson S (2019) The representational capacity of action-value networks for multi-agent reinforcement learning. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, international foundation for autonomous agents and multiagent systems, Richland, SC, AAMAS '19, pp 1862–1864. http://dl.acm.org/citation.cfm?id=3306127.3331944

Celikyilmaz A, Bosselut A, He X, Choi Y (2018) Deep communicating agents for abstractive summarization. CoRR arxiv: abs/1803.10357,

Chang Y, Ho T, Kaelbling LP (2004) All learning is local: Multi-agent learning in global reward games. In: Thrun S, Saul LK, Schölkopf B (eds) Advances in neural information processing systems 16, MIT Press, pp 807–814. http://papers.nips.cc/paper/2476-all-learning-is-local-multi-agent-learning-in-global-reward-games.pdf

Chen Y, Zhou M, Wen Y, Yang Y, Su Y, Zhang W, Zhang D, Wang J, Liu H (2018) Factorized q-learning for large-scale multi-agent systems. CoRR arxiv: abs/1809.03738

Chen YF, Liu M, Everett M, How JP (2016) Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. CoRR. arxiv: abs/1609.07845,

Chentanez N, Barto AG, Singh SP (2005) Intrinsically motivated reinforcement learning. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems 17, MIT Press, pp 1281–1288. http://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf

Choi E, Lazaridou A, de Freitas N (2018) Multi-agent compositional communication learning from raw visual input. In: International conference on learning representations. https://openreview.net/forum?id=rknt2Be0-

Chu T, Chinchali S, Katti S (2020) Multi-agent reinforcement learning for networked system control. In: International conference on learning representations. https://openreview.net/forum?id=Syx7A3NFvH

Chu T, Wang J, Codecà L, Li Z (2020) Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans Intell Transp Syst 21(3):1086–1095

Chu X, Ye H (2017) Parameter sharing deep deterministic policy gradient for cooperative multi-agent reinforcement learning. CoRR arxiv: abs/1710.00336

Claus C, Boutilier C (1998) The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of the fifteenth national conference on artificial intelligence and tenth innovative applications of artificial intelligence conference, AAAI 98, IAAI 98, July 26–30, 1998, Madison, Wisconsin, USA, pp 746–752. http://www.aaai.org/Library/AAAI/1998/aaai98-106.php

Crandall JW, Goodrich MA (2011) Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. Mach Learn 82(3):281–314. https://doi.org/10.1007/s10994-010-5192-9

Da Silva FL, Costa AHR (2017) Accelerating multiagent reinforcement learning through transfer learning. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, AAAI Press, AAAI'17, pp 5034–5035. http://dl.acm.org/citation.cfm?id=3297863.3297988

Da Silva FL, Costa AHR (2019) A survey on transfer learning for multiagent reinforcement learning systems. J Artif Int Res 64(1):645–703. https://doi.org/10.1613/jair.1.11396

Da Silva FL, Glatt R, Costa AHR (2017) Simultaneously learning and advising in multiagent reinforcement learning. In: Proceedings of the 16th conference on autonomous agents and multiagent systems, international foundation for autonomous agents and multiagent systems, Richland, SC, AAMAS '17, pp 1100–1108. http://dl.acm.org/citation.cfm?id=3091210.3091280

Da Silva FL, Warnell G, Costa AHR, Stone P (2019) Agents teaching agents: a survey on inter-agent transfer learning. Auton Agent Multi-Agent Syst 34(1):9. https://doi.org/10.1007/s10458-019-09430-0

Das A, Kottur S, Moura JMF, Lee S, Batra D (2017) Learning cooperative visual dialog agents with deep reinforcement learning. In: The IEEE international conference on computer vision (ICCV)

Das A, Gervet T, Romoff J, Batra D, Parikh D, Rabbat M, Pineau J (2019) TarMAC: Targeted multi-agent communication. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of machine learning research, vol 97, pp 1538–1546. http://proceedings.mlr.press/v97/das19a.html

Dayan P, Hinton GE (1993) Feudal reinforcement learning. In: Hanson SJ, Cowan JD, Giles CL (eds) Advances in neural information processing systems 5, Morgan-Kaufmann, pp 271–278. http://papers.nips.cc/paper/714-feudal-reinforcement-learning.pdf

De Cote EM, Lazaric A, Restelli M (2006) Learning to cooperate in multi-agent social dilemmas. In: Proceedings of the fifth international joint conference on autonomous agents and multiagent systems, ACM, New York, NY, USA, AAMAS '06, pp 783–785. https://doi.org/10.1145/1160633.1160770

Diallo EAO, Sugiyama A, Sugawara T (2017) Learning to coordinate with deep reinforcement learning in doubles pong game. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA), pp 14–19. https://doi.org/10.1109/ICMLA.2017.0-184

Dibangoye J, Buffet O (2018) Learning to act in decentralized partially observable MDPs. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 1233–1242. http://proceedings.mlr.press/v80/dibangoye18a.html

Dobbe R, Fridovich-Keil D, Tomlin C (2017) Fully decentralized policies for multi-agent systems: an information theoretic approach. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 2941–2950. http://papers.nips.cc/paper/6887-fully-decentralized-policies-for-multi-agent-systems-an-information-theoretic-approach.pdf

Duan Y, Schulman J, Chen X, Bartlett PL, Sutskever I, Abbeel P (2016) Rl: fast reinforcement learning via slow reinforcement learning. CoRR arxiv: abs/1611.02779,

Eccles T, Bachrach Y, Lever G, Lazaridou A, Graepel T (2019) Biases for emergent communication in multi-agent reinforcement learning. In: Wallach H, Larochelle H, Beygelzimer A, Alche-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems 32, Curran Associates, Inc., pp 13111–13121. http://papers.nips.cc/paper/9470-biases-for-emergent-communication-in-multi-agent-reinforcement-learning.pdf

Everett R, Roberts S (2018) Learning against non-stationary agents with opponent modelling and deep reinforcement learning. In: 2018 AAAI Spring symposium series

Evtimova K, Drozdov A, Kiela D, Cho K (2018) Emergent communication in a multi-modal, multi-step referential game. In: International conference on learning representations. https://openreview.net/forum?id=rJGZq6g0-

Finn C, Levine S (2018) Meta-learning and universality: deep representations and gradient descent can approximate any learning algorithm. In: International conference on learning representations. https://openreview.net/forum?id=HyjC5yWCW

Foerster J, Assael IA, de Freitas N, Whiteson S (2016) Learning to communicate with deep multi-agent reinforcement learning. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29, Curran Associates, Inc., pp 2137–2145. http://papers.nips.cc/paper/6042-learning-to-communicate-with-deep-multi-agent-reinforcement-learning.pdf

Foerster J, Nardelli N, Farquhar G, Afouras T, Torr PHS, Kohli P, Whiteson S (2017) Stabilising experience replay for deep multi-agent reinforcement learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 1146–1155. http://proceedings.mlr.press/v70/foerster17b.html

Foerster J, Chen RY, Al-Shedivat M, Whiteson S, Abbeel P, Mordatch I (2018a) Learning with opponent-learning awareness. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 122–130. http://dl.acm.org/citation.cfm?id=3237383.3237408

Foerster J, Farquhar G, Afouras T, Nardelli N, Whiteson S (2018b) Counterfactual multi-agent policy gradients. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17193

Foerster J, Song F, Hughes E, Burch N, Dunning I, Whiteson S, Botvinick M, Bowling M (2019) Bayesian action decoder for deep multi-agent reinforcement learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 1942–1951. http://proceedings.mlr.press/v97/foerster19a.html

Fulda N, Ventura D (2007) Predicting and preventing coordination problems in cooperative q-learning systems. In: Proceedings of the 20th international joint conference on artifical intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'07, pp 780–785

García J, Fern, o Fernández (2015) A comprehensive survey on safe reinforcement learning. J Mach Learn Res 16(42):1437–1480. http://jmlr.org/papers/v16/garcia15a.html

Ghavamzadeh M, Mahadevan S, Makar R (2006) Hierarchical multi-agent reinforcement learning. Auton Agent Multi-Agent Syst. https://doi.org/10.1007/s10458-006-7035-4

Gleave A, Dennis M, Wild C, Kant N, Levine S, Russell S (2020) Adversarial policies: Attacking deep reinforcement learning. In: International conference on learning representations. https://openreview.net/forum?id=HJgEMpVFwB

Goldman CV, Zilberstein S (2004) Decentralized control of cooperative systems: categorization and complexity analysis. J Artif Int Res 22(1):143–174. http://dl.acm.org/citation.cfm?id=1622487.1622493

Grover A, Al-Shedivat M, Gupta J, Burda Y, Edwards H (2018) Learning policy representations in multiagent systems. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 1802–1811. http://proceedings.mlr.press/v80/grover18a.html

Guestrin C, Koller D, Parr R (2002) Multiagent planning with factored mdps. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing systems 14, MIT Press, pp 1523–1530. http://papers.nips.cc/paper/1941-multiagent-planning-with-factored-mdps.pdf

Gupta JK, Egorov M, Kochenderfer M (2017) Cooperative multi-agent control using deep reinforcement learning. In: Sukthankar G, Rodriguez-Aguilar JA (eds) autonomous agents and multiagent systems. Springer, Cham, pp 66–83

Hadfield-Menell D, Milli S, Abbeel P, Russell SJ, Dragan A (2017) Inverse reward design. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 6765–6774. http://papers.nips.cc/paper/7253-inverse-reward-design.pdf

Han D, Boehmer W, Wooldridge M, Rogers A (2019) Multi-agent hierarchical reinforcement learning with dynamic termination. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, pp 2006–2008. http://dl.acm.org/citation.cfm?id=3306127.3331992

Hansen EA, Bernstein D, Zilberstein S (2004) Dynamic programming for partially observable stochastic games. In: AAAI

Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248

Hausknecht M, Stone P (2015) Deep recurrent q-learning for partially observable mdps. https://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11673

Havrylov S, Titov I (2017) Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 2149–2159. http://papers.nips.cc/paper/6810-emergence-of-language-with-multi-agent-games-learning-to-communicate-with-sequences-of-symbols.pdf

He H, Boyd-Graber J, Kwok K, III HD (2016) Opponent modeling in deep reinforcement learning. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd international conference on machine learning, PMLR, New York, New York, USA, Proceedings of Machine Learning Research, vol 48, pp 1804–1813. http://proceedings.mlr.press/v48/he16.html

He H, Chen D, Balakrishnan A, Liang P (2018) Decoupling strategy and generation in negotiation dialogues. CoRR arxiv: abs/1808.09637,

Heinrich J, Silver D (2016) Deep reinforcement learning from self-play in imperfect-information games. CoRR arxiv: abs/1603.01121, Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D (2018) Deep reinforcement learning that matters. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669

Hernandez-Leal P, Kaisers M, Baarslag T, de Cote EM (2017) A survey of learning in multiagent environments: dealing with non-stationarity. CoRR arxiv: abs/1707.09183,

Hernandez-Leal P, Kartal B, Taylor ME (2019) Agent modeling as auxiliary task for deep reinforcement learning. CoRR arxiv: abs/1907.09597,

Hernandez-Leal P, Kartal B, Taylor ME (2019) A survey and critique of multiagent deep reinforcement learning. Auton Agent Multi-Agent Syst 33(6):750–797. https://doi.org/10.1007/s10458-019-09421-1

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hong Z, Su S, Shann T, Chang Y, Lee C (2017) A deep policy inference q-network for multi-agent systems. CoRR arxiv: abs/1712.07893,

Hoshen Y (2017) Vain: Attentional multi-agent predictive modeling. In: Proceedings of the 31st international conference on neural information processing systems, Curran Associates Inc., USA, NIPS'17, pp 2698–2708. http://dl.acm.org/citation.cfm?id=3294996.3295030

Houthooft R, Chen X, Chen X, Duan Y, Schulman J, De Turck F, Abbeel P (2016) Vime: variational information maximizing exploration. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in Neural Information Processing Systems 29, Curran Associates, Inc., pp 1109–1117. http://papers.nips.cc/paper/6591-vime-variational-information-maximizing-exploration.pdf

Hu J, Wellman MP (1998) Multiagent reinforcement learning: theoretical framework and an algorithm. In: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, ICML '98, pp 242–250. http://dl.acm.org/citation.cfm?id=645527.657296

Hu J, Wellman MP (2003) Nash q-learning for general-sum stochastic games. J Mach Learn Res 4:1039–1069

Hughes E, Leibo JZ, Phillips M, Tuyls K, Dueñez Guzman E, García Castañeda A, Dunning I, Zhu T, McKee K, Koster R, Roff H, Graepel T (2018) Inequity aversion improves cooperation in intertemporal social dilemmas. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 3326–3336. http://papers.nips.cc/paper/7593-inequity-aversion-improves-cooperation-in-intertemporal-social-dilemmas.pdf

Iqbal S, Sha F (2019) Actor-attention-critic for multi-agent reinforcement learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of machine learning research, vol 97, pp 2961–2970. http://proceedings.mlr.press/v97/iqbal19a.html

Islam R, Henderson P, Gomrokchi M, Precup D (2017) Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. CoRR arxiv: abs/1708.04133,

Jaderberg M, Czarnecki WM, Dunning I, Marris L, Lever G, Castañeda AG, Beattie C, Rabinowitz NC, Morcos AS, Ruderman A, Sonnerat N, Green T, Deason L, Leibo JZ, Silver D, Hassabis D, Kavukcuoglu K, Graepel T (2019) Human-level performance in 3d multiplayer games with population-based reinforcement learning. Science 364(6443):859–865

Jain U, Weihs L, Kolve E, Rastegari M, Lazebnik S, Farhadi A, Schwing AG, Kembhavi A (2019) Two body problem: Collaborative visual task completion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)

Jaques N, Lazaridou A, Hughes E, Gülçehre Ç, Ortega PA, Strouse D, Leibo JZ, de Freitas N (2018) Intrinsic social motivation via causal influence in multi-agent RL. CoRR arxiv: abs/1810.08647,

Jaques N, Lazaridou A, Hughes E, Gulcehre C, Ortega P, Strouse D, Leibo JZ, De Freitas N (2019) Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: International conference on machine learning, pp 3040–3049

Jiang J, Lu Z (2018) Learning attentional communication for multi-agent cooperation. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 7254–7264. http://papers.nips.cc/paper/7956-learning-attentional-communication-for-multi-agent-cooperation.pdf

Johnson M, Hofmann K, Hutton T, Bignell D (2016) The malmo platform for artificial intelligence experimentation. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, AAAI Press, IJCAI'16, pp 4246–4247. http://dl.acm.org/citation.cfm?id=3061053.3061259

Jorge E, Kågebäck M, Gustavsson E (2016) Learning to play guess who? and inventing a grounded language as a consequence. CoRR arxiv: abs/1611.03218,

Juliani A, Berges V, Vckay E, Gao Y, Henry H, Mattar M, Lange D (2018) Unity: a general platform for intelligent agents. CoRR arxiv: abs/1809.02627, Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4(1):237–285. http://dl.acm.org/citation.cfm?id=1622737.1622748

Kasai T, Tenmoto H, Kamiya A (2008) Learning of communication codes in multi-agent reinforcement learning problem. In: 2008 IEEE conference on soft computing in industrial applications, pp 1–6

Kim W, Cho M, Sung Y (2019) Message-dropout: An efficient training method for multi-agent deep reinforcement learning. In: Proceedings of the AAAI conference on artificial intelligence 33(01):6079–6086

Kirby S (2002) Natural language from artificial life. Artif Life 8(2):185–215. https://doi.org/10.1162/106454602320184248

Kok JR, Vlassis N (2006) Collaborative multiagent reinforcement learning by payoff propagation. J Mach Learn Res 7:1789–1828. http://dl.acm.org/citation.cfm?id=1248547.1248612

Kollock P (1998) Social dilemmas: the anatomy of cooperation. Annu Rev Sociol 24(1):183–214. https://doi.org/10.1146/annurev.soc.24.1.183

Kong X, Xin B, Liu F, Wang Y (2017) Revisiting the master-slave architecture in multi-agent deep reinforcement learning. CoRR arxiv: abs/1712.07305,

Kraemer L, Banerjee B (2016) Multi-agent reinforcement learning as a rehearsal for decentralized planning. Neurocomputing 190:82–94

Kumar S, Shah P, Hakkani-Tür D, Heck LP (2017) Federated control with hierarchical multi-agent deep reinforcement learning. CoRR arxiv: abs/1712.08266,

Lanctot M, Zambaldi V, Gruslys A, Lazaridou A, Tuyls K, Perolat J, Silver D, Graepel T (2017) A unified game-theoretic approach to multiagent reinforcement learning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 4190–4203. http://papers.nips.cc/paper/7007-a-unified-game-theoretic-approach-to-multiagent-reinforcement-learning.pdf

Lange PAV, Joireman J, Parks CD, Dijk EV (2013) The psychology of social dilemmas: a review. Organ Behav Hum Decis Process 120(2):125–141

Lauer M, Riedmiller M (2000) An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In: In Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, pp 535–542

Laurent GJ, Matignon L, Fort-Piat NL (2011) The world of independent learners is not markovian. Int J Knowl-Based Intell Eng Syst 15(1):55–64. http://dl.acm.org/citation.cfm?id=1971886.1971887

Lazaridou A, Baroni M (2020) Emergent multi-agent communication in the deep learning era. ArXiv arxiv: abs/2006.02419

Lazaridou A, Peysakhovich A, Baroni M (2017) Multi-agent cooperation and the emergence of (natural) language. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. https://openreview.net/forum?id=Hk8N3Sclg

Lazaridou A, Hermann KM, Tuyls K, Clark S (2018) Emergence of linguistic communication from referential games with symbolic and pixel input. In: International conference on learning representations. https://openreview.net/forum?id=HJGv1Z-AW

Le HM, Yue Y, Carr P, Lucey P (2017) Coordinated multi-agent imitation learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 1995–2003. http://proceedings.mlr.press/v70/le17a.html

Lee J, Cho K, Weston J, Kiela D (2017) Emergent translation in multi-agent communication. CoRR arxiv: abs/1710.06922,

Lee Y, Yang J, Lim JJ (2020) Learning to coordinate manipulation skills via skill behavior diversification. In: International conference on learning representations. https://openreview.net/forum?id=ryxB2lBtvH

Leibo JZ, Zambaldi V, Lanctot M, Marecki J, Graepel T (2017) Multi-agent reinforcement learning in sequential social dilemmas. In: Proceedings of the 16th conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '17, pp 464–473. http://dl.acm.org/citation.cfm?id=3091125.3091194

Leibo JZ, Hughes E, Lanctot M, Graepel T (2019) Autocurricula and the emergence of innovation from social interaction: a manifesto for multi-agent intelligence research. CoRR arxiv: abs/1903.00742,

Lerer A, Peysakhovich A (2017) Maintaining cooperation in complex social dilemmas using deep reinforcement learning. CoRR arxiv: abs/1707.01068,

Letcher A, Foerster J, Balduzzi D, Rocktäschel T, Whiteson S (2019) Stable opponent shaping in differentiable games. In: International conference on learning representations. https://openreview.net/forum?id=SyGjjsC5tQ

Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. Journal of Machine Learning Research 17(1):1334–1373. http://dl.acm.org/citation.cfm?id=2946645.2946684

Lewis M, Yarats D, Dauphin YN, Parikh D, Batra D (2017) Deal or no deal? end-to-end learning for negotiation dialogues. CoRR arxiv: abs/1706.05125,

Li F, Bowling M (2019) Ease-of-teaching and language structure from emergent communication. In: Wallach H, Larochelle H, Beygelzimer A, Alche-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems 32, Curran Associates, Inc., pp 15851–15861. http://papers.nips.cc/paper/9714-ease-of-teaching-and-language-structure-from-emergent-communication.pdf

Li S, Wu Y, Cui X, Dong H, Fang F, Russell S (2019a) Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. Proc AAAI Conf Artif Intell 33(01):4213–4220

Li X, Sun M, Li P (2019b) Multi-agent discussion mechanism for natural language generation. Proc AAAI Conf Artif Intell 33(01):6096–6103

Li Y (2018) Deep reinforcement learning. CoRR arxiv: abs/1810.06339,

Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2016) Continuous control with deep reinforcement learning. In: ICLR (Poster). http://arxiv.org/arxiv: abs/1509.02971

Lin K, Zhao R, Xu Z, Zhou J (2018) Efficient large-scale fleet management via multi-agent deep reinforcement learning. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, ACM, New York, NY, USA, KDD '18, pp 1774–1783. https://doi.org/10.1145/3219819.3219993,

Lin X, Beling PA, Cogill R (2018) Multiagent inverse reinforcement learning for two-person zero-sum games. IEEE Trans Games 10(1):56–68. https://doi.org/10.1109/TCIAIG.2017.2679115

Littman M (2001) Value-function reinforcement learning in markov games. Cogn Syst Res 2:55–66

Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the eleventh international conference on international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML'94, pp 157–163. http://dl.acm.org/citation.cfm?id=3091574.3091594

Liu IJ, Yeh RA, Schwing AG (2020) Pic: Permutation invariant critic for multi-agent deep reinforcement learning. In: PMLR, proceedings of machine learning research, vol 100, pp 590–602. http://proceedings.mlr.press/v100/liu20a.html

Liu S, Lever G, Heess N, Merel J, Tunyasuvunakool S, Graepel T (2019) Emergent coordination through competition. In: International conference on learning representations. https://openreview.net/forum?id=BkG8sjR5Km

Long Q, Zhou Z, Gupta A, Fang F, Wu Y, Wang X (2020) Evolutionary population curriculum for scaling multi-agent reinforcement learning. In: International conference on learning representations. https://openreview.net/forum?id=SJxbHkrKDH

Lowe R, WU Y, Tamar A, Harb J, Pieter Abbeel O, Mordatch I (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 6379–6390. http://papers.nips.cc/paper/7217-multi-agent-actor-critic-for-mixed-cooperative-competitive-environments.pdf

Lowe R, Foerster JN, Boureau Y, Pineau J, Dauphin YN (2019) On the pitfalls of measuring emergent communication. CoRR arxiv: abs/1903.05168,

Luketina J, Nardelli N, Farquhar G, Foerster JN, Andreas J, Grefenstette E, Whiteson S, Rocktäschel T (2019) A survey of reinforcement learning informed by natural language. CoRR arxiv: abs/1906.03926,

Luong NC, Hoang DT, Gong S, Niyato D, Wang P, Liang Y, Kim DI (2019) Applications of deep reinforcement learning in communications and networking: a survey. IEEE Communications Surveys Tutorials pp 1–1. https://doi.org/10.1109/COMST.2019.2916583

Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. Nature 397(6719):498–500. https://doi.org/10.1038/17290

Lyu X, Amato C (2020) Likelihood quantile networks for coordinating multi-agent reinforcement learning. In: Proceedings of the 19th international conference on autonomous agents and multiagent systems, pp 798–806

Ma J, Wu F (2020) Feudal multi-agent deep reinforcement learning for traffic signal control. In: Seghrouchni AEF, Sukthankar G, An B, Yorke-Smith N (eds) Proceedings of the 19th international conference on autonomous agents and multiagent systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020, International Foundation for Autonomous Agents and Multiagent Systems, pp 816–824. https://dl.acm.org/doi/arxiv: abs/10.5555/3398761.3398858

Makar R, Mahadevan S, Ghavamzadeh M (2001) Hierarchical multi-agent reinforcement learning. In: Proceedings of the fifth international conference on autonomous agents, ACM, New York, NY, USA, AGENTS '01, pp 246–253. https://doi.org/10.1145/375735.376302,

Matignon L, Laurent GJ, Le Fort-Piat N (2007) Hysteretic q-learning : an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In: 2007 IEEE/RSJ international conference on intelligent robots and systems, pp 64–69

Matignon L, Jeanpierre L, Mouaddib AI (2012a) Coordinated multi-robot exploration under communication constraints using decentralized markov decision processes. https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5038

Matignon L, Gj Laurent, Le fort piat N, (2012b) Review: independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. Knowl Eng Rev 27(1):1–31. https://doi.org/10.1017/S0269888912000057

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518:529 EP –. https://doi.org/10.1038/nature14236

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd international conference on machine learning, PMLR, New York, New York, USA, Proceedings of machine learning research, vol 48, pp 1928–1937. http://proceedings.mlr.press/v48/mniha16.html

Moerland TM, Broekens J, Jonker CM (2018) Emotion in reinforcement learning agents and robots: a survey. Mach Learn 107(2):443–480. https://doi.org/10.1007/s10994-017-5666-0

Mordatch I, Abbeel P (2018) Emergence of grounded compositional language in multi-agent populations. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17007

Nair R, Tambe M, Yokoo M, Pynadath D, Marsella S (2003) Taming decentralized pomdps: towards efficient policy computation for multiagent settings. In: Proceedings of the 18th international joint conference on artificial intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'03, pp 705–711. http://dl.acm.org/citation.cfm?id=1630659.1630762

Narvekar S, Sinapov J, Leonetti M, Stone P (2016) Source task creation for curriculum learning. In: Proceedings of the 2016 international conference on autonomous agents & multiagent systems, international foundation for autonomous agents and multiagent systems, Richland, SC, AAMAS '16, pp 566–574. http://dl.acm.org/citation.cfm?id=2936924.2937007

Nedic A, Ozdaglar A (2009) Distributed subgradient methods for multi-agent optimization. IEEE Trans Autom Control 54(1):48–61

Ng AY, Russell SJ (2000) Algorithms for inverse reinforcement learning. In: Proceedings of the seventeenth international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '00, pp 663–670. http://dl.acm.org/citation.cfm?id=645529.657801

Ng AY, Harada D, Russell S (1999) Policy invariance under reward transformations: theory and application to reward shaping. In: In Proceedings of the sixteenth international conference on machine learning, Morgan Kaufmann, pp 278–287

Nguyen DT, Kumar A, Lau HC (2017a) Collective multiagent sequential decision making under uncertainty. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14891

Nguyen DT, Kumar A, Lau HC (2017b) Policy gradient with value function approximation for collective multiagent planning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 4319–4329. http://papers.nips.cc/paper/7019-policy-gradient-with-value-function-approximation-for-collective-multiagent-planning.pdf

Nguyen DT, Kumar A, Lau HC (2018) Credit assignment for collective multiagent rl with global rewards. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 8102–8113. http://papers.nips.cc/paper/8033-credit-assignment-for-collective-multiagent-rl-with-global-rewards.pdf

Nguyen TT, Nguyen ND, Nahavandi S (2020) Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. IEEE Trans Cybern 50(9):3826–3839

Oliehoek FA, Amato C (2016) A Concise Introduction to Decentralized POMDPs, 1st edn. Springer Publishing Company, Berlin

Oliehoek FA, Spaan MTJ, Vlassis N (2008) Optimal and approximate q-value functions for decentralized pomdps. J Artif Int Res 32(1):289–353. http://dl.acm.org/citation.cfm?id=1622673.1622680

Omidshafiei S, Pazis J, Amato C, How JP, Vian J (2017) Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of machine learning research, vol 70, pp 2681–2690. http://proceedings.mlr.press/v70/omidshafiei17a.html

Omidshafiei S, Kim DK, Liu M, Tesauro G, Riemer M, Amato C, Campbell M, How JP (2019) Learning to teach in cooperative multiagent reinforcement learning. Proc AAAI Conf Artif Intelli 33(01):6128–6136

Oroojlooyjadid A, Hajinezhad D (2019) A review of cooperative multi-agent deep reinforcement learning. ArXiv arxiv: abs/1908.03963

Oudeyer PY, Kaplan F (2007) What is intrinsic motivation? A typology of computational approaches. Front Neurorobotics 1:6–6

Palmer G, Tuyls K, Bloembergen D, Savani R (2018) Lenient multi-agent deep reinforcement learning. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 443–451. http://dl.acm.org/citation.cfm?id=3237383.3237451

Palmer G, Savani R, Tuyls K (2019) Negative update intervals in deep multi-agent reinforcement learning. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, pp 43–51

Panait L, Luke S (2005) Cooperative multi-agent learning: the state of the art. Auton Agent Multi-Agent Syst 11(3):387–434. https://doi.org/10.1007/s10458-005-2631-2

Panait L, Sullivan K, Luke S (2006) Lenient learners in cooperative multiagent systems. In: Proceedings of the fifth international joint conference on autonomous agents and multiagent systems, association for computing machinery, New York, NY, USA, AAMAS '06, pp 801–803. https://doi.org/10.1145/1160633.1160776,

Papoudakis G, Christianos F, Rahman A, Albrecht SV (2019) Dealing with non-stationarity in multi-agent deep reinforcement learning. CoRR arxiv: abs/1906.04737,

Pathak D, Agrawal P, Efros AA, Darrell T (2017) Curiosity-driven exploration by self-supervised prediction. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 2778–2787. http://proceedings.mlr.press/v70/pathak17a.html

Peng P, Yuan Q, Wen Y, Yang Y, Tang Z, Long H, Wang J (2017) Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games. CoRR arxiv: abs/1703.10069,

Pérolat J, Leibo JZ, Zambaldi V, Beattie C, Tuyls K, Graepel T (2017) A multi-agent reinforcement learning model of common-pool resource appropriation. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 3643–3652. http://papers.nips.cc/paper/6955-a-multi-agent-reinforcement-learning-model-of-common-pool-resource-appropriation.pdf

Peysakhovich A, Lerer A (2018) Prosocial learning agents solve generalized stag hunts better than selfish ones. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, international Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 2043–2044. http://dl.acm.org/citation.cfm?id=3237383.3238065

Pinto L, Davidson J, Sukthankar R, Gupta A (2017) Robust adversarial reinforcement learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of machine learning research, vol 70, pp 2817–2826. http://proceedings.mlr.press/v70/pinto17a.html

Pinyol I, Sabater-Mir J (2013) Computational trust and reputation models for open multi-agent systems: a review. Artif Intell Rev 40(1):1–25. https://doi.org/10.1007/s10462-011-9277-z

Potter MA, De Jong KA (1994) A cooperative coevolutionary approach to function optimization. In: Davidor Y, Schwefel HP, Männer R (eds) Parallel problem solving from nature - PPSN III. Springer, Berlin, pp 249–257

Qu G, Wierman A, Li N (2020) Scalable reinforcement learning of localized policies for multi-agent networked systems. PMLR, The Cloud, Proceedings of machine learning research, vol 120, pp 256–266. http://proceedings.mlr.press/v120/qu20a.html

Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SMA, Botvinick M (2018) Machine theory of mind. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4218–4227. http://proceedings.mlr.press/v80/rabinowitz18a.html

Raghu M, Irpan A, Andreas J, Kleinberg B, Le Q, Kleinberg J (2018) Can deep reinforcement learning solve Erdos-Selfridge-Spencer games? In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4238–4246. http://proceedings.mlr.press/v80/raghu18a.html

Raileanu R, Denton E, Szlam A, Fergus R (2018) Modeling others using oneself in multi-agent reinforcement learning. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4257–4266. http://proceedings.mlr.press/v80/raileanu18a.html

Ramchurn SD, Huynh D, Jennings NR (2004) Trust in multi-agent systems. Knowl Eng Rev 19(1):1–25. https://doi.org/10.1017/S0269888904000116

Rashid T, Samvelyan M, Schroeder C, Farquhar G, Foerster J, Whiteson S (2018) QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 4295–4304. http://proceedings.mlr.press/v80/rashid18a.html

Russell S, Zimdars AL (2003) Q-decomposition for reinforcement learning agents. In: Proceedings of the twentieth international conference on international conference on machine learning, AAAI Press, ICML'03, pp 656–663. http://dl.acm.org/citation.cfm?id=3041838.3041921

Schaul T, Horgan D, Gregor K, Silver D (2015) Universal value function approximators. In: Proceedings of the 32nd international conference on international conference on machine learning - volume 37, JMLR.org, ICML'15, pp 1312–1320

Schmidhuber J (2010) Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Trans Auton Ment Dev 2(3):230–247. https://doi.org/10.1109/TAMD.2010.2056368

Schmidhuber J, Zhao J, Wiering M (1996) Simple principles of metalearning. Tech. rep

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. CoRR arxiv: abs/1707.06347,

Sen S, Weiss G (1999) Multiagent systems. MIT Press, Cambridge, MA, USA. http://dl.acm.org/citation.cfm?id=305606.305612

Sequeira P, Melo FS, Prada R, Paiva A (2011) Emerging social awareness: exploring intrinsic motivation in multiagent learning. In: 2011 IEEE international conference on development and learning (ICDL), vol 2, pp 1–6. https://doi.org/10.1109/DEVLRN.2011.6037325

Shalev-Shwartz S, Shammah S, Shashua A (2016) Safe, multi-agent, reinforcement learning for autonomous driving. CoRR arxiv: abs/1610.03295,

Shapley LS (1953) Stochastic games. Proc Nat Acad Sci 39(10):1095–1100

Shoham Y, Leyton-Brown K (2008) Multiagent systems: algorithmic, game-theoretic, and logical foundations. Cambridge University Press, USA

Shoham Y, Powers R, Grenager T (2003) Multi-agent reinforcement learning: a critical survey. Tech. rep

Silva FLD, Taylor ME, Costa AHR (2018) Autonomously reusing knowledge in multiagent reinforcement learning. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18, International joint conferences on artificial intelligence organization, pp 5487–5493. https://doi.org/10.24963/ijcai.2018/774,

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of go with deep neural networks and tree search. Nature 529:484 EP –. https://doi.org/10.1038/nature16961

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K, Hassabis D (2018) A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science 362(6419):1140–1144

Singh A, Jain T, Sukhbaatar S (2019) Learning when to communicate at scale in multiagent cooperative and competitive tasks. In: International conference on learning representations. https://openreview.net/forum?id=rye7knCqK7

Son K, Kim D, Kang WJ, Hostallero DE, Yi Y (2019) Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In: International conference on machine learning, pp 5887–5896

Song J, Ren H, Sadigh D, Ermon S (2018) Multi-agent generative adversarial imitation learning. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, Curran Associates, Inc., vol 31, pp 7461–7472. https://proceedings.neurips.cc/paper/2018/file/240c945bb72980130446fc2b40fbb8e0-Paper.pdf

Song Y, Wang J, Lukasiewicz T, Xu Z, Xu M, Ding Z, Wu L (2019) Arena: A general evaluation platform and building toolkit for multi-agent intelligence. CoRR arxiv: abs/1905.08085,

Spooner T, Savani R (2020) Robust market making via adversarial reinforcement learning. In: Proceedings of the 19th international conference on autonomous agents and multiagent systems, pp 2014–2016

Srinivasan S, Lanctot M, Zambaldi V, Perolat J, Tuyls K, Munos R, Bowling M (2018) Actor-critic policy optimization in partially observable multiagent environments. In: Bengio S, Wallach H, Larochelle

H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 3422–3435. http://papers.nips.cc/paper/7602-actor-critic-policy-optimization-in-partially-observable-multiagent-environments.pdf

Stone P, Veloso M (2000) Multiagent systems: a survey from a machine learning perspective. Auton Robots 8(3):345–383. https://doi.org/10.1023/A:1008942012299

Strouse D, Kleiman-Weiner M, Tenenbaum J, Botvinick M, Schwab DJ (2018) Learning to share and hide intentions using information regularization. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 10249–10259. http://papers.nips.cc/paper/8227-learning-to-share-and-hide-intentions-using-information-regularization.pdf

Sukhbaatar S, szlam a, Fergus R (2016) Learning multiagent communication with backpropagation. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29, Curran Associates, Inc., pp 2244–2252. http://papers.nips.cc/paper/6398-learning-multiagent-communication-with-backpropagation.pdf

Sukhbaatar S, Kostrikov I, Szlam A, Fergus R (2017) Intrinsic motivation and automatic curricula via asymmetric self-play. CoRR arxiv: abs/1703.05407,

Sunehag P, Lever G, Gruslys A, Czarnecki WM, Zambaldi V, Jaderberg M, Lanctot M, Sonnerat N, Leibo JZ, Tuyls K, Graepel T (2018) 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, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, pp 2085–2087. http://dl.acm.org/citation.cfm?id=3237383.3238080

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Adaptive computation and machine learning, MIT Press. http://www.worldcat.org/oclc/37293240

Sutton RS, Precup D, Singh S (1999) Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. Artif Intell 112(1):181–211

Svetlik M, Leonetti M, Sinapov J, Shah R, Walker N, Stone P (2017) Automatic curriculum graph generation for reinforcement learning agents. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14961

Tacchetti A, Song HF, Mediano PAM, Zambaldi V, Kramár J, Rabinowitz NC, Graepel T, Botvinick M, Battaglia PW (2019) Relational forward models for multi-agent learning. In: International conference on learning representations. https://openreview.net/forum?id=rJlEojAqFm

Tampuu A, Matiisen T, Kodelja D, Kuzovkin I, Korjus K, Aru J, Aru J, Vicente R (2017) Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE 12(4):1–15. https://doi.org/10.1371/journal.pone.0172395

Tan M (1993) Multi-agent reinforcement learning: Independent vs. cooperative agents. In: In Proceedings of the tenth international conference on machine learning, Morgan Kaufmann, pp 330–337

Tang H, Hao J, Lv T, Chen Y, Zhang Z, Jia H, Ren C, Zheng Y, Fan C, Wang L (2018) Hierarchical deep multiagent reinforcement learning. CoRR arxiv: abs/1809.09332,

Taylor A, Dusparic I, Cahill V (2013) Transfer learning in multi-agent systems through parallel transfer. In: in Workshop on theoretically grounded transfer learning at the 30th international conference on machine learning (Poster

Taylor ME, Stone P (2009) Transfer learning for reinforcement learning domains: a survey. J Mach Learn Res 10:1633–1685. http://dl.acm.org/citation.cfm?id=1577069.1755839

Tesauro G (2004) Extending q-learning to general adaptive multi-agent systems. In: Thrun S, Saul LK, Schölkopf B (eds) Advances in neural information processing systems 16, MIT Press, pp 871–878. http://papers.nips.cc/paper/2503-extending-q-learning-to-general-adaptive-multi-agent-systems.pdf

Tumer K, Wolpert DH (2004) Collectives and the design of complex systems. Springer, Berlin

Tuyls K, Weiss G (2012) Multiagent learning: basics, challenges, and prospects. AI Mag 33(3):41

Vezhnevets AS, Osindero S, Schaul T, Heess N, Jaderberg M, Silver D, Kavukcuoglu K (2017) FeUdal networks for hierarchical reinforcement learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 3540–3549. http://proceedings.mlr.press/v70/vezhnevets17a.html

Vezhnevets AS, Wu Y, Leblond R, Leibo JZ (2019) Options as responses: grounding behavioural hierarchies in multi-agent RL. CoRR arxiv: abs/1906.01470,

Vinyals O, Ewalds T, Bartunov S, Georgiev P, Vezhnevets AS, Yeo M, Makhzani A, Küttler H, Agapiou J, Schrittwieser J, Quan J, Gaffney S, Petersen S, Simonyan K, Schaul T, van Hasselt H, Silver D, Lillicrap TP, Calderone K, Keet P, Brunasso A, Lawrence D, Ekermo A, Repp J, Tsing R (2017) Starcraft II: a new challenge for reinforcement learning. CoRR arxiv: abs/1708.04782, Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P, Oh J, Horgan D, Kroiss M, Danihelka I, Huang A, Sifre L, Cai T, Agapiou JP, Jaderberg M, Vezhnevets AS, Leblond R, Pohlen T, Dalibard V, Budden D, Sulsky Y, Molloy J, Paine TL, Gulcehre C, Wang Z, Pfaff T, Wu Y, Ring R, Yogatama D, Wünsch D, McKinney K, Smith O, Schaul T, Lillicrap T, Kavukcuoglu K, Hassabis D, Apps C, Silver D (2019) Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature 575(7782):350–354. https://doi.org/10.1038/s41586-019-1724-z

Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, Blundell C, Kumaran D, Botvinick M (2016a) Learning to reinforcement learn. CoRR arxiv: abs/1611.05763,

Wang JX, Hughes E, Fernando C, Czarnecki WM, Duéñez Guzmán EA, Leibo JZ (2019) Evolving intrinsic motivations for altruistic behavior. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, pp 683–692. http://dl.acm.org/citation.cfm?id=3306127.3331756

Wang S, Wan J, Zhang D, Li D, Zhang C (2016b) Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. Comput Netw 101:158–168. https://doi.org/10.1016/j.comnet.2015.12.017. http://www.sciencedirect.com/science/article/pii/S1389128615005046, industrial Technologies and Applications for the Internet of Things

Wang T, Dong H, Lesser VR, Zhang C (2020a) ROMA: multi-agent reinforcement learning with emergent roles. CoRR arxiv: abs/2003.08039

Wang T, Wang J, Wu Y, Zhang C (2020b) Influence-based multi-agent exploration. In: International conference on learning representations. https://openreview.net/forum?id=BJgy96EYvr

Wang T, Wang J, Zheng C, Zhang C (2020c) Learning nearly decomposable value functions via communication minimization. In: International conference on learning representations. https://openreview.net/forum?id=HJx-3grYDB

Wei E, Luke S (2016) Lenient learning in independent-learner stochastic cooperative games. J Mach Learn Res 17(84):1–42. http://jmlr.org/papers/v17/15-417.html

Wei E, Wicke D, Freelan D, Luke S (2018) Multiagent soft q-learning. https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/view/17508

Wei Ren, Beard RW, Atkins EM (2005) A survey of consensus problems in multi-agent coordination. In: Proceedings of the 2005, American control conference, 2005., pp 1859–1864 vol. 3. https://doi.org/10.1109/ACC.2005.1470239

Weiß G (1995) Distributed reinforcement learning. In: Steels L (ed) The biology and technology of intelligent autonomous agents. Springer, Berlin, pp 415–428

Weiss G (ed) (1999) Multiagent systems: a modern approach to distributed artificial intelligence. MIT Press, Cambridge

Wiegand RP (2004) An analysis of cooperative coevolutionary algorithms. PhD thesis, USA, aAI3108645

Wolpert DH, Tumer K (1999) An introduction to collective intelligence. CoRR cs.LG/9908014. http://arxiv.org/arxiv: abs/cs.LG/9908014

Wu C, Rajeswaran A, Duan Y, Kumar V, Bayen AM, Kakade S, Mordatch I, Abbeel P (2018) Variance reduction for policy gradient with action-dependent factorized baselines. In: International conference on learning representations. https://openreview.net/forum?id=H1tSsb-AW

Yang E, Gu D (2004) Multiagent reinforcement learning for multi-robot systems: a survey. Tech. rep

Yang J, Nakhaei A, Isele D, Fujimura K, Zha H (2020) Cm3: Cooperative multi-goal multi-stage multi-agent reinforcement learning. In: International conference on learning representations. https://openreview.net/forum?id=S1lEX04tPr

Yang T, Meng Z, Hao J, Zhang C, Zheng Y (2018a) Bayes-tomop: a fast detection and best response algorithm towards sophisticated opponents. CoRR arxiv: abs/1809.04240,

Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J (2018b) Mean field multi-agent reinforcement learning. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 5571–5580. http://proceedings.mlr.press/v80/yang18d.html

Yu C, Zhang M, Ren F (2013) Emotional multiagent reinforcement learning in social dilemmas. In: Boella G, Elkind E, Savarimuthu BTR, Dignum F, Purvis MK (eds) PRIMA 2013: principles and practice of multi-agent systems. Springer, Berlin, pp 372–387

Yu H, Shen Z, Leung C, Miao C, Lesser VR (2013) A survey of multi-agent trust management systems. IEEE Access 1:35–50. https://doi.org/10.1109/ACCESS.2013.2259892

Yu L, Song J, Ermon S (2019) Multi-agent adversarial inverse reinforcement learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, PMLR, Long Beach, California, USA, Proceedings of machine learning research, vol 97, pp 7194–7201. http://proceedings.mlr.press/v97/yu19e.html

Zhang K, Yang Z, Basar T (2018) Networked multi-agent reinforcement learning in continuous spaces. In: 2018 IEEE conference on decision and control (CDC), pp 2771–2776

Zhang K, Yang Z, Liu H, Zhang T, Basar T (2018) Fully decentralized multi-agent reinforcement learning with networked agents. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of machine learning research, vol 80, pp 5872–5881. http://proceedings.mlr.press/v80/zhang18n.html

Zhang K, Yang Z, Başar T (2019) Multi-agent reinforcement learning: a selective overview of theories and algorithms. ArXiv arxiv: abs/1911.10635

Zhang W, Bastani O (2019) Mamps: Safe multi-agent reinforcement learning via model predictive shielding. ArXiv arxiv: abs/1910.12639

Zheng Y, Meng Z, Hao J, Zhang Z (2018a) Weighted double deep multiagent reinforcement learning in stochastic cooperative environments. In: Geng X, Kang BH (eds) PRICAI 2018: trends in artificial intelligence. Springer International Publishing, Cham, pp 421–429

Zheng Y, Meng Z, Hao J, Zhang Z, Yang T, Fan C (2018b) A deep bayesian policy reuse approach against non-stationary agents. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 954–964. http://papers.nips.cc/paper/7374-a-deep-bayesian-policy-reuse-approach-against-non-stationary-agents.pdf

Zhu H, Kirley M (2019) Deep multi-agent reinforcement learning in a common-pool resource system. In: 2019 IEEE congress on evolutionary computation (CEC), pp 142–149. https://doi.org/10.1109/CEC.2019.8790001

Zhu Z, Biyik E, Sadigh D (2020) Multi-agent safe planning with gaussian processes. ArXiv arxiv: abs/2008.04452