

Deep Reinforcement Learning for Multi-Agent Systems: A Review of Challenges, Solutions and Applications

Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi

Institute for Intelligent Systems Research and Innovation

Deakin University, Waurn Ponds, Victoria, Australia

E-mail: thanh.nguyen@deakin.edu.au.

Tel: +61 3 52278281. Fax: +61 3 52271046.

Abstract

Reinforcement learning (RL) algorithms have been around for decades and employed to solve various sequential decision-making problems. These algorithms however have faced great challenges when dealing with high-dimensional environments. The recent development of deep learning has enabled RL methods to drive optimal policies for sophisticated and capable agents, which can perform efficiently in these challenging environments. This paper addresses an important aspect of deep RL related to situations that require multiple agents to communicate and cooperate to solve complex tasks. A survey of different approaches to problems related to multi-agent deep RL (MADRL) is presented, including non-stationarity, partial observability, continuous state and action spaces, multi-agent training schemes, multi-agent transfer learning. The merits and demerits of the reviewed methods will be analyzed and discussed, with their corresponding applications explored. It is envisaged that this review provides insights about various MADRL methods and can lead to future development of more robust and highly useful multi-agent learning methods for solving real-world problems.

Keywords: review, survey, deep learning, deep reinforcement learning, robotics, multi-agent, partial observability, non-stationary, continuous action space.

1 Introduction

Reinforcement learning was instigated by a *trial and error* (TE) procedure, conducted by Thorndike in an experiment on cat's behaviors in 1898 [113]. In 1954, Minsky [72] designed the first neural computer named Stochastic Neural-Analog Reinforcement Calculators (SNARCs), which simulated the rat's brain to solve the maze puzzle. SNARCs remarked the uplift of TE learning to a computational period. Almost two decades later, Klopff [51] integrated the mechanism of *temporal-difference* (TD) learning from psychology into the computational model of TE learning. That integration succeeded in making TE learning a feasible approach to large systems. In 1989, Watkins and Dayan [116] brought the theory of optimal control [6] including *Bellman equation* and *Markov decision process* together with temporal-difference learning to form a well-known *Q-learning*. Since then, Q-learning has been applied to solve various real-world problems, but it is unable to solve high-dimensional problems where the number of calculations increases drastically with number of inputs. This problem, known as the *curse of dimensionality*, exceeds the computational constraint of conventional computers. In 2015, Mnih et al. [73] made an important breakthrough by combining *deep learning* with RL to partially overcome the curse of dimensionality. Deep RL has become a normative approach in artificial intelligence, attracting significant attention from the research community since then. Milestones of the development of RL are presented in Fig. 1, which span from the trial and error method to deep RL.

RL originates from animal learning in psychology and thus it can mimic human learning ability to select actions that maximize long-term profit in their interactions with the environment. Therefore, RL can be used to develop an agent that is comparable to the human performance. For instance, RL has been widely used in robotics and autonomous systems, e.g. Mahadevan and Connell [70] designed a robot that can push cubes (1992); Schaal [98] created a humanoid robot that can effectively solve the pole-balancing task (1997); Benbrahim and Franklin [7] made a biped robot that can learn to walk without any knowledge of the

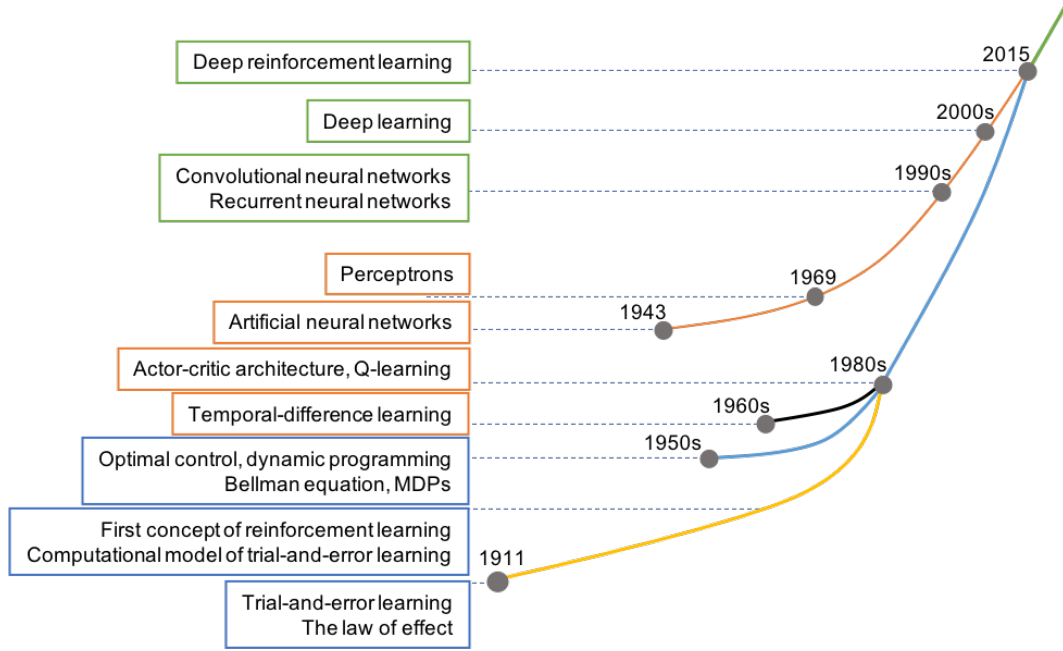


Fig. 1: Emergence of deep RL through different essential milestones.

environment (1997); Riedmiller et al. [95] built a soccer robot team (2009); and Muelling et al. [76] trained a robot to play table tennis (2013).

Modern RL is truly marked by the success of deep RL in 2015 when Mnih et al. [73] made use of a structure named *deep Q-network* (DQN) in creating an agent that outperformed a professional player in a series of 49 classic Atari games [5]. In 2016, Google’s DeepMind created a self-taught AlphaGo program that could beat the best professional Go players, including China’s Ke Jie and Korea’s Lee Sadol [107]. Deep RL has also been used to solve MuJuCo physic problems [19] and 3D maze games [4]. In 2017, OpenAI announced a bot that could beat the best professional gamer on the online game Dota 2, which is supposed to be more complicated than the Go game. These fates provide the necessary impetus to enterprise corporations such as Google, Tesla, and Uber in their race to make self-driving cars. More importantly, deep RL has become a promising approach to solving complex real-world problems and has also been a huge contribution to the field of artificial intelligence.

RL is a research theme that distincts from other related concepts in artificial intelligence. Historically, there had been a confusion between RL and *supervised learning* (SL) since the 1960s. It was not until 1981 that Sutton and Barto [110] shed a light on the discrepancy between the two learning methods. Concisely, SL is learning from data that define input and corresponding output (often called “labelled” data) by an external supervisor, whereas RL is learning by interacting with the unknown environment. In the former case, it maybe infeasible to collect all possible behaviors in the real world to feed the algorithm. However, in the latter case, an RL agent conducts a TE procedure to gain experiences and improve itself over time. Furthermore, RL is not an *unsupervised learning* (UL) method. UL is learning to explore the hidden structure of data where output information is unknown (“unlabelled” data). In contrast, RL is a goal-directed learning, i.e., it constructs a learning model that clearly specifies output to maximize the long-term profit. Finally, deep RL distinguishes with deep learning method. **Deep learning uses multi-layer neural networks to learn a problem in different levels of abstraction [17]. Deep RL leverages deep learning as an approximator to deal with high-dimensional data. This fact makes deep RL a promising approach to solving complex real-world problems.**

As real-world problems have become increasingly complicated, there are many situations where a single deep RL agent is not able to cope with. In such situations, the applications of a *multi-agent system* (MAS) are indispensable. **In an MAS, agents must compete or cooperate to obtain the best overall results.** Examples of such systems include multi-player online games, **cooperative robots in the production factories**, traffic

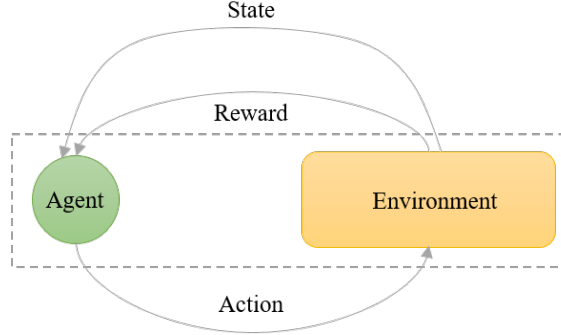


Fig. 2: A single agent interacting with its environment.

control systems, and autonomous military systems like unmanned aerial vehicles, surveillance, and spacecraft. Among many applications of deep RL in the literature, there is a large number of studies using deep RL in MAS, henceforth *multi-agent deep RL* (MADRL). Extending from a single agent domain to a multi-agent environment creates several challenges. Previous surveys considered different perspectives, for example, Busoniu et al. [10] examined the stability and adaptation aspects of agents, Bloembergen et al. [8] analysed the evolutionary dynamics, Hernandez-Leal et al. [41] considered emergent behaviors, communication and cooperation learning perspectives, and Silva et al. [105] reviewed methods for knowledge reuse autonomy in *multi-agent RL* (MARL). This paper presents an overview of technical challenges in multi-agent learning as well as deep RL approaches to these challenges. We cover numerous MADRL perspectives, including non-stationarity, partial observability, multi-agent training schemes, transfer learning in MAS, and continuous state and action spaces in multi-agent learning. Applications of MADRL in various fields are also reviewed and analysed in the current study. In the last section, we present extensive discussions and interesting future research directions of MADRL.

2 Background: Reinforcement Learning

2.1 Preliminary

RL is a TE learning 1) by interacting directly with the environment 2) to self-teach over time and 3) eventually achieve designating goal. Specifically, RL defines any decision maker (learner) as an *agent* and anything outside the agent as an *environment*. The interactions between agent and environment are described via three essential elements: state s , action a , and reward r , as illustrated in Fig. 2 [110]. The state of the environment at time-step t is denoted as s_t . Thereby, the agent examines s_t and performs a corresponding action a_t . The environment then alters its state s_t to s_{t+1} and provides a feedback reward r_{t+1} to the agent. For instance, Fig. 3 illustrates one of the earliest problems in RL literature, a 2D pole-balancing task. In this problem, a state of the environment at time-step t can be presented by a 4-tuple $s_t = [x_c, v_c, \alpha_p, \omega_p]_t$, where x_c denotes *x-coordinate of the cart in Cartesian coordinate system* O_{xy} , v_c presents velocity of the cart along the track, α_p is the angle created by the pole and axis O_y , and ω_p indicates the angular velocity of the pole around center I . The agent can produce two possible actions at each time-step t : exert a unit force ($|\vec{F}| = 1$) to the cart along axis O_x from left to right $a_t = \vec{F}$ or from right to left $a_t = -\vec{F}$. The agent is given a feedback reward $r_{t+1} = +1$ for every action that can keep the pole upright and $r_{t+1} = 0$ otherwise. Therefore, the agent's goal is to keep the pole upright as long as possible and ultimately maximize the accumulated feedback reward.

Typically, the interactions between agent and environment can be presented by a series of states, actions, and rewards: $s_0, a_0, r_1, s_1, a_1, \dots, r_n, s_n$. Although n can approach to infinity, we often limit n in practice by defining a terminal state $s_n = s_T$. In this case, a series of states, actions, and rewards from initial state to terminal state is called an *episode*. For example, in pole-balancing task, we can define a terminal state as if $|\alpha_p| > 10^\circ$ or $|x_c| > X_{max}$.

The next step is to formalize the agent's decision by defining a concept of *policy*. A policy π is a mapping

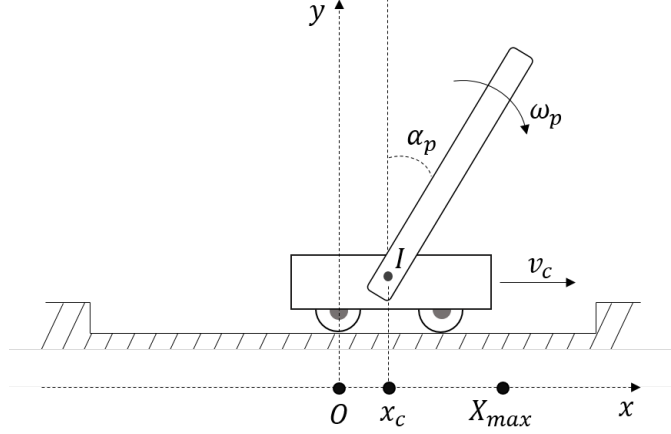


Fig. 3: A 2D pole-balancing task.

function from any **perceived** state s to action a taken from that state. A policy is *deterministic* if the probability of choosing an action a from s : $p(a|s) = 1$ for all state s . In contrast, the policy is *stochastic* if there exists a state s so that $p(a|s) < 1$. **In either case, we can define the policy π as a probability distribution of candidate actions that will be selected from a certain state as below:**

$$\pi = \Psi(s) = \left\{ p(a_i|s) \mid \forall a_i \in \Delta_\pi \wedge \sum_i p(a_i|s) = 1 \right\}, \quad (1)$$

where Δ_π represents all candidate actions (*action space*) of policy π . For clarity, we assume that the action space is discrete because the continuous case can be straightforwardly inferred by using integral notation. Furthermore, we presume that the next state s_{t+1} and feedback reward r_{t+1} are entirely determined by the current state-action pair (s_t, a_t) regardless of the history. **Any RL problem satisfies this “memoryless” condition is known as *Markov decision process* (MDP).** Therefore, the *dynamics* (model) of an RL problem is completely specified by giving all *transition probabilities* $p(a_i|s)$. Based on (1), we can define a deterministic policy π_d :

$$\pi_d = \Psi_d(s) = \begin{cases} 1, & a_i = a(s) \wedge a(s) \in \Delta_{\pi_d} \\ 0, & \forall a_i \in \Delta_{\pi_d} \wedge a_i \neq a(s) \end{cases},$$

where $a(s)$ denotes the designating action taken at state s . Deterministic policy is desirable in practical application because it has a predictable behavior, which is a crucial factor for designing an effective RL algorithm. In practice, we can derive a deterministic policy π_d from a stochastic policy π using the following rule:

$$\mathbf{R}_1 : \pi \mapsto \pi_d = \Psi_d(s) = \begin{cases} 1, & a_i = a_j \wedge j = \arg \max_k \pi(s, a_k) \\ 0, & \forall a_i \in \Delta_\pi \wedge a_i \neq a_j \end{cases}, \quad (2)$$

where $\pi(s, a_k)$ denotes the probability of taking action $a_k \in \Delta_\pi$ in state s using policy π and $\Delta_{\pi_d} = \Delta_\pi$.

Initially, the agent is assigned a random policy π_0 . It adjusts the policy π_0 to improve itself by interacting with the environment in a TE learning manner. In this respect, we call that policy π_{t+1} is better than policy π_t and denoted as $\pi_{t+1} > \pi_t$. Therefore, we have a series of policies improved over time as follows:

$$\pi_0 < \pi_1 < \dots < \pi_t < \pi_{t+1} < \dots < \pi^*.$$

This process, named *policy improvement*, is repeated until the agent cannot find any policy better than the *optimal policy* π^* . By this definition, however, we still do not know exactly how to compare two policies and decide which one is better. In the next subsection, we will review other metrics that can be used to evaluate a policy and then we can use these metrics to compare how “good” between different policies.

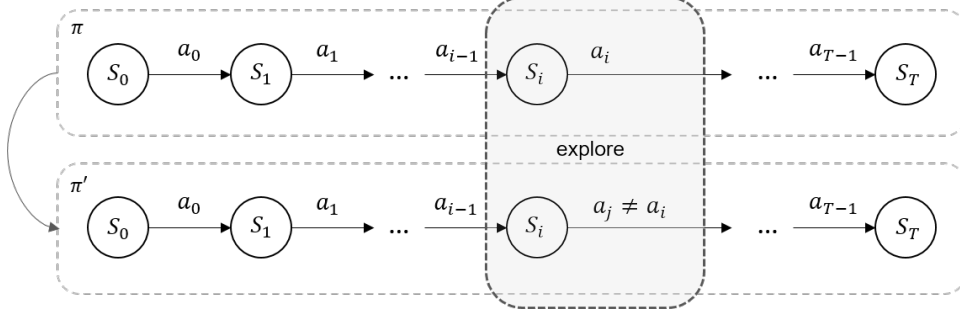


Fig. 4: A naive approach to finding a better policy π' from π .

2.2 Bellman equation

Remind that the agent receives a feedback reward r_{t+1} for every time-step t until it reaches the terminal state s_T . However, the immediate reward r_{t+1} does not represent the long-term profit, we instead leverage a generalized *return value* R_t at time-step t :

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots + \gamma^{T-t-1} r_T = \sum_{i=0}^{T-t-1} \gamma^i r_{t+i+1}, \quad (3)$$

where γ is a discounted factor so that $0 \leq \gamma < 1$. The agent becomes farsighted when γ approaches to 1 and vice versa the agent becomes shortsighted when γ is close to 0. Apparently, we often select γ closing to 1 in practice.

The next step is to define a *value function* that is used to evaluate how “good” of a certain state s or a certain state-action pair (s, a) . Specifically, the value function of state s under policy π is calculated by obtaining expected return value from s [110]: $V_\pi(s) = \mathbb{E}[R_t | s_t = s, \pi]$. Likewise, the value function of state-action pair (s, a) is $Q_\pi(s, a) = \mathbb{E}[R_t | s_t = s, a_t = a, \pi]$. We can leverage value functions to compare how “good” between two policies π and π' using the following rule [110]:

$$\pi \leq \pi' \iff \left[V_\pi(s) \leq V_{\pi'}(s) \ \forall s \right] \vee \left[Q_\pi(s, a) \leq Q_{\pi'}(s, a) \ \forall (s, a) \right]. \quad (4)$$

Based on (3), we can expand $V_\pi(s)$ and $Q_\pi(s, a)$ to present the relationship between two consecutive states $s = s_t$ and $s' = s_{t+1}$ as below [110]:

$$V_\pi(s) = \sum_a \pi(s, a) \sum_{s'} p(s' | s, a) \left(\mathbb{W}_{s \rightarrow s' | a} + \gamma V_\pi(s') \right), \quad (5)$$

and

$$Q_\pi(s, a) = \sum_{s'} p(s' | s, a) \left(\mathbb{W}_{s \rightarrow s' | a} + \gamma V_\pi(s') \right), \quad (6)$$

where $\mathbb{W}_{s \rightarrow s' | a} = \mathbb{E}[r_{t+1} | s_t = s, a_t = a, s_{t+1} = s']$. Solving (5) or (6), we can find value function $V(s)$ or $Q(s, a)$, respectively. Equations (5) and (6) are called Bellman equations and widely used in policy improvement. For example, Fig. 4 describes a naive approach to improve a deterministic policy π . By using a TE approach, we can “explore” a different action $a_j \neq a_i$ taken at state s_i so that $Q_\pi(s_i, a_j) > Q_\pi(s_i, a_i)$. We include a_j as a new action taken at s_i in a derived policy π' while keeping other pairs of state-action unchanged. Based on (4) and (6), we can infer that $\pi < \pi'$. Therefore, it is straightforward to select a “greedy” action a_j so that $Q_\pi(s_i, a_j)$ attains maximum values. Finally, we can derive an improved policy π' from π using the following rule:

$$\mathbf{R}_2 : \pi \mapsto \pi' = \Psi'(s) = \begin{cases} 1, & a_i = a_j \wedge j = \arg \max_k Q_\pi(s, a_k) \\ 0, & \forall a_i \in \Delta_\pi \wedge a_i \neq a_j \end{cases}. \quad (7)$$

This process is iterated for all pairs of (s_i, a_i) until we find an optimal solution π^* . The idea indeed can be generalized to any stochastic policy π .

Instead of repeating policy improvement process, we can estimate directly the value function of optimal policy π^* using the following *optimality Bellman equation* [110]:

$$V_{\pi^*}(s) = \max_a \sum_{s'} p(s'|s, a) \left(\mathbb{W}_{s \rightarrow s'|a} + \gamma V_{\pi^*}(s') \right), \quad (8)$$

and

$$Q_{\pi^*}(s, a) = \sum_{s'} p(s'|s, a) \left(\mathbb{W}_{s \rightarrow s'|a} + \gamma \max_{a'} Q_{\pi^*}(s', a') \right). \quad (9)$$

After solving optimality Bellman equations (8) or (9), we can derive an optimal deterministic policy π^* using (7).

Although we can use *dynamic programming* to approximate the solutions of Bellman equations, it requires the complete dynamics information of the problem. Therefore, it is infeasible due to the lack of memory and computational power of conventional computer when the number of states is large. In the next subsection, we will review two *model-free* RL methods (require no knowledge of transition probabilities $p(a_i|s)$) to approximate the value functions.

2.3 RL methods

In this section, we review two well-known learning schemes in RL: Monte-Carlo and temporal-difference learning. These methods do not require the dynamics information of the environment, i.e., they can deal with larger state-space problems than trivial dynamic programming approaches.

2.3.1 Monte-Carlo method

Monte-Carlo (MC) method estimates value function by repeatedly generating episodes and recording average return at each state or each state-action pair. Therefore, the state-value function is calculated as follows:

$$V_{\pi}^{MC}(s) = \lim_{i \rightarrow +\infty} \mathbb{E} \left[r^i(s_t) \mid s_t = s, \pi \right],$$

where $r^i(s_t)$ denotes observed return at state s_t in episode i th. Similarly, we have value function of state-action pair:

$$Q_{\pi}^{MC}(s, a) = \lim_{i \rightarrow +\infty} \mathbb{E} \left[r^i(s_t, a_t) \mid s_t = s, a_t = a, \pi \right].$$

MC method does not require any knowledge of transition probabilities, i.e., MC method is model-free. However, this approach has made two essential assumptions to ensure the convergence happens: 1) the number of episodes is large and 2) every state and every action must be visited with a significant number of times. To make this “exploration” possible, we often use ϵ -greedy strategy in policy improvement instead of (7):

$$\mathbf{R}_3 : \pi \mapsto \pi' = \Psi(s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\Delta_{\pi}(s)|}, & a_i = a_j \wedge j = \arg \max_k Q_{\pi}(s, a_k) \\ \frac{\epsilon}{|\Delta_{\pi}(s)|}, & \forall a_i \in \Delta_{\pi} \wedge a_i \neq a_j \end{cases}, \quad (10)$$

where $|\Delta_{\pi}(s)|$ denotes number of candidate actions taken in state s and $0 < \epsilon < 1$. Generally, MC algorithms are divided into two groups: *on-policy* and *off-policy*. In on-policy methods, we use policy π for both evaluation and exploration purpose. Therefore, the policy π must be stochastic or *soft*. In contrast, off-policy uses different policy $\pi' \neq \pi$ to generate the episodes and hence π can be deterministic. Although off-policy is desirable due to its simplicity, on-policy method is more stable when working with continuous state-space problems and when using together with a function approximator (such as neural networks) [114].

Table 1: Characteristics of RL methods

Category	Pros	Cons
Model-free	<ul style="list-style-type: none"> • Dynamics of environment is unknown • Deal with larger state-space environments 	<ul style="list-style-type: none"> ◦ Requires “exploration” condition
On-policy	<ul style="list-style-type: none"> • Stable when using with function approximator • Suitable with continuous state-space problems 	<ul style="list-style-type: none"> ◦ Policy must be stochastic
Off-policy	<ul style="list-style-type: none"> • Simplify algorithm design • Can tackle with different kinds of problems • Policy can be deterministic 	<ul style="list-style-type: none"> ◦ Unstable when using with function approximator
Bootstrapping	<ul style="list-style-type: none"> • Learn faster in most cases 	<ul style="list-style-type: none"> ◦ Not as good as nonbootstrapping methods on mean square error

2.3.2 Temporal-difference method

Similar to MC, *temporal-difference* (TD) learning is also learning from experiences (model-free method). However, unlike MC, TD learning does not wait until the end of episode to make an update. It makes an update on every step within the episode by leveraging 1-step Bellman equation (5) and hence possibly providing faster convergence:

$$\mathbf{U}_1 : V^i(s_t) \leftarrow \alpha V^{i-1}(s_t) + (1 - \alpha) \left(r_{t+1} + \gamma V^{i-1}(s_{t+1}) \right),$$

where α is step-size parameter and $0 < \alpha < 1$. TD learning uses previous estimated values V^{i-1} to update the current ones V^i , which is known as *bootstrapping* method. Basically, bootstrapping method learns faster than non-bootstrapping ones in most of the cases [110]. TD learning is also divided into two categories: on-policy TD control (*Sarsa*) and off-policy TD control (*Q-learning*). In *Sarsa*, the algorithm estimates value function of state-action pair based on (6):

$$\mathbf{U}_2 : Q^i(s_t, a_t) \leftarrow \alpha Q^{i-1}(s_t, a_t) + (1 - \alpha) \left(r_{t+1} + \gamma Q^{i-1}(s_{t+1}, a_{t+1}) \right).$$

On the other hand, *Q-learning* uses 1-step optimality Bellman equation (9) to perform the update, i.e., *Q-learning* directly approximates value function of optimal policy:

$$\mathbf{U}_3 : Q^i(s_t, a_t) \leftarrow \alpha Q^{i-1}(s_t, a_t) + (1 - \alpha) \left(r_{t+1} + \gamma \overbrace{\max_{a_{t+1}^j} Q^{i-1}(s_{t+1}, a_{t+1}^j)}^{\text{deterministic subpolicy}} \right). \quad (11)$$

We notice that the operator *max* in update rule (11) substitutes for a deterministic policy. This strongly explains why *Q-learning* is off-policy.

In practice, MC and TD learning often use table memory structure (*tabular method*) to save value function of each state or each state-action pair. This makes them inefficient due to lack of memory in solving complicated problems where number of states is large. Therefore, *actor-critic* (AC) architecture is designed to subdue this limitation. Specifically, AC includes two separate memory structures for an agent: *actor* and *critic*. Actor structure is used to select a suitable action according to the observed state and transfer to critic structure for evaluation. Critic structure uses the following TD error to decide future tendency of the selected action:

$$\delta(a_t) = \beta \left(r_{t+1} + \gamma V(s_{t+1}) \right) - (1 - \beta) V(s_t),$$

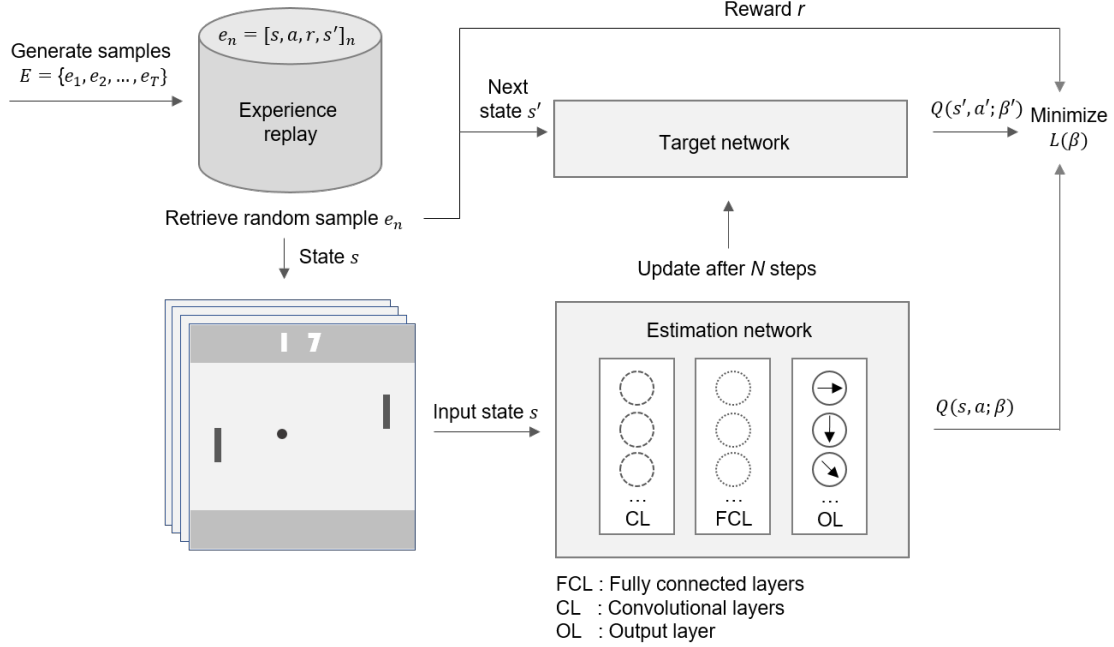


Fig. 5: Deep Q-network structure.

where $0 < \beta < 1$; and if $\delta(a_t) > 0$, the tendency to select the action a_t in the future is high and vice versa. Furthermore, AC can be on-policy or off-policy depending on the implementation details. Table 1 and Table 2 summarize characteristics of RL methods and the comparisons between different RL methods, respectively.

Table 2: Comparisons between RL methods

Category	Dynamic programming	On-policy MC	Off-policy MC	Sarsa	Q-learning	AC
Model-free		✓	✓	✓	✓	✓
On-policy		✓		✓		✓
Off-policy			✓		✓	✓
Bootstrapping	✓			✓	✓	✓

3 Deep RL: Single Agent

3.1 Deep Q-Network

Deep RL is a broad term that indicates the combination between deep learning and RL to deal with high-dimensional environments [3, 65, 79]. In 2015, Mnih et al. [73] at the first time announced the success of this combination by creating an autonomous agent that can play competently a series of 49 Atari games. Concisely, the authors proposed a novel structure named *deep Q-network* (DQN) that leverages the *convolutional neural network* (CNN) [56] to directly interpret graphical representation of input state s from the environment. The output of DQN produces Q-values of all possible actions $a \in \Delta_\tau$ taken at state s , where Δ_τ denotes action space [80]. Therefore, DQN can be seen as a policy network τ , parameterized by β , which is continually trained so as to approximate optimal policy. Mathematically, DQN uses Bellman equation to minimize the loss function $\mathcal{L}(\beta)$ as below:

$$\mathcal{L}(\beta) = \mathbb{E} \left[\left(r + \gamma \max_{a'} Q(s', a' | \beta) - Q(s, a | \beta) \right)^2 \right]. \quad (12)$$

However, using neural network to approximate value function is proved to be unstable and may result

in divergence due to the bias originated from correlative samples [114]. To make the samples uncorrelated, Mnih et al. [73] created a *target network* τ' , parameterized by β' , which is updated in every N steps from estimation network τ . Moreover, generated samples are stored in an *experience replay memory*. Samples are then retrieved randomly from the experience replay and fed into the training process, as described in Fig. 5. Therefore, equation (12) can be rewritten as:

$$\begin{cases} \mathcal{L}_{\text{DQN}}(\beta) = \mathbb{E} \left[(r + \gamma \max_{a'} Q(s', a' | \beta') - Q(s, a | \beta))^2 \right] \\ \beta' \leftarrow \beta \text{ for every } N \text{ steps} \end{cases} \quad (13)$$

Although DQN basically solved a challenging problem in RL, the curse of dimensionality, this is just a rudimental step in solving completely real-world applications. DQN has numerous drawbacks, which can be remedied by different schemes, from a simple form to complicated modifications that we will discuss in the next subsection.

3.2 DQN variants

The first and simplest form of DQN’s variant is *double deep Q-network* (DDQN) proposed by Hasselt [34, 35]. The idea of DDQN is to separate the selection of “greedy” action from action evaluation. In this way, DDQN expects to reduce the overestimation of Q-values in the training process. In other words, the *max* operator in equation (13) is decoupled into two different operators, as represented by the following loss function:

$$\mathcal{L}_{\text{DDQN}}(\beta) = \mathbb{E} \left[\left(r + \gamma Q(s', \arg \max_{a'} Q(s', a' | \beta) | \beta') - Q(s, a | \beta) \right)^2 \right]. \quad (14)$$

Empirical experiment results on 57 Atari games show that the normalized performance of DDQN without tuning is two times greater than DQN and three times greater than DQN when tuning.

Secondly, experience replay in DQN plays an important role to break the correlations between samples, and at the same time, remind “rare” samples that the policy network may rapidly forget. However, the fact that selecting randomly samples from experience replay does not completely separate sample data. Specifically, we prefer rare and goal-related samples to appear more frequent than redundancy ones. Therefore, Schaul et al. [99] proposed a *prioritized experience replay* that gives priority to a sample i based on its absolute value of TD error:

$$p_i = |\delta_i| = |r_i + \gamma \max_a Q(s_i, a | \beta') - Q(s_{i-1}, a_{i-1} | \beta)| \quad (15)$$

Prioritized experience replay when combining with DDQN provides stable convergence of policy network and achieves a performance up to five times greater than DQN in terms of normalized mean score on 57 Atari games.

DQN’s policy evaluation process is struggle to work in “redundant” situations, i.e., there are more than two candidate actions that can be selected without getting any negative results. For instance, when driving a car and there are no obstacles ahead, we can follow either the left lane or the right lane. If there is an obstacle ahead in the left lane, we must be in the right lane to avoid crashing. Therefore, it is more efficient if we focus only on the road and obstacles ahead. To resolve such situations, Wang et al. [115] proposed a novel network architecture named *dueling network*. In dueling architecture, there are two collateral networks that coexist: one network, parameterized by θ , estimates state-value function $V(s | \theta)$ and the other one, parameterized by θ' , estimates advantage action function $A(s, a | \theta')$. The two networks are then aggregated using the following equation to approximate Q-value function:

$$Q(s, a | \theta, \theta') = V(s | \theta) + \left(A(s, a | \theta') - \frac{1}{|\Delta_\pi|} \sum_{a'} A(s, a' | \theta') \right).$$

Because dueling network represents action-value function, it was combined with DDQN and prioritized experience replay to boost the performance of the agent up to six times more than standard DQN on the Atari domain [115].

Another drawback of DQN is that it uses a history of four frames as an input to policy network. DQN is therefore inefficient to solve problems where the current state depends on a significant amount of history

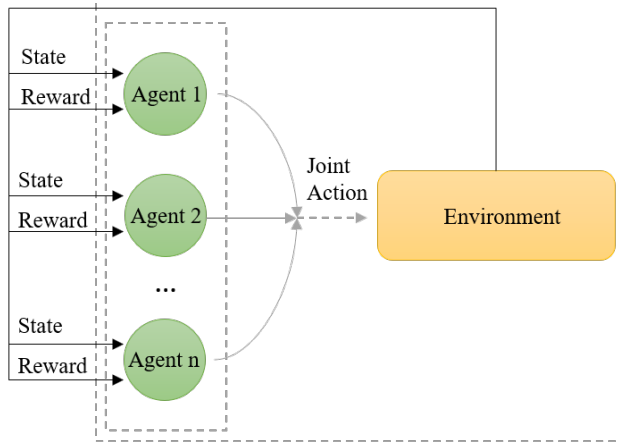


Fig. 6: Multiple agents interacting with the same environment.

information such as Double Dunk or Frostbite. These games are often called partially observable MDP problems. The straightforward solution is to replace the fully-connected layer right after the last convolutional layer of the policy network with a recurrent *long short term memory*, as described in [36]. This DQN’s variant named *deep recurrent Q-network* (DRQN) outperforms standard DQN up to 700 percent in games Double Dunk and Frostbite. Furthermore, Lample and Chaplot [60] successfully created an agent that beats an average player on Doom, a 3D FPS (first-person shooter) environment by adding a game feature layer in DRQN. Another interesting variant of DRQN is *deep attention recurrent Q-network* (DARQN) [108]. In that paper, Sorokin et al. add attention mechanism into DRQN so that the network can focus only on important regions in the game, allowing smaller network’s parameters and hence speeding the training process. As a result, DARQN achieves a score of 7263 compared with 1284 and 1421 of DQN and DRQN on game Seaquest, respectively.

4 Deep RL: Multi-Agent

MASs have attracted great attention because they are able to solve complex tasks through the cooperation of individual agents. Within a MAS, agents communicate with each other and interact with the environment (Fig. 6). In a multi-agent learning domain, the MDP is generalized to a stochastic game, or a Markov game. Let denote n as the number of agents, S as a discrete set of environmental states, and $A_i, i = 1, 2, \dots, n$ as a set of actions for each agent. The joint action set for all agents is defined by $A = A_1 \times A_2 \times \dots \times A_n$. The state transition probability function is represented by $p : S \times A \times S \rightarrow [0, 1]$ and the reward function is specified as $r : S \times A \times S \rightarrow \mathbb{R}^n$. The value function of each agent is dependent on the joint action and joint policy, which is characterized by $V^\pi : S \times A \rightarrow \mathbb{R}^n$. The following subsections describe challenges and MADRL solutions as well as their applications to solve real-world problems.

4.1 MADRL: Challenges and Solutions

4.1.1 Non-stationarity

Controlling multiple agents poses several additional challenges as compared to single agent setting such as the *heterogeneity* of agents, how to define suitable collective goals or the scalability to large number of agents that requires design of compact representations, and more importantly the non-stationarity problem. In a single agent environment, an agent is concerning only the outcome of its own actions. In a multi-agent domain, an agent observes not only the outcomes of its own action but also the behavior of other agents. Learning among the agents is complex because all agents potentially interact with each other and learn concurrently. The interactions among multiple agents constantly reshape the environment and lead to non-stationarity. In this case, learning among the agents sometimes causes changes in the policy of an agent, and can affect the optimal policy of other agents. The estimated potential rewards of an action would be

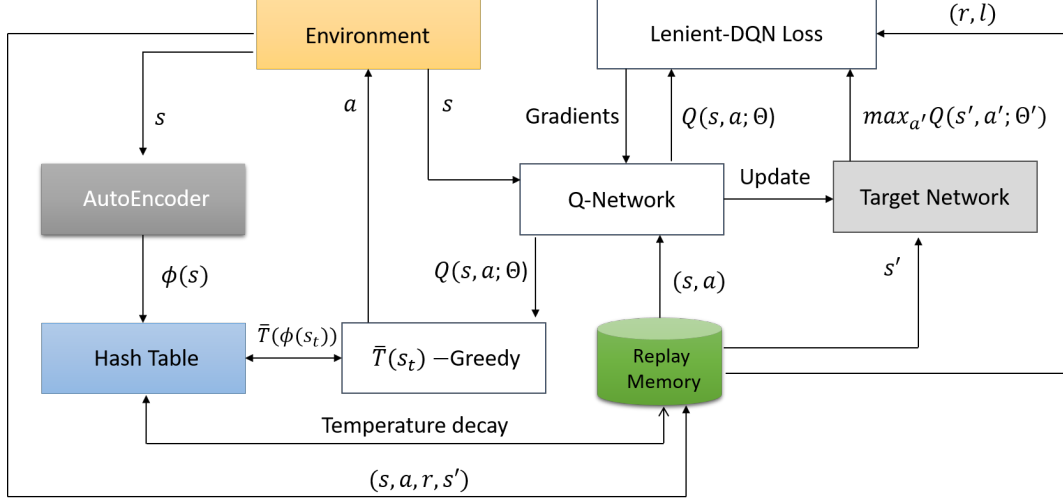


Fig. 7: Architecture of LDQN.

inaccurate and therefore, good policies at a given point in the multi-agent setting could not remain so in the future. The convergence theory of Q-learning applied in single agent setting is not guaranteed to most multi-agent problems as the Markov property does not hold anymore in the non-stationary environment [40]. Therefore, collecting and processing information must be performed with certain recurrence while ensuring that it does not affect the agents' stability. The exploration-exploitation dilemma could be more involved under multi-agent settings.

The popular *independent Q-learning* [112] or experience replay based DQN [73] was not designed for the non-stationary environments. Castaneda [12] proposed two variants of DQN, namely deep repeated update Q-network (DRUQN) and deep loosely coupled Q-network (DLCQN), to deal with the non-stationarity problem in MAS. The DRUQN is developed based on the repeated update Q-learning (RUQL) model introduced in [1, 2]. It aims to avoid policy bias by updating the action value inversely proportional to the likelihood of selecting an action. On the other hand, DLCQN relies on the loosely coupled Q-learning proposed in [119], which specifies and adjusts an independence degree for each agent using its negative rewards and observations. Through this independence degree, the agent learns to decide whether it needs to act independently or cooperate with other agents in different circumstances. Likewise, Diallo et al. [18] extended DQN to a *multi-agent concurrent DQN* and demonstrated that this method can converge in a non-stationary environment. Foerster et al. [25] alternatively introduced two methods for stabilising experience replay of DQN in MADRL. The first method uses the importance sampling approach to naturally decay obsolete data whilst the second method disambiguates the age of the samples retrieved from the replay memory using a fingerprint.

Recently, to deal with non-stationarity due to concurrent learning of multiple agents in MAS, Palmer et al. [86] presented a method namely *lenient-DQN* (LDQN) that applies leniency with decaying temperature values to adjust policy updates sampled from the experience replay memory (Fig. 7). That method is applied to the coordinated multi-agent object transportation problems and its performance is compared with the *hysteretic-DQN* (HDQN) [85]. The experimental results demonstrate the superiority of LDQN against HDQN in terms of convergence to optimal policies in a stochastic reward environment. The notion of leniency along with a scheduled replay strategy were also incorporated into the *weighted double deep Q-network* (WDDQN) in [120] to deal with non-stationarity in MAS. Experiments show the better performance of WDDQN against double DQN in two multi-agent environments with stochastic rewards and large state space.

4.1.2 Partial observability

In real-world applications, there are many circumstances where agents only have partial observability of the environment. In other words, complete information of states pertaining to the environment is not known

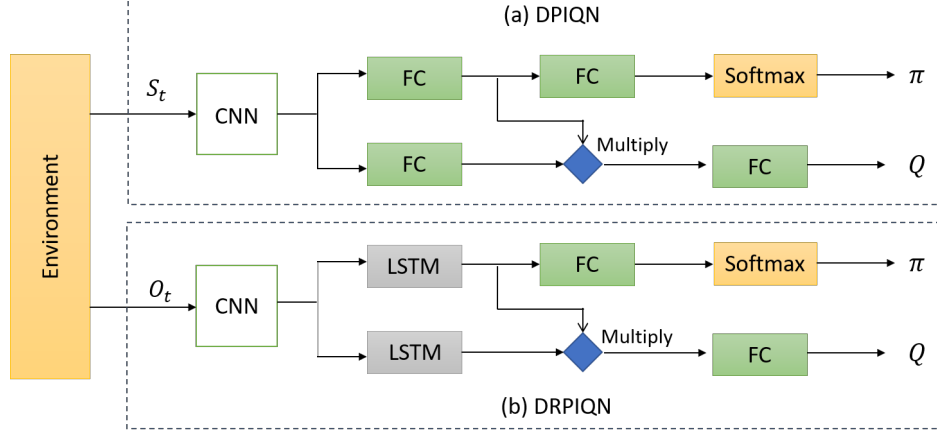


Fig. 8: Architecture of DPIQN and DRPIQN.

to the agents when they interact with the environment. In such situations, the agents observe partial information about the environment, and need to make the best decision during each time step. This type of problem can be modelled using the *partially observable Markov decision process* (POMDP).

In the current literature, a number of deep RL models have been proposed to handle POMDP. Hausknecht and Stone [36] proposed *deep recurrent Q-network* (DRQN) based on a long short term memory network. With the recurrent structure, the DRQN-based agents are able to learn the improved policy in a robust sense in the partially observable environment. Unlike DQN, DRQN approximates $Q(o, a)$, which is a Q-function with observation o and action a , by a recurrent neural network. DRQN treats a hidden state of the network h_{t-1} as an internal state. The DRQN therefore is characterized by the Q-function $(o_t, h_{t-1}, a; \theta_i)$ where θ_i is the parameters of the network at the i th training step. In [24], DRQN is extended to *deep distributed recurrent Q-network* (DDRQN) to handle multi-agent POMDP problems. The success of DDRQN is relied on three notable features, i.e. last-action inputs, inter-agent weight sharing, and disabling experience replay. The first feature, i.e. last-action inputs, requires the provision of previous action of each agent as input to its next step. The inter-agent weight sharing means that all agents use weights of only one network, which is learned during the training process. The disabling experience replay simply excludes the experience replay feature of DQN. DDRQN therefore learns a Q-function of the form $Q(o_t^m, h_{t-1}^m, m, a_{t-1}^m, a_t^m; \theta_i)$ where each agent receives its own index m as input. Weight sharing decreases learning time because it reduces the number of parameters to be learned. Although each agent has different observation and hidden state, this approach however assumes that agents have the same set of actions. To address complicated problems, autonomous agents often have different sets of actions. For example, UAVs manoeuvre in the air whilst robots operate on the ground. Therefore, action spaces of UAVs and robots are different and thus the inter-agent weight sharing feature cannot be applied.

Scaling to a system of many agents in partial observable domains is a challenging problem. Gupta et al. [30] extended the *curriculum learning* technique to an MAS, which integrates with three classes of deep RL, including policy gradient, temporal-difference error, and actor-critic methods. The curriculum principle is to start learning to complete simple tasks first to accumulate knowledge before proceeding to perform complicated tasks. This is suitable with an MAS environment where fewer agents initially collaborate before extending to accommodate more and more agents to complete increasingly difficult tasks. Experimental results show the vitality of the curriculum learning method in scaling deep RL algorithms to complex multi-agent problems.

Hong et al. [42] introduced a *deep policy inference Q-network* (DPIQN) to model multi-agent systems and its enhanced version *deep recurrent policy inference Q-network* (DRPIQN) to cope with partial observability. Both DPIQN and DRPIQN are learned by adapting network's attention to policy features and their own Q-values at various stages of the training process (Fig. 8). Experiments show the better overall performance of both DPIQN and DRPIQN over the baseline DQN and DRQN [36]. Also in the context of partial observability, but extended to multi-task, multi-agent problems, Omidshafiei et al. [85] proposed a

method called *multi-task multi-agent RL* (MT-MARL) that integrates hysteretic learners [71], DRQNs [36], distillation [96], and concurrent experience replay trajectories (CERTs), which are a decentralized extension of experience replay strategy proposed in [73]. The agents are not explicitly provided with task identity (thus partial observability) whilst they cooperatively learn to complete a set of *decentralized POMDP* tasks with sparse rewards. This method however has a disadvantage that cannot perform in an environment with heterogeneous agents.

Apart from partial observability, there are circumstances that agents must deal with extremely noisy observations, which are weakly correlated with the true state of the environment. Kilinc and Montana [48] introduced a method denoted as MADDPG-M that combines DDPG and a communication medium to address these circumstances. Agents need to decide whether their observations are informative to share with other agents and the communication policies are learned concurrently with the main policies through experience. Recently, Foerster et al. [26] proposed *Bayesian action decoder* (BAD) algorithm for learning multiple agents with cooperative partial observable settings. A new concept, namely public belief MDP, is introduced based on BAD that employs an approximate Bayesian update to attain a public belief with publicly observable features in the environment. BAD relies on a factorised and approximate belief state to discover conventions to enable agents to learn optimal policies efficiently. This is closely relevant to *theory of mind* that humans normally use to interpret others' actions. Experimental results on a proof-of-principle two-step matrix game and the cooperative partial-information card game Hanabi demonstrate the efficiency and superiority of the proposed method against traditional policy gradient algorithms.

4.1.3 MAS training schemes

The direct extension of single agent deep RL to multi-agent environment is to learn each agent independently by considering other agents as part of the environment as the independent Q-learning algorithm proposed in [111]. This method is vulnerable to overfitting [61] and computationally expensive, and therefore the number of agents involved is limited. An alternative and popular approach is the *centralized learning and decentralized execution* where a group of agents can be trained simultaneously by applying a centralized method via an open communication channel [55]. Decentralized policies where each agent can take actions based on its local observations have an advantage under partial observability and in limited communications during execution. Centralized learning of decentralized policies has become a standard paradigm in multi-agent settings because the learning process may happen in a simulator and a laboratory where there are no communication constraints, and extra state information is available [55, 27].

Gupta et al. [30] examined three different training schemes for a MAS, which consists of centralized learning, concurrent learning and parameter sharing. Centralized policy attempts to obtain a joint action from joint observations of all agents whilst the concurrent learning trains agents simultaneously using the joint reward signal. In the latter, each agent learns its own policy independently based on private observation. Alternatively, the parameter sharing scheme allows agents to be trained simultaneously using the experiences of all agents although each agent can obtain unique observations. With the ability to execute decentralized policies, parameter sharing can be used to extend a single agent deep RL algorithm to accommodate a system of many agents. Particularly, the combination of *parameter sharing and TRPO*, namely PS-TRPO, has been proposed in [30], and briefly summarized in Algorithm 1. The PS-TRPO has demonstrated great performance when dealing with high-dimensional observations and continuous action spaces under partial observability.

Foerster et al. [23] introduced *reinforced inter-agent learning* (RIAL) and *differentiable inter-agent learning* (DIAL) methods based on the centralized learning approach to improve agent learning communication. In RIAL, deep Q-learning has a recurrent structure to address the partial observability issue, in which independent Q-learning offers individual agents to learn their own network parameters. DIAL pushes gradients from one agent to another through a channel, allowing end-to-end backpropagation across agents. Likewise, Sukhbaatar et al. [109] developed communication neural net (CommNet) allowing dynamic agents to learn continuous communication alongside their policy for fully cooperative tasks. Unlike CommNet, He et al. [38] proposed a method namely *deep reinforcement opponent network* (DRON) that encodes observation of the opponent agent into DQN to jointly learn a policy and behaviors of opponents without domain knowledge.

Kong et al. [53, 54] incorporated both decentralized and centralized perspectives into the hierarchical master-slave architecture to form a model named *master-slave multi-agent RL* (MS-MARL) to solve the

Algorithm 1 PS-TRPO

- 1: Initialize parameters of policy network Θ_0 , and trust region size Δ
 - 2: **for** $i \leftarrow 0, 1, \dots$ **do**
 - 3: Generate trajectories for all agents as $\tau \sim \pi_{\theta_i}$ using the policy with shared parameters.
 - 4: For each agent m , compute the advantage values $A_{\pi_{\theta_i}}(o^m, m, a^m)$ with m is the agent index.
 - 5: Search $\pi_{\theta_{i+1}}$ that maximizes $L(\theta) = E_{o \sim p_{\theta_k}, a \sim \pi_{\theta_k}} \left[\frac{\pi_{\theta}(a|o, m)}{\pi_{\theta_k}(a|o, m)} A_{\theta_k}(o, m, a) \right]$
subject to $\bar{D}_{KL}(\pi_{\theta_i} \parallel \pi_{\theta_{i+1}}) \leq \Delta$ where D_{KL} is the KL divergence between distributions of two policies, and p_{θ} are the discounted frequencies of state visitation caused by π_{θ} .
 - 6: **end for**
-

communication problem in MAS. The master agent receives and collectively processes messages from slave agents and then generates unique instructive messages to each slave agent. Slave agents use their own information and instructive messages from the master agent to take actions. This model significantly reduces the communication burden within a MAS compared to the peer-peer architecture, especially when the system has many agents.

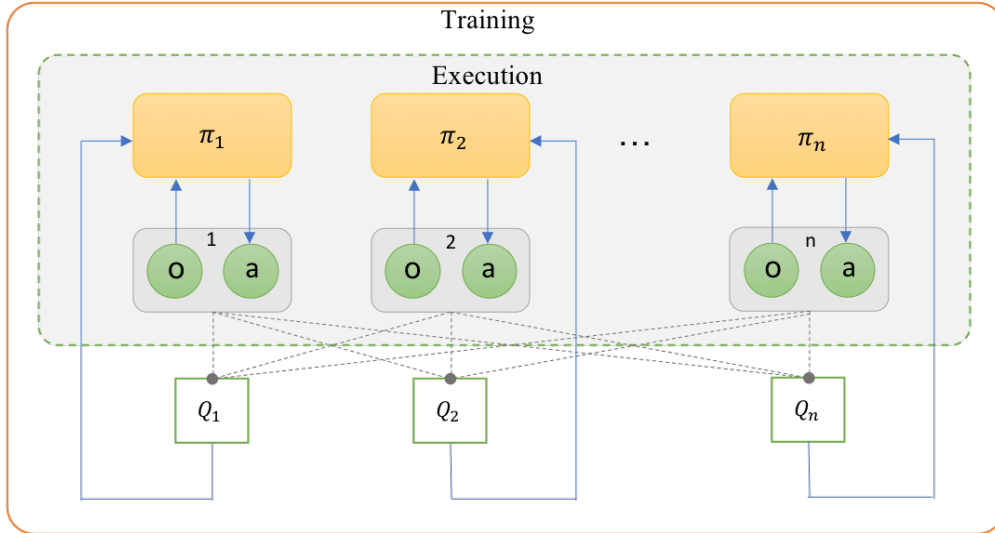


Fig. 9: Centralized learning and decentralized execution based MADDPG where policies of agents are learned by the centralized critic with augmented information from other agents' observations and actions.

Lowe et al. [68] proposed *multi-agent deep deterministic policy gradient* (MADDPG) method based on the actor-critic policy gradient algorithms. MADDPG features the centralized learning and decentralized execution paradigm in which the critic uses extra information to ease training process whilst actors take actions based on their own local observations. Fig. 9 illustrates the multi-agent decentralized actor and centralized critic components of MADDPG where only actors are used during the execution phase.

Recently, Foerster et al. [27] introduced another multi-agent actor-critic method, namely counterfactual multi-agent (COMA), which was also relied on the centralized learning and decentralized execution scheme. Unlike MADDPG [68], COMA can handle the *multi-agent credit assignment* problem [33] where agents are difficult to work out their contribution to the team's success from global rewards generated by joint actions in cooperative settings. COMA however has a disadvantage that focuses only on discrete action space whilst MADDPG is able to learn continuous policies effectively.

4.1.4 Continuous action spaces

Most deep RL models can only be applied to discrete spaces [66]. For example, DQN [73] is limited only to problems with discrete and low-dimensional action spaces, although it can handle high-dimensional obser-

vation spaces. DQN aims to find action that has maximum action-value, and therefore requires an iterative optimization process at every step in the continuous action (state) spaces. Discretising the action space is a possible solution to adapt deep RL methods to continuous domains. However, this creates many problems, notably is the curse of dimensionality: the exponential increase of action numbers against the number of degrees of freedom.

Schulman et al. [101] proposed *trust region policy optimization* (TRPO) method, which can be extended to continuous states and actions, for optimizing stochastic control policies in the domain of robotic locomotion and image-based game playing. Lillicrap et al. [66] introduced an off-policy algorithm, namely *deep deterministic policy gradient* (DDPG), which utilizes the actor-critic architecture [52, 74] to handle the continuous action spaces. Based on the deterministic policy gradient (DPG) [106], DDPG deterministically maps states to specific actions using a parameterized actor function while keeping DQN learning on the critic side. This approach however requires a large number of training episodes to find solutions, as found common in model-free reinforcement methods. Heess et al. [39] extended DDPG to *recurrent DPG* (RDPG) to handle problems with continuous action spaces under partial observability, where the true state is not available to the agents when making decisions. Recently, Gupta et al. [30] introduced the PS-TRPO method for multi-agent learning (see Algorithm 1). This method is based on the foundation of TRPO so that it can deal with continuous action spaces effectively.

4.1.5 Transfer Learning for MADRL

Training the Q-network or generally a deep RL model of a single agent is often very computationally expensive. This problem is significantly severe for a system of multiple agents. To improve the performance and reduce computational costs during training process of multiple deep RL models, several studies have promoted transfer learning for deep RL.

Rusu et al. [96, 97] proposed *policy distillation* method and *progressive neural networks* to promote transfer learning in the context of deep RL. These methods however are computationally complex and expensive [49]. Yin and Pan [117] likewise introduced another policy distillation architecture to apply knowledge transfer for deep RL. That method reduces training time and outperforms DQNs but its exploration strategy is still not efficient. Parisotto et al. [87] proposed the *actor-mimic* method for multi-task and transfer learning that improves learning speed of a deep policy network. The network can obtain an expert performance on many games simultaneously, although its model is not so complex. That method however requires a sufficient level of similarity between source and target tasks and is vulnerable to negative transfer.

Egorov [20] reformulated a multi-agent environment into an image like representation and utilized CNNs to estimate Q-values for each agent in question. That approach can address the scalability problem in MAS when the transfer learning method can be used to speed up the training process. A policy network trained on a different but related environment is used for learning process of other agents to reduce computational expenses. Experiments carried out on the pursuit-evasion problem [14] show the effectiveness of the transfer learning approach in the multi-agent domain.

Table 3 presents a summary of reviewed papers that address different multi-agent learning challenges. It can be seen that many extensions of DQN have been proposed in the literature whilst policy-based or actor-critic methods have not adequately been explored in multi-agent environments.

Table 3: Multi-agent learning challenges and their solving methods

Challenges	Value-based	Actor-critic	Policy-based
Partial observability	DRQN [36]; DDRQN [24]; RIAL and DIAL [23]; Action-specific DRQN [121]; MT-MARL [85]; PS-DQN [30]; RL as a Rehearsal (RLaR) [55]	PS-DDPG and PS-A3C [30]; MADDPG-M [48]	DPIQN and DRPIQN [42]; PS-TRPO [30]; Bayesian action decoder (BAD) [26]
Non-stationarity	DRUQN and DLCQN [12]; Multi-agent concurrent DQN [18]; Recurrent DQN-based multi-agent importance sampling and fingerprints [25]; Hysteretic-DQN [85]; Lenient-DQN [86]; WDDQN [120]	MADDPG [68]; PS-A3C [30]	PS-TRPO [30]
Continuous action spaces		Recurrent DPG [39]; DDPG [66]	TRPO [101]; PS-TRPO [30]
Multi-agent training schemes	Multi-agent extension of DQN [111]; RIAL and DIAL [23]; CommNet [109]; DRON [38]; MS-MARL [53, 54]; Linearly fuzzified joint Q-function for MAS [69]	MADDPG [68]; COMA [27]	
Transfer learning in MAS	Policy distillation [96]; Multi-task policy distillation [117]; Multi-agent DQN [20]	Progressive networks [97]	Actor-Mimic [87]

4.2 MADRL Applications

Since the success of deep RL marked by the DQN proposed in [73], many algorithms have been proposed to integrate deep learning to multi-agent learning. These algorithms can solve complex problems in various fields. This section provides a survey of these applications with a focus on the integration of deep learning and MARL. Table 4 summarizes the features and limitations of approaches to these applications.

Prasad and Dusparic [92] introduced a MADRL model to deal with energy sharing problem in a zero-energy community that comprises a collection of zero energy buildings, which have the total energy use over a year smaller than or equal to the renewables generation within each building. A deep RL agent is used to characterize each building to learn appropriate actions in sharing energy with other buildings. The global reward is modelled by the negative of the community energy status as $reward = -(\sum_{i=1}^n c(h_i) - g(h_i))$, where $c(h_i)$ and $g(h_i)$ are the energy consumed and energy generated by the i th building. The community monitoring service is introduced to manage the group membership activities such as joining and leaving the group or maintaining a list of active agents. Experiments show the superiority of the proposed model compared to the random action selection strategy in terms of net zero energy balance as a community. A limitation of the proposed approach lies in the episodic learning manner so that agent’s behaviors cannot be observed in an online fashion. Further drawbacks include that the current experiments were not performed on a large scale, i.e. ten houses at maximum, and energy price scheme is not considered.

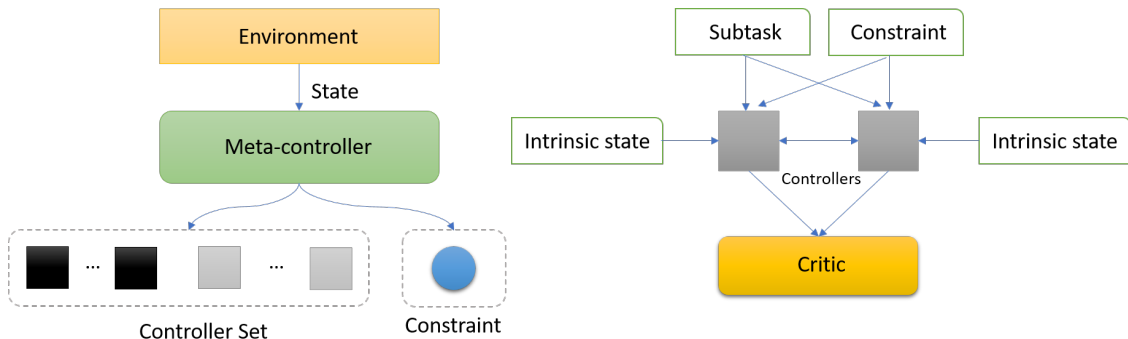


Fig. 10: Federated control with hierarchical MADRL method.

A combination of hierarchical RL and MADRL methods was developed to coordinate and control multiple agents in problems where agents’ privacy is prioritized [58]. Such distributed scheduling problems could be a multi-task dialogue where an automated assistant needs to help a user to plan for several independent tasks, for example, purchase a train ticket to the city, book a movie ticket, and make a dinner reservation in a restaurant. Each of these tasks is handled by a decentralized controller whilst the assistant is a meta-controller which benefits from *temporal abstractions* to lessen the communication complexity, and thus able to find a globally consistent solution for the user (Fig. 10). On the other hand, Leibo et al. [62] introduced a *sequential social dilemma* (SSD) model based on general-sum Markov games under partial observability to address the evolution of cooperation in MAS. Being able to capture sequential structure of real-world social dilemmas, SSD is an extension of *matrix game social dilemma* (MGSD) that has been applied to various phenomena in social science and biology [16, 8, 119]. The general-sum modelling requires solving algorithms to either track different potential equilibria for each agent or be able to find cyclic strategy consisting of multiple policies learned by using different state space sweeps [122, 89]. DQN is utilized to characterize self-interested independent learning agents to find equilibria of the SSD, which cannot be solved by the standard evolution and learning methods used for MGSD [50]. Perolat et al. [90] also demonstrated the application of MADRL in social science phenomenon, i.e. the common-pool resource (CPR) appropriation. The proposed method comprises a spatially and temporally dynamic CPR environment [46] and an MAS of independent self-interested DQNs. Through the RL trial and error process, the CPR appropriation problem is solved by self-organization that adjusts the incentives felt by independent individual agents over time.

Huttenrauch et al. [45] formulated swarm systems as a special case of the decentralized POMDP [84] and used an actor-critic deep RL approach to control a group of cooperative agents. The Q-function is learned using a global state information, which can be the view of a camera capturing the scene in swarm robotics

Table 4: A summary of typical MADRL applications in different fields

Applications	Basic DLR	Features	Limitations
Federated control [58]	Hierarchical-DQN (h-DQN) [57]	<ul style="list-style-type: none"> • Divide the control problem into disjoint subtasks and leverage <i>temporal abstractions</i>. • Use <i>meta-controller</i> to guide decentralized controllers. • Able to solve distributed scheduling problems such as multi-task dialogue, urban traffic control. 	<ul style="list-style-type: none"> • Does not address the non-stationarity problem. • Number of agents is currently limited at six. • Meta-controller's optimal policy becomes complicated and inefficient when the number of agents increases.
Large-scale fleet management [67]	Actor-critic and DQN	<ul style="list-style-type: none"> • Reallocate vehicles ahead of time to balance the transport demands and supplies. • Geographic context and collaborative context are integrated to coordinate agents. • Two proposed algorithms, <i>contextual multi-agent actor-critic</i> and <i>contextual deep Q-learning</i>, can achieve explicit coordination among thousands of agents. 	<ul style="list-style-type: none"> • Can only deal with discrete actions and each agent has a small (simplified) action space. • Assume that agents in the same region at the same time interval (i.e. same spatial-temporal state) are homogeneous.
Swarm systems [45]	DQN and DDPG	<ul style="list-style-type: none"> • Agents can only observe local environment (partial observability) but not the global state. • Use guided approach for multi-agent learning where actors make decisions based on locally sensed information whilst critic has central access to global state. 	<ul style="list-style-type: none"> • Can only work with homogeneous agents. • Unable to converge to meaningful policies in huge dimensionality and partial observed problem.
Traffic lights control [11]	DDDQN and IDQN	<ul style="list-style-type: none"> • Learning multiple agents is performed using IDQN where each agent is modelled by DDDQN. • First approach to address heterogeneous multi-agent learning in urban traffic control. • Fingerprint technique is used to stabilize the experience replay memory to handle non-stationarity. 	<ul style="list-style-type: none"> • The proposed deep RL approach learns ineffectively in high traffic conditions. • The fingerprint does not improve the performance of experience replay although the latter is required for efficient learning.
Task and resources allocation [83]	CommNet [109]	<ul style="list-style-type: none"> • Propose distributed task allocation where agents can request help from cooperating neighbors. • Three types of agents are defined: manager, participant and mediator. • Communication protocol is learned simultaneously with agents' policies through CommNet. 	<ul style="list-style-type: none"> • May not be able to deal with heterogeneous agents. • Computational deficiencies regarding the decentralization and reallocation characteristics. • Experiments only on small state action spaces.
Energy sharing optimization [92]	DQN	<ul style="list-style-type: none"> • Each building is characterized by a DRL agent to learn appropriate actions independently. • Agents' actions include: consume and store excess energy, request neighbor or supply grid for additional energy, grant or deny requests. • Agents collaborate via shared or global rewards to achieve a common goal, i.e. zero-energy status. 	<ul style="list-style-type: none"> • Agents' behaviors cannot be observed in an online fashion. • Limited number of houses, currently ten houses at maximum were experimented. • Energy price is not considered.
Keepaway soccer [59]	DQN	<ul style="list-style-type: none"> • Low-dimensional state space, described by only 13 variables. • Heterogeneous MAS, each agent has different experience replay memory and different network policy. 	<ul style="list-style-type: none"> • Number of agents is limited, currently setting with 3 keepers vs. 2 takers. • Heterogeneous learning speed is significantly lower than homogeneous case.
Action markets [100]	DQN	<ul style="list-style-type: none"> • Agents can trade their atomic actions in exchange for environmental reward. • Reduce greedy behavior and thus negative effects of individual reward maximization. • The proposed approach significantly increases the overall reward compared to methods without action trading. 	<ul style="list-style-type: none"> • Agents cannot find prices for actions by themselves because they are given at design time. • Strongly assume that agents cannot cheat on each other by making offers which they do not hold afterwards.
Sequential social dilemma (SSD) [62]	DQN	<ul style="list-style-type: none"> • Introduce an SSD model to extend MGSD to capture sequential structure of real-world social dilemmas. • Describe SSDs as general-sum Markov games with partial observations. • Multi-agent DQN is used to find equilibria of SSD problems. 	<ul style="list-style-type: none"> • Assume agent's learning is independent and regard the others as part of the environment. • Agents do not recursively reason about one another's learning.
Common-pool resource (CPR) appropriation [90]	DQN	<ul style="list-style-type: none"> • Introduce a new CPR appropriation model using an MAS containing spatially and temporally environment dynamics. • Use the <i>descriptive agenda</i> [104] to describe the behaviors emerging when agents learn in the presence of other learning agents. • Simulate multiple independent agents with each learned by DQN. 	<ul style="list-style-type: none"> • Single agent DQN is extended to multi-agent environment where the Markov assumption is no longer hold. • Agents do not do anything of <i>rational negotiation</i>, e.g. bargaining, building consensus, or making appeals.

examples. The group can perform complicated tasks such as search and rescue or distributed assembly although individual agent has limited sensory capability. That model has a drawback as it assumes agents to be homogenous. Calvo and Dusparic [11] proposed the use of IDQN to address the heterogeneity problem in multi-agent environment of urban traffic light control. Each agent is learned by the *dueling double deep Q-network* (DDDQN), which integrates dueling networks, double DQN and prioritized experience replay. Heterogenous agents are trained independently and simultaneously, considering other agents as part of the environment. The non-stationarity of the multi-agent environment is dealt with by a technique of

fingerprinting that disambiguates the age of training samples and stabilizes the replay memory.

A special application of DQN to the heterogeneous MAS where the state space is low-dimensional was presented in [59]. Experiments are performed on a multi-agent Keepaway soccer problem whose state comprises only 13 variables. To handle heterogeneity, each DQN agent is set up with different experience replay memory and different neural network. Agents cannot communicate with each other but only can observe others' behaviors. While DQNs can enhance results in terms of game score in the heterogeneous team learning settings in low-dimensional environments, their learning process is significantly slower than those in the homogeneous cases.

Establishing communication channels among agents during learning is an important step in designing and constructing MADRL algorithms. Nguyen et al. [82] characterized the communication channel via human knowledge represented by images and allow deep RL agents to communicate using these shared images. The *asynchronous advantage actor-critic* (A3C) algorithm [74] is used to learn optimal policy for each agent, which can be extended to multiple heterogeneous agents. On the other hand, Nouredine et al. [83] introduced a method, namely task allocation process using cooperative deep RL, to allow multiple agents to interact with each other and allocate resources and tasks effectively. Agents can request help from their cooperative neighbors in a loosely coupled distributed multi-agent environment. The CommNet model [109] is used to facilitate communications among agents, which are characterized by DRQN [36]. Experimental results demonstrate the great capability of that method in handling complicated task allocation problem. One of the drawbacks of that algorithm however is its limited ability in managing heterogeneous agents. In addition, its decentralization and reallocation characteristics also pose disadvantages in terms of computational time and communication overhead.

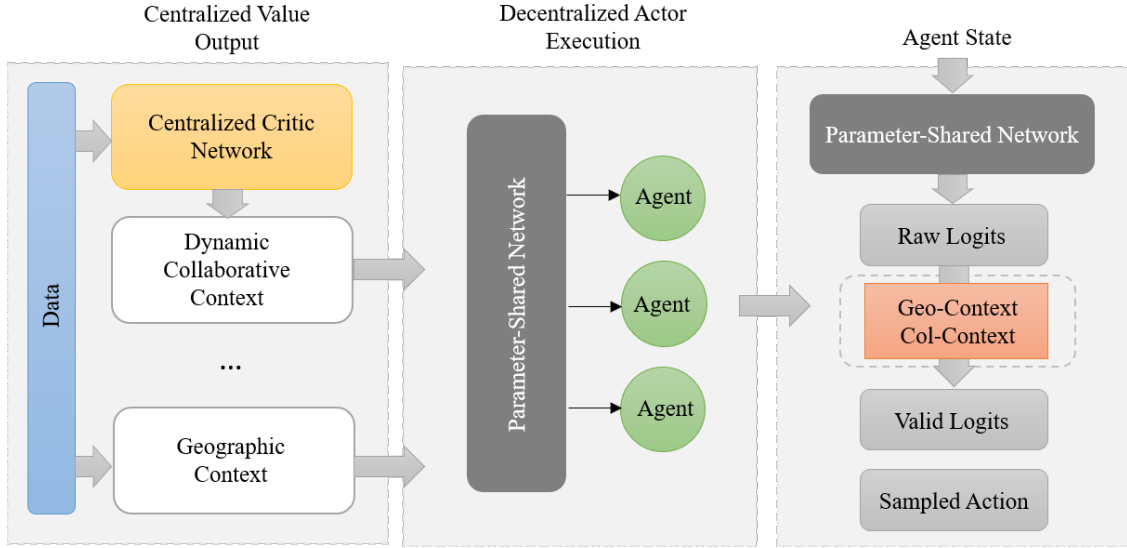


Fig. 11: The contextual multi-agent actor-critic architecture proposed in [67] where decentralized execution is coordinated using the centralized value network's outputs as illustrated in the left part whilst the right part shows how context is embedded to the policy network.

Lin et al. [67] addressed the large-scale fleet management using MADRL by proposing two algorithms, namely *contextual deep Q-learning* and *contextual multi-agent actor-critic*. These algorithms aim to balance the difference between demand and supply by reallocating transportation resources that help to reduce traffic congestion and increase transportation efficiency. The contextual multi-agent actor-critic model is illustrated in Fig. 11 where a parameter-shared policy network is used to coordinate the agents, which represent available vehicles or equivalently the idle drivers.

Recently, Schmid et al. [100] introduced an interesting approach to an MAS where agents can trade their actions in exchange for other resources, e.g. environmental rewards. The action trading was inspired by the fundamental theorem of welfare economics that competitive markets adjust towards the Pareto efficiency. Specifically, agents need to extend their action spaces and learn two policies simultaneously: one for the

original stochastic reward and another for trading environmental reward. The behavior market realized from the action trading helps mitigate greedy behavior (like the tit-for-tat game theoretic strategy proposed in [63]), enable agents to incentivize other agents and reduce the negative effects of individual reward maximization. Simulation results on the iterated matrix game and the Coin game show the effectiveness of the action trading method as it increases the social welfare, measured in terms of overall rewards of all agents.

5 Conclusions and Research Directions

This paper presents an overview of different challenges in multi-agent learning and solutions to these challenges using deep RL methods. We group the surveyed papers into five categories, including non-stationarity, partial observability, multi-agent training schemes, multi-agent transfer learning, and continuous state and action spaces. We have highlighted advantages and disadvantages of the approaches to address the challenges. Applications of MADRL methods in different fields are also reviewed thoroughly. We have found that the integration of deep learning into traditional MARL methods has been able to solve many complicated problems, such as urban traffic light control, energy sharing problem in a zero-energy community, large-scale fleet management, task and resources allocation, swarm robotics, and social science phenomena. The results indicate that deep RL-based methods provide a viable approach to handling complicated tasks in the MAS domain.

Learning from demonstration including *imitation learning* and *inverse RL* has demonstrated effectiveness in single agent deep RL [91]. On one hand, imitation learning tries to map between states to actions as a supervised approach. It directly generalizes the expert strategy to unvisited states so that it is close to a multi-class classification problem in cases of finite action set. On the other hand, inverse RL agent needs to infer a reward function from the expert demonstrations. Inverse RL assumes that the expert policy is optimal regarding the unknown reward function [31, 32]. These methods however have not yet been explored fully in multi-agent environments. Both imitation learning and inverse RL have great potential for applications in MAS. It is expected that they can reduce the learning time and improve the effectiveness of MAS. A very straightforward challenge arose from these applications is the requirement of multiple experts who are able to demonstrate the tasks collaboratively. Furthermore, the communication and reasoning capabilities of experts are difficult to be characterized and modelled by autonomous agents in the MAS domain. These pose important research questions towards extensions of imitation learning and inverse RL to MADRL methods. In addition, for complicated tasks or behaviors which are difficult for humans to demonstrate, there is a need of alternative methods that allow human preferences to be integrated into deep RL [13, 81, 82].

Deep RL has considerably facilitated autonomy, which allows to deploy many applications in robotics or autonomous vehicles. The most common drawback of deep RL models however is the ability to interact with human through human-machine teaming technologies. In complex and adversarial environments, there is a critical need for human intellect teamed with technology because humans alone cannot sustain the volume, and machines alone cannot issue creative responses when new situations are introduced. Recent advances of *human-on-the-loop* architecture [78] can be fused with MADRL to integrate humans and autonomous agents to deal with complex problems. In the conventional *human-in-the-loop* setting, agents perform their assigned tasks autonomously for a period, then stop and wait for human commands before continuing in this rate-limited fashion. In human-on-the-loop, agents execute their tasks autonomously until completion, with a human in a monitoring or supervisory role reserving the ability to intervene in operations carried out by agents. A human-on-the-loop-based architecture can be fully autonomous if human supervisors allow task completion by agents entirely on their own [78].

Model-free deep RL has been able to solve many complicated problems both in single agent and multi-agent domains. This category of methods however requires a huge number of samples and long learning time to achieve a good performance. *Model-based* methods have demonstrated effectiveness in terms of sample efficiency, transferability and generality in various problems using single as well as multi-agent models. Although the deep learning extensions of the model-based methods have been studied recently in single agent domain, e.g. [64, 29, 22, 77, 103, 15], these extensions have not been investigated widely in the multi-agent domain. This creates a research gap that could be developed to a research direction in model-based MADRL. In addition, dealing with high-dimensional observations using model-based approaches or combining elements of model-based planning and model-free policy is another active, exciting but under-explored research area.

Scaling to large systems, especially dealing with many heterogeneous agents, has been a primary challenge in RL research domain since its first days. As the world dynamics become more and more complex, this challenge has always been required to resolve. Since agents have common behaviors such as actions, domain knowledge, and goals (homogeneous agents), the scalability can be achievable by (partially) centralized training and decentralized execution [94, 28]. In the heterogeneous setting with many agents, the key challenge is how to provide the most optimal solution and maximize the task completion success based on self-learning with effective coordinative and cooperative strategies among the agents. This is more problematic in hostile environments where communications among agents are limited and in scenarios that involve more heterogeneous agents as the credit assignment problem become increasingly difficult. A research direction to address these difficulties is well worth an investigation.

Regarding applications of multi-agent learning, there have been many studies using traditional MARL methods to solve various problems such as controlling a group of autonomous vehicles or drones [43], robot soccer [102], controlling traffic signals [75], coordinating collaborative bots in factories and warehouse [47], controlling electrical power networks [93] or optimizing distributed sensor networks [37], automated trading [88], machine bidding in competitive e-commerce and financial markets [9], resource management [44], transportation [21], and phenomena of social sciences [62]. Since the emergence of DQN [73], efforts to extend traditional RL to deep RL in the multi-agent domain have been found in the literature but they are still very limited (see Table 4 for applications available in the current literature). Many applications of MARL can now be solved effectively by MADRL based on its high-dimension handling capability. Therefore, there is a need of further empirical research to apply MADRL methods to effectively solve complex real-world problems such as the aforementioned applications.

References

- [1] Abdallah, S., and Kaisers, M. (2013, May). Addressing the policy-bias of Q-learning by repeating updates. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1045-1052).
- [2] Abdallah, S., and Kaisers, M. (2016). Addressing environment non-stationarity by repeating Q-learning updates. *The Journal of Machine Learning Research*, 17(1), 1582-1612.
- [3] Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). Deep reinforcement learning: a brief survey. *IEEE Signal Processing Magazine*, 34(6), 26-38.
- [4] Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Kttler, H., ... and Schrittwieser, J. (2016). DeepMind lab. *arXiv preprint arXiv:1612.03801*.
- [5] Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: an evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47, 253-279.
- [6] Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8), 716-719.
- [7] Benbrahim, H., and Franklin, J. A. (1997). Biped dynamic walking using reinforcement learning. *Robotics and Autonomous Systems*, 22(3-4), 283-302.
- [8] Bloembergen, D., Tuyls, K., Hennes, D., and Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: a survey. *Journal of Artificial Intelligence Research*, 53, 659-697.
- [9] Brandouy, O., Mathieu, P., and Veryzhenko, I. (2011, January). On the design of agent-based artificial stock markets. In *International Conference on Agents and Artificial Intelligence* (pp. 350-364).
- [10] Busoniu, L., Babuska, R., and De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 38 (2), 156-172.

- [11] Calvo, J. J. A., and Dusparic, I. (2018). Heterogeneous multi-agent deep reinforcement learning for traffic lights control. *The 26th Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 1-12), Dublin, Ireland.
- [12] Castaneda, A. O. (2016). *Deep Reinforcement Learning Variants of Multi-Agent Learning Algorithms* (Master’s Thesis, School of Informatics, University of Edinburgh).
- [13] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (pp. 4299-4307).
- [14] Chung, T. H., Hollinger, G. A., and Isler, V. (2011). Search and pursuit-evasion in mobile robotics. *Autonomous Robots*, 31(4), 299.
- [15] Corneil, D., Gerstner, W., and Brea, J. (2018). Efficient model-based deep reinforcement learning with variational state tabulation. *arXiv preprint arXiv:1802.04325*.
- [16] de Cote, E. M., Lazaric, A., and Restelli, M. (2006, May). Learning to cooperate in multi-agent social dilemmas. In *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 783-785). ACM.
- [17] Deng, L., and Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends in Signal Processing*, 7(34), 197-387.
- [18] Diallo, E. A. O., Sugiyama, A., and Sugawara, T. (2017, December). Learning to coordinate with deep reinforcement learning in doubles Pong game. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on* (pp. 14-19). IEEE.
- [19] Duan, Y., Chen, X., Houthoofd, R., Schulman, J., and Abbeel, P. (2016, June). Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning* (pp. 1329-1338).
- [20] Egorov, M. (2016). Multi-agent deep reinforcement learning. Stanford University.
- [21] Fernandez-Gauna, B., Etxeberria-Agiriano, I., and Grana, M. (2015). Learning multirobot hose transportation and deployment by distributed round-robin Q-learning. *PloS One*, 10(7), e0127129.
- [22] Finn, C., and Levine, S. (2017, May). Deep visual foresight for planning robot motion. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on* (pp. 2786-2793). IEEE.
- [23] Foerster, J., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems* (pp. 2137-2145).
- [24] Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). Learning to communicate to solve riddles with deep distributed recurrent Q-networks. *arXiv preprint arXiv:1602.02672*.
- [25] Foerster, J., Nardelli, N., Farquhar, G., Afouras, T., Torr, P. H., Kohli, P., and Whiteson, S. (2017, July). Stabilising experience replay for deep multi-agent reinforcement learning. In *International Conference on Machine Learning* (pp. 1146-1155).
- [26] Foerster, J. N., Song, F., Hughes, E., Burch, N., Dunning, I., Whiteson, S., ... and Bowling, M. (2018). Bayesian action decoder for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1811.01458*.
- [27] Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. (2018). Counterfactual multi-agent policy gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 2974-2982).
- [28] Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018, July). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (pp. 122-130).

- [29] Gu, S., Lillicrap, T., Sutskever, I., and Levine, S. (2016, June). Continuous deep Q-learning with model-based acceleration. In *International Conference on Machine Learning* (pp. 2829-2838).
- [30] Gupta, J. K., Egorov, M., and Kochenderfer, M. (2017, May). Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems* (pp. 66-83).
- [31] Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems* (pp. 3909-3917).
- [32] Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., and Dragan, A. (2017). Inverse reward design. In *Advances in Neural Information Processing Systems* (pp. 6765-6774).
- [33] Harati, A., Ahmadabadi, M. N., and Araabi, B. N. (2007). Knowledge-based multiagent credit assignment: a study on task type and critic information. *IEEE Systems Journal*, 1(1), 55-67.
- [34] Hasselt, H. V. (2010). Double Q-learning. In *Advances in Neural Information Processing Systems* (pp. 2613-2621).
- [35] Hasselt, H. V., Guez, A., and Silver, D. (2016, February). Deep reinforcement learning with double Q-Learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 2094-2100). AAAI Press.
- [36] Hausknecht, M., and Stone, P. (2015). Deep recurrent Q-learning for partially observable MDPs. *CoRR*, abs/1507.06527, 7(1).
- [37] He, J., Peng, J., Jiang, F., Qin, G., and Liu, W. (2015). A distributed Q learning spectrum decision scheme for cognitive radio sensor network. *International Journal of Distributed Sensor Networks*, 11(5), 301317.
- [38] He, H., Boyd-Graber, J., Kwok, K., and Daum III, H. (2016, June). Opponent modeling in deep reinforcement learning. In *International Conference on Machine Learning* (pp. 1804-1813).
- [39] Heess, N., Hunt, J. J., Lillicrap, T. P., and Silver, D. (2015). Memory-based control with recurrent neural networks. *arXiv preprint arXiv:1512.04455*.
- [40] Hernandez-Leal, P., Kaisers, M., Baarslag, T., and de Cote, E. M. (2017). A survey of learning in multiagent environments: dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*.
- [41] Hernandez-Leal, P., Kartal, B., and Taylor, M. E. (2018). Is multiagent deep reinforcement learning the answer or the question? a brief survey. *arXiv preprint arXiv:1810.05587*.
- [42] Hong, Z. W., Su, S. Y., Shann, T. Y., Chang, Y. H., and Lee, C. Y. (2018, July). A deep policy inference Q-network for multi-agent systems. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1388-1396).
- [43] Hung, S. M., and Givigi, S. N. (2017). A Q-learning approach to flocking with UAVs in a stochastic environment. *IEEE Transactions on Cybernetics*, 47(1), 186-197.
- [44] Hussin, M., Hamid, N. A. W. A., and Kasmiran, K. A. (2015). Improving reliability in resource management through adaptive reinforcement learning for distributed systems. *Journal of Parallel and Distributed Computing*, 75, 93-100.
- [45] Huttenrauch, M., Sodic, A., and Neumann, G. (2017). Guided deep reinforcement learning for swarm systems. *arXiv preprint arXiv:1709.06011*.
- [46] Janssen, M. A., Holahan, R., Lee, A., and Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science*, 328(5978), 613-617.

- [47] Kattepur, A., Rath, H. K., Simha, A., and Mukherjee, A. (2018, April). Distributed optimization in multi-agent robotics for industry 4.0 warehouses. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 808-815). ACM.
- [48] Kilinc, O., and Montana, G. (2018). Multi-agent deep reinforcement learning with extremely noisy observations. *arXiv preprint arXiv:1812.00922*.
- [49] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... and Hassabis, D. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521-3526.
- [50] Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., and Tenenbaum, J. B. (2016, January). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1679-1684).
- [51] Klopff, A. (1972). Brain function and adaptive systems: a heterostatic theory. Air Force Cambridge Research Laboratories Special Report.
- [52] Konda, V. R., and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems* (pp. 1008-1014).
- [53] Kong, X., Xin, B., Liu, F., and Wang, Y. (2017). Effective master-slave communication on a multi-agent deep reinforcement learning system. *Hierarchical Reinforcement Learning Workshop at the 31st Conference on NIPS*, Long Beach, CA, USA.
- [54] Kong, X., Xin, B., Liu, F., and Wang, Y. (2017). Revisiting the master-slave architecture in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1712.07305*.
- [55] Kraemer, L., and Banerjee, B. (2016). Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 190, 82-94.
- [56] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097-1105).
- [57] Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation. In *Advances in Neural Information Processing Systems* (pp. 3675-3683).
- [58] Kumar, S., Shah, P., Hakkani-Tur, D., and Heck, L. (2017). Federated control with hierarchical multi-agent deep reinforcement learning. *arXiv preprint arXiv:1712.08266*.
- [59] Kurek, M., and Jakowski, W. (2016, September). Heterogeneous team deep Q-learning in low-dimensional multi-agent environments. In *Computational Intelligence and Games (CIG), 2016 IEEE Conference on* (pp. 1-8). IEEE.
- [60] Lample, G., and Chaplot, D. S. (2017, February). Playing FPS games with deep reinforcement learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 2140-2146).
- [61] Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Perolat, J., Silver, D., and Graepel, T. (2017). A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems* (pp. 4193-4206).
- [62] Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017, May). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems* (pp. 464-473).
- [63] Lerer, A., and Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.

- [64] Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1), 1334-1373.
- [65] Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.
- [66] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.
- [67] Lin, K., Zhao, R., Xu, Z., and Zhou, J. (2018). Efficient large-scale fleet management via multi-agent deep reinforcement learning. *arXiv preprint arXiv:1802.06444*.
- [68] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems* (pp. 6379-6390).
- [69] Luviano-Cruz, D., Garcia-Luna, F., Perez-Dominguez, L., and Gadi, S. (2018). Multi-agent reinforcement learning using linear fuzzy model applied to cooperative mobile robots. *Symmetry*, 10(10), 461.
- [70] Mahadevan, S., and Connell, J. (1992). Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, 55(2-3), 311-365.
- [71] Matignon, L., Laurent, G., and Le Fort-Piat, N. (2007, October). Hysteretic Q-Learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 64-69).
- [72] Minsky, M. L. (1954). Theory of neural-analog reinforcement systems and its application to the brain model problem. Princeton University.
- [73] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... and Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
- [74] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... and Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning* (pp. 1928-1937).
- [75] Mousavi, S. S., Schukat, M., and Howley, E. (2017). Traffic light control using deep policy-gradient and value-function-based reinforcement learning. *IET Intelligent Transport Systems*, 11(7), 417-423.
- [76] Mulling, K., Kober, J., Kroemer, O., and Peters, J. (2013). Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, 32(3), 263-279.
- [77] Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. (2018, May). Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 7559-7566). IEEE.
- [78] Nahavandi, S. (2017). Trusted autonomy between humans and robots: toward human-on-the-loop in robotics and autonomous systems. *IEEE Systems, Man, and Cybernetics Magazine*, 3(1), 10-17.
- [79] Nguyen, N. D., Nguyen, T., and Nahavandi, S. (2017). System design perspective for human-level agents using deep reinforcement learning: a survey. *IEEE Access*, 5, 27091-27102.
- [80] Nguyen, T. (2018). A multi-objective deep reinforcement learning framework. *arXiv preprint arXiv:1803.02965*.
- [81] Nguyen, N. D., Nahavandi, S., and Nguyen, T. (2018). A human mixed strategy approach to deep reinforcement learning. In *2018 IEEE International Conference on Systems, Man, and Cybernetics*, (pp. 4023-4028).
- [82] Nguyen, T., Nguyen, N. D., and Nahavandi, S. (2018). Multi-agent deep reinforcement learning with human strategies. *arXiv preprint arXiv:1806.04562*.

- [83] Nouredine, D., Gharbi, A. and Ahmed, S. (2017). Multi-agent deep reinforcement learning for task allocation in dynamic environment. In *Proceedings of the 12th International Conference on Software Technologies (ICSOFT)*, pp. 17-26.
- [84] Oliehoek, F. A. (2012). Decentralized POMDPs. In *Reinforcement Learning* (pp. 471-503). Springer, Berlin, Heidelberg.
- [85] Omidshafiei, S., Pazis, J., Amato, C., How, J. P., and Vian, J. (2017, July). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning* (pp. 2681-2690).
- [86] Palmer, G., Tuyls, K., Bloembergen, D., and Savani, R. (2018, July). Lenient multi-agent deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems* (pp. 443-451).
- [87] Parisotto, E., Ba, J. L., and Salakhutdinov, R. (2015). Actor-mimic: deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.
- [88] Pendharkar, P. C., and Cusatis, P. (2018). Trading financial indices with reinforcement learning agents. *Expert Systems with Applications*, 103, 1-13.
- [89] Perolat, J., Piot, B., Scherrer, B., and Pietquin, O. (2016, May). On the Use of Non-Stationary Strategies for Solving Two-Player Zero-Sum Markov Games. In *The 19th International Conference on Artificial Intelligence and Statistics* (pp. 893-901).
- [90] Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems* (pp. 3643-3652).
- [91] Piot, B., Geist, M., and Pietquin, O. (2017). Bridging the gap between imitation learning and inverse reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(8), 1814-1826.
- [92] Prasad, A., and Dusparic, I. (2018). Multi-agent deep reinforcement learning for zero energy communities. *arXiv preprint arXiv:1810.03679*.
- [93] Rahman, M. S., Mahmud, M. A., Pota, H. R., Hossain, M. J., and Orchi, T. F. (2015). Distributed multi-agent-based protection scheme for transient stability enhancement in power systems. *International Journal of Emerging Electric Power Systems*, 16(2), 117-129.
- [94] Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. (2018). QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485*.
- [95] Riedmiller, M., Gabel, T., Hafner, R., and Lange, S. (2009). Reinforcement learning for robot soccer. *Autonomous Robots*, 27(1), 55-73.
- [96] Rusu, A. A., Colmenarejo, S. G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., ... and Hadsell, R. (2015). Policy distillation. *arXiv preprint arXiv:1511.06295*.
- [97] Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., ... and Hadsell, R. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.
- [98] Schaal, S. (1997). Learning from demonstration. In *Advances in Neural Information Processing Systems* (pp. 1040-1046).
- [99] Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.
- [100] Schmid, K., Belzner, L., Gabor, T., and Phan, T. (2018, October). Action markets in deep multi-agent reinforcement learning. In *International Conference on Artificial Neural Networks* (pp. 240-249).

- [101] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015, June). Trust region policy optimization. In *International Conference on Machine Learning* (pp. 1889-1897).
- [102] Schwab, D., Zhu, Y., and Veloso, M. (2018, July). Zero shot transfer learning for robot soccer. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 2070-2072).
- [103] Serban, I. V., Sankar, C., Pieper, M., Pineau, J., and Bengio, Y. (2018). The bottleneck simulator: a model-based deep reinforcement learning approach. *arXiv preprint arXiv:1807.04723*.
- [104] Shoham, Y., Powers, R., and Grenager, T. (2007). If multi-agent learning is the answer, what is the question?. *Artificial Intelligence*, 171(7), 365-377.
- [105] Silva, F.L., Taylor, M.E., and Costa, A.H.R. (2018). Autonomously reusing knowledge in multiagent reinforcement learning. In *The 27th International Joint Conference on Artificial Intelligence*, pp. 5487-5493.
- [106] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014, January). Deterministic policy gradient algorithms. In *International Conference on Machine Learning* (pp. 387-395).
- [107] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... and Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.
- [108] Sorokin, I., Seleznev, A., Pavlov, M., Fedorov, A., and Ignateva, A. (2015). Deep attention recurrent Q-network. *arXiv preprint arXiv:1512.01693*.
- [109] Sukhbaatar, S., Szlam, A., and Fergus, R. (2016). Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems* (pp. 2244-2252).
- [110] Sutton, R. S., and Barto, A. G. (1998). Reinforcement learning: An introduction. MIT press.
- [111] Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., and Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS One*, 12(4), e0172395.
- [112] Tan, M. (1993). Multi-agent reinforcement learning: independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning* (pp. 330-337).
- [113] Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associate processes in animals. *American Psychologist*, 53(10), 1125-1127.
- [114] Tsitsiklis, J. N., and Van Roy, B. (1997). Analysis of temporal-difference learning with function approximation. In *Advances in Neural Information Processing Systems* (pp. 1075-1081).
- [115] Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., and Freitas, N. (2016, June). Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning* (pp. 1995-2003).
- [116] Watkins, C. J., and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3-4), 279-292.
- [117] Yin, H., and Pan, S. J. (2017, January). Knowledge transfer for deep reinforcement learning with hierarchical experience replay. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 1640-1646).
- [118] Yu, C., Zhang, M., Ren, F., and Tan, G. (2015). Multiagent learning of coordination in loosely coupled multiagent systems. *IEEE Transactions on Cybernetics*, 45(12), 2853-2867.
- [119] Yu, C., Zhang, M., Ren, F., and Tan, G. (2015). Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE Transactions on Neural Networks and Learning Systems*, 26(12), 3083-3096.

- [120] Zheng, Y., Meng, Z., Hao, J., and Zhang, Z. (2018, August). Weighted double deep multiagent reinforcement learning in stochastic cooperative environments. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 421-429).
- [121] Zhu, P., Li, X., and Poupart, P. (2017). On improving deep reinforcement learning for POMDPs. *arXiv preprint arXiv:1704.07978*.
- [122] Zinkevich, M., Greenwald, A., and Littman, M. L. (2006). Cyclic equilibria in Markov games. In *Advances in Neural Information Processing Systems* (pp. 1641-1648).