A Survey and Critique of Multiagent Deep Reinforcement Learning[☆]

Pablo Hernandez-Leal, Bilal Kartal and Matthew E. Taylor
{pablo.hernandez,bilal.kartal,matthew.taylor}@borealisai.com

Borealis AI
Edmonton, Canada

Abstract

Deep reinforcement learning (RL) has achieved outstanding results in recent years. This has led to a dramatic increase in the number of applications and methods. Recent works have explored learning beyond single-agent scenarios and have considered multiagent learning (MAL) scenarios. Initial results report successes in complex multiagent domains, although there are several challenges to be addressed. The primary goal of this article is to provide a clear overview of current multiagent deep reinforcement learning (MDRL) literature. Additionally, we complement the overview with a broader analysis: (i) we revisit previous key components, originally presented in MAL and RL, and highlight how they have been adapted to multiagent deep reinforcement learning settings. (ii) We provide general guidelines to new practitioners in the area: describing lessons learned from MDRL works, pointing to recent benchmarks, and outlining open avenues of research. (iii) We take a more critical tone raising practical challenges of MDRL (e.g., implementation and computational demands). We expect this article will help unify and motivate future research to take advantage of the abundant literature that exists (e.g., RL and MAL) in a joint effort to promote fruitful research in the multiagent community.

1. Introduction

Almost 20 years ago Stone and Veloso's seminal survey [1] laid the groundwork for defining the area of multiagent systems (MAS) and its open problems in the context of AI. About ten years ago, Shoham, Powers, and Grenager [2] noted that the literature on multiagent learning (MAL) was growing and it was not possible to enumerate all relevant articles. Since then, the number of published MAL works continues to steadily rise, which led to different surveys on the area, ranging from analyzing the basics of MAL and their challenges [3, 4, 5], to addressing specific subareas: game theory and MAL [2, 6], cooperative scenarios [7, 8], and evolutionary dynamics of MAL [9]. In just the last couple of years, three surveys related to MAL have been published: learning in non-stationary environments [10], agents modeling agents [11], and transfer learning in multiagent RL [12].

The research interest in MAL has been accompanied by successes in artificial intelligence, first, in single-agent video games [13]; more recently, in two-player games, for example, playing

[☆]Earlier versions of this work had the title: "Is multiagent deep reinforcement learning the answer or the question? A brief survey"

Go [14, 15], poker [16, 17], and games of two competing teams, e.g., DOTA 2 [18] and StarCraft II [19].

While different techniques and algorithms were used in the above scenarios, in general, they are all a combination of techniques from two main areas: reinforcement learning (RL) [20] and deep learning [21, 22].

RL is an area of machine learning where an agent learns by interacting (i.e., taking actions) within a dynamic environment. However, one of the main challenges to RL, and traditional machine learning in general, is the need for manually designing quality features on which to learn. Deep learning enables efficient representation learning, thus allowing the automatic discovery of features [21, 22]. In recent years, deep learning has had successes in different areas such as computer vision and natural language processing [21, 22]. One of the key aspects of deep learning is the use of *neural networks* (NNs) that can find compact representations in high-dimensional data [23].

In deep reinforcement learning (DRL) [23, 24] deep neural networks are trained to approximate the optimal policy and/or the value function. In this way the deep NN, serving as function approximator, enables powerful generalization. One of the key advantages of DRL is that it enables RL to scale to problems with high-dimensional state and action spaces. However, most existing successful DRL applications so far have been on visual domains (e.g., Atari games), and there is still a lot of work to be done for more realistic applications [25, 26] with complex dynamics, which are not necessarily vision-based.

DRL has been regarded as an important component in constructing general AI systems [27] and has been successfully integrated with other techniques, e.g., search [14], planning [28], and more recently with multiagent systems, with an emerging area of *multiagent deep reinforcement learning (MDRL)*[29, 30].[1]

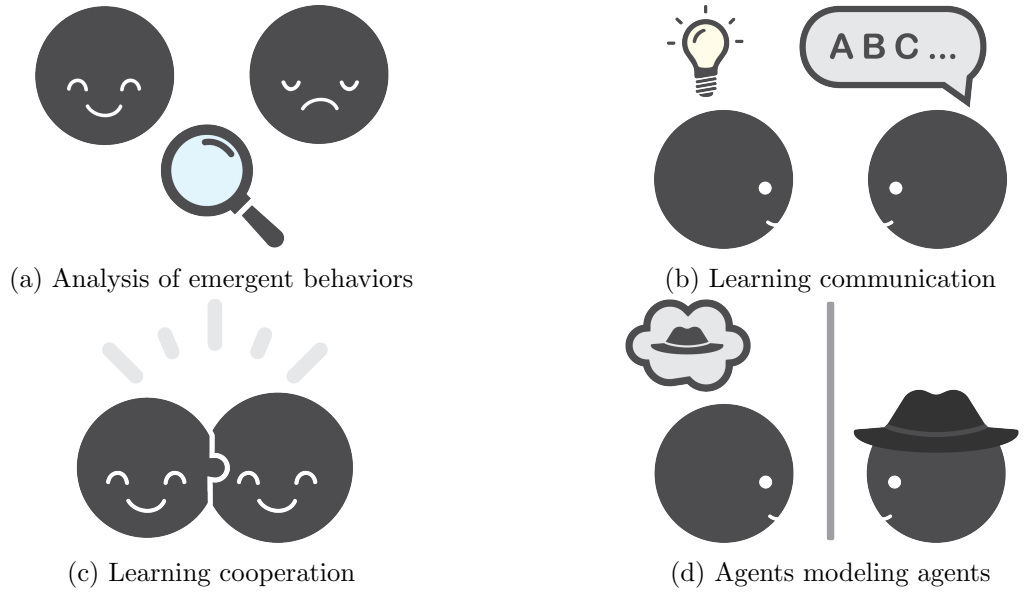
Learning in multiagent settings is fundamentally more difficult than the single-agent case due to the presence of multiagent pathologies, e.g., the moving target problem (non-stationarity) [2, 5, 10], curse of dimensionality [2, 5], multiagent credit assignment [31, 32], global exploration [8], and relative overgeneralization [33, 34, 35]. Despite this complexity, top AI conferences like AAAI, ICML, ICLR, IJCAI and NeurIPS, and specialized conferences such as AAMAS, have published works reporting successes in MDRL. In light of these works, we believe it is pertinent to first, have an overview of the recent MDRL works, and second, understand how these recent works relate to the existing literature.

This article contributes to the state of the art with a brief survey of the current works in MDRL in an effort to complement existing surveys on multiagent learning [36, 10], cooperative learning [7, 8], agents modeling agents [11], knowledge reuse in multiagent RL [12], and (single-agent) deep reinforcement learning [23, 37].

First, we provide a short review of key algorithms in RL such as Q-learning and REINFORCE (see Section 2.1). Second, we review DRL highlighting the challenges in this setting and reviewing recent works (see Section 2.2). Third, we present the multiagent setting and give an overview of key challenges and results (see Section 3.1). Then, we present the identified four categories to group recent MDRL works (see Figure 1):

- Analysis of emergent behaviors: evaluate single-agent DRL algorithms in multiagent scenarios (e.g., Atari games, social dilemmas, 3D competitive games).
- Learning communication: agents learn communication protocols to solve cooperative tasks.

[1] We have noted inconsistency in abbreviations such as: D-MARL, MADRL, deep-multiagent RL and MA-DRL.



(a) Analysis of emergent behaviors

(b) Learning communication

(c) Learning cooperation

(d) Agents modeling agents

Figure 1: Categories of different MDRL works. (a) Analysis of emergent behaviors: evaluate single-agent DRL algorithms in multiagent scenarios. (b) Learning communication: agents learn with actions and through messages. (c) Learning cooperation: agents learn to cooperate using only actions and (local) observations. (d) Agents modeling agents: agents reason about others to fulfill a task (e.g., cooperative or competitive). For a more detailed description see Sections 3.3–3.6 and Tables 1–4.

- Learning cooperation: agents learn to cooperate using only actions and (local) observations.
- Agents modeling agents: agents reason about others to fulfill a task (e.g., best response learners).

For each category we provide a description as well as outline the recent works (see Section 3.2 and Tables 1–4). Then, we take a step back and reflect on how these new works relate to the existing literature. In that context, first, we present examples on how methods and algorithms originally introduced in RL and MAL were successfully been scaled to MDRL (see Section 4.1). Second, we provide some pointers for new practitioners in the area by describing general *lessons learned* from the existing MDRL works (see Section 4.2) and point to recent multiagent benchmarks (see Section 4.3). Third, we take a more critical view and describe practical challenges in MDRL, such as reproducibility, hyperparameter tunning, and computational demands (see Section 4.4). Then, we outline some open research questions (see Section 4.5). Lastly, we present our conclusions from this work (see Section 5).

Our goal is to outline a recent and active area (i.e., MDRL), as well as to motivate future research to take advantage of the ample and existing literature in multiagent learning. We aim to enable researchers with experience in either DRL or MAL to gain a common understanding about recent works, and open problems in MDRL, and to avoid having scattered sub-communities with little interaction [2, 10, 11, 38].

2. Single-agent learning

This section presents the formalism of reinforcement learning and its main components before outlining *deep* reinforcement learning along with its particular challenges and recent algorithms.

For a more detailed description we refer the reader to excellent books and surveys on the area [39, 20, 23, 40, 24].

2.1. Reinforcement learning

RL formalizes the interaction of an agent with an environment using a Markov decision process (MDP) [41]. An MDP is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, R, T, \gamma \rangle$ where $\mathcal{S}$ represents a finite set of states. $\mathcal{A}$ represents a finite set of actions. The transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ determines the probability of a transition from any state $s \in \mathcal{S}$ to any state $s' \in \mathcal{S}$ given any possible action $a \in \mathcal{A}$. The reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ defines the immediate and possibly stochastic reward that an agent would receive given that the agent executes action a while in state s and it is transitioned to state s', $\gamma \in [0, 1]$ represents the discount factor that balances the trade-off between immediate rewards and future rewards.

MDPs are adequate models to obtain optimal decisions in *single* agent fully observable environments.[2] Solving an MDP will yield a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, which is a mapping from states to actions. An optimal policy π^* is the one that maximizes the expected discounted sum of rewards. There are different techniques for solving MDPs assuming a complete description of all its elements. One of the most common techniques is the value iteration algorithm [44], which requires a complete and accurate representation of states, actions, rewards, and transitions. However, this may be difficult to obtain in many domains. For this reason, RL algorithms often learn from experience interacting with the environment in discrete time steps.

Q-learning. One of the most well known algorithms for RL is Q-learning [45]. It has been devised for stationary, single-agent, fully observable environments with discrete actions. A Q-learning agent keeps the estimate of its expected payoff starting in state s, taking action a as $\hat{Q}(s, a)$. Each tabular entry $\hat{Q}(s, a)$ is an estimate of the corresponding optimal Q^* function that maps state-action pairs to the discounted sum of future rewards starting with action a at state s and following the optimal policy thereafter. Each time the agent transitions from a state s to a state s' via action a receiving payoff r, the Q table is updated as follows:

$$\hat{Q}(s, a) \leftarrow \hat{Q}(s, a) + \alpha[(r + \gamma \max_{a'} \hat{Q}(s', a')) - \hat{Q}(s, a)] \tag{1}$$

with the learning rate $\alpha \in [0, 1]$. Q-learning is proven to converge to Q^* if state and action spaces are discrete and finite, the sum of the learning rates goes to infinity (so that each state-action pair is visited *infinitely* often) and that the sum of the squares of the learning rates is finite (which is required to show that the convergence is with probability one) [46, 45, 47, 48, 49, 50, 51]. The convergence of single-step on-policy RL algorithms, i.e, SARSA ($\lambda = 0$), for both decaying exploration (greedy in the limit with infinite exploration) and persistent exploration (selecting actions probabilistically according to the ranks of the Q values) was demonstrated by Singh et al. [52]. Furthermore, Van Seijen [53] has proven convergence for Expected SARSA (see Section 3.1 for convergence results in multiagent domains).

[2] A Partially Observable Markov Decision Process (POMDP) [42, 43] explicitly models environments where the agent no longer sees the true system state and instead receives an *observation* (generated from the underlying system state).

REINFORCE (Monte Carlo policy gradient). In contrast to value-based methods, which do not try to optimize directly over a policy space [54], policy gradient methods can learn parameterized policies without using intermediate value estimates.

Policy parameters are learned by following the gradient of some performance measure with gradient descent [55]. For example, REINFORCE [56] uses estimated return by Monte Carlo (MC) methods with full episode trajectories to learn policy parameters θ, with $\pi(a; s, \theta) \approx \pi(a; s)$, as follows

$$\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla \pi(A_t; S_t, \theta_t)}{\pi(A_t; S_t, \theta_t)} \tag{2}$$

where G_t represents the return, α is the learning rate, and $A_t \sim \pi$. A main limitation is that policy gradient methods can have high variance [54].

The policy gradient update can be generalized to include a comparison to an arbitrary *baseline* of the state [56]. The baseline, $b(s)$, can be any function, as long as it does not vary with the action; the baseline leaves the expected value of the update unchanged, but it can have an effect on its variance [20]. A natural choice for the baseline is a learned state-value function, this reduces the variance, and it is bias-free if learned by MC.[3] Moreover, when using the state-value function for bootstrapping (updating the value estimate for a state from the estimated values of subsequent states) it assigns credit (reducing the variance but introducing bias), i.e., criticizes the policy's action selections. Thus, in actor-critic methods [54], the actor represents the policy, i.e., action-selection mechanism, whereas a critic is used for the value function learning. In the case when the critic learns a state-action function (Q function) and a state value function (V function), an *advantage function* can be computed by subtracting state values from the state-action values [20, 60]. The advantage function indicates the relative quality of an action compared to other available actions computed from the baseline, i.e., state value function. An example of an actor-critic algorithm is Deterministic Policy Gradient (DPG) [61]. In DPG [61] the critic follows the standard Q-learning and the actor is updated following the gradient of the policy's performance [62], DPG was later extended to DRL (see Section 2.2) and MDRL (see Section 3.5). For multiagent learning settings the variance is further increased as all the agents' rewards depend on the rest of the agents, and it is formally shown that as the number of agents increase, the probability of taking a correct gradient direction decreases exponentially [63]. Recent MDRL works addressed this high variance issue, e.g., COMA [64] and MADDPG [63] (see Section 3.5).

Policy gradient methods have a clear connection with deep reinforcement learning since *the policy might be represented by a neural network* whose input is a representation of the state, whose output are action selection probabilities or values for continuous control [65], and whose weights are the policy parameters.

2.2. Deep reinforcement learning

While tabular RL methods such as Q-learning are successful in domains that do not suffer from the curse of dimensionality, there are many limitations: learning in large state spaces can be prohibitively slow, methods do not generalize (across the state space), and state representations need to be hand-specified [20]. Function approximators tried to address those limitations, using

[3] Action-dependant baselines had been proposed [57, 58], however, a recent study by Tucker et al. [59] found that in many works the reason of good performance was because of bugs or errors in the code, rather than the proposed method itself.

for example, decision trees [66], tile coding [67], radial basis functions [68], and locally weighted regression [69] to approximate the value function.

Similarly, these challenges can be addressed by using deep learning, i.e., neural networks [69, 66] as function approximators. For example, $Q(s, a; \theta)$ can be used to approximate the state-action values with θ representing the neural network weights. This has two advantages, first, deep learning helps to generalize across states improving the sample efficiency for large state-space RL problems. Second, deep learning can be used to reduce (or eliminate) the need for manually designing features to represent state information [21, 22].

However, extending deep learning to RL problems comes with additional challenges including non-i.i.d. (not independently and identically distributed) data. Many supervised learning methods assume that training data is from an i.i.d. stationary distribution [70, 22, 71]. However, in RL, training data consists of highly correlated sequential agent-environment interactions, which violates the *independence* condition. Moreover, RL training data distribution is non-stationary as the agent actively learns while exploring different parts of the state space, violating the condition of sampled data being *identically distributed* [72].

In practice, using function approximators in RL requires making crucial representational decisions and poor design choices can result in estimates that diverge from the optimal value function [73, 69, 74, 75, 76, 77]. In particular, function approximation, bootstrapping, and off-policy learning are considered the three main properties that when combined, can make the learning to diverge and are known as *the deadly triad* [77, 20]. Recently, some works have shown that non-linear (i.e., deep) function approximators poorly estimate the value function [78, 59, 79] and another work found problems with Q-learning using function approximation (over/under-estimation, instability and even divergence) due to the *delusional* bias: "delusional bias occurs whenever a backed-up value estimate is derived from action choices that are not realizable in the underlying policy class"[80]. Additionally, convergence results for reinforcement learning using function approximation are still scarce [74, 81, 82, 83, 80]; in general, stronger convergence guarantees are available for policy-gradient methods [55] than for value-based methods [20].

Below we mention how the existing DRL methods aim to address these challenges when briefly reviewing value-based methods, such as DQN [13]; policy gradient methods, like Proximal Policy Optimization (PPO) [60]; and actor-critic methods like Asynchronous Advantage Actor-Critic (A3C) [84]. We refer the reader to recent surveys on single-agent DRL [23, 37, 24] for a more detailed discussion of the literature.

Value-based methods. The major breakthrough work combining deep learning with Q-learning was the Deep Q-Network (DQN) [13]. DQN uses a deep neural network for function approximation [87][4] (see Figure 2) and maintains an *experience replay* (ER) buffer [89, 90] to store interactions $\langle s, a, r, s' \rangle$. DQN keeps an additional copy of neural network parameters, θ^-, for the target network in addition to the θ parameters to stabilize the learning, i.e., to alleviate the non-stationary data distribution.[5] For each training iteration i, DQN minimizes the mean-squared error (MSE) between the Q-network and its target network using the loss function:

[4]Before DQN, many approaches used neural networks for representing the Q-value function [88], such as Neural Fitted Q-learning [87] and NEAT+Q [75].

[5]Double Q-learning [91] originally proposed keeping two Q functions (estimators) to reduce the overestimation bias in RL, while still keeping the convergence guarantees, later it was extended to DRL in Double DQN [92] (see Section 4.1).

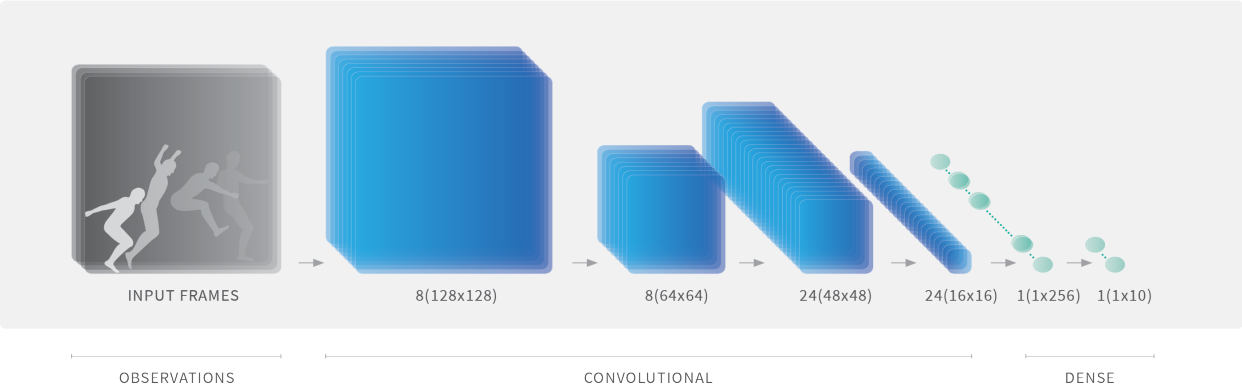


Figure 2: Deep Q-Network (DQN) [13]: Inputs are four stacked frames; the network is composed of several layers: *Convolutional* layers employ filters to learn features from high-dimensional data with a much smaller number of neurons and *Dense* layers are fully-connected layers. The last layer represents the actions the agent can take (in this case, 10 possible actions). Deep Recurrent Q-Network (DRQN) [85], which extends DQN to partially observable domains [42], is identical to this setup except the penultimate layer (1×256 Dense layer) is replaced with a recurrent LSTM layer [86].

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}[(r + \gamma max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i))^2] \tag{3}$$

where target network parameters θ^- are set to Q-network parameters θ periodically and mini-batches of $\langle s, a, r, s' \rangle$ tuples are sampled from the ER buffer, as depicted in Figure 3.

The ER buffer provides stability for learning as random batches sampled from the buffer helps alleviating the problems caused by the non-i.i.d. data. However, it comes with disadvantages, such as higher memory requirements and computation per real interaction [93]. The ER buffer is mainly used for off-policy RL methods as it can cause a mismatch between buffer content from earlier policy and from the current policy for on-policy methods [93]. Extending the ER buffer for the multiagent case is not trivial, see Sections 3.5, 4.1 and 4.2. Recent works were designed to reduce the problem of catastrophic forgetting (this occurs when the trained neural network performs poorly on previously learned tasks due to a non-stationary training distribution [94, 95]) and the ER buffer, in DRL [96] and MDRL [97].

DQN has been extended in many ways, for example, by using double estimators [91] to reduce the overestimation bias with Double DQN [92] (see Section 4.1) and by decomposing the Q-function with a *dueling*-DQN architecture [98], where two streams are learned, one estimates state values and another one advantages, those are combined in the final layer to form Q values (this method improved over Double DQN).

In practice, DQN is trained using an input of four stacked frames (last four frames the agent has encountered). If a game requires a memory of more than four frames it will appear non-Markovian to DQN because the future game states (and rewards) do not depend only on the input (four frames) but rather on the history [99]. Thus, DQN's performance declines when given incomplete state observations (e.g., one input frame) since DQN assumes full state observability.

Real-world tasks often feature incomplete and noisy state information resulting from *partial observability* (see Section 2.1). Deep Recurrent Q-Networks (DRQN) [85] proposed using *recurrent neural networks*, in particular, Long Short-Term Memory (LSTMs) cells [86] in DQN, for this setting. Consider the architecture in Figure 2 with the first dense layer after convolution replaced

Figure 3: Representation of a DQN agent that uses an experience replay buffer [89, 90] to keep $\langle s, a, r, s' \rangle$ tuples for minibatch updates. The Q-values are parameterized with a NN and a policy is obtained by selecting (greedily) over those at every timestep.

by a layer of LSTM cells. With this addition, DRQN has memory capacity so that it can even work with only one input frame rather than a stacked input of consecutive frames. This idea has been extended to MDRL, see Figure 6 and Section 4.2. There are also other approaches to deal with partial observability such as finite state controllers [100] (where action selection is performed according to the complete observation history) and using an initiation set of options conditioned on the previously employed option [101].

Policy gradient methods. For many tasks, particularly for physical control, the action space is continuous and high dimensional where DQN is not suitable. Deep Deterministic Policy Gradient (DDPG) [65] is a model-free off-policy actor-critic algorithm for such domains, based on the DPG algorithm [61] (see Section 2.1). Additionally, it proposes a new method for updating the networks, i.e., the target network parameters slowly change (this could also be applicable to DQN), in contrast to the hard reset (direct weight copy) used in DQN. Given the off-policy nature, DDPG generates exploratory behavior by adding sampled noise from some noise processes to its actor policy. The authors also used batch normalization [102] to ensure generalization across many different tasks without performing manual normalizations. However, note that other works have shown batch normalization can cause divergence in DRL [103, 104].

Asynchronous Advantage Actor-Critic (A3C) [93] is an algorithm that employs a *parallelized* asynchronous training scheme (using multiple CPU threads) for efficiency. It is an on-policy RL method that does not use an experience replay buffer. A3C allows multiple workers to simultaneously interact with the environment and compute gradients locally. All the workers pass their computed local gradients to a global NN which performs the optimization and synchronizes with the workers asynchronously (see Figure 4). There is also the Advantage Actor-Critic (A2C) method [105] that combines all the gradients from all the workers to update the global NN *synchronously*. The loss function for A3C is composed of two terms: policy loss (actor), $\mathcal{L}_\pi$, and value loss (critic), $\mathcal{L}_v$. A3C parameters are updated using the *advantage* function $A(s_t, a_t; \theta_v) = Q(s, a) - V(s)$, commonly used to reduce variance (see Section 2.1). An entropy loss for the policy, $H(\pi)$, is also commonly added, which helps to improve exploration by discouraging premature convergence to suboptimal deterministic policies [93]. Thus, the loss function is given by: $\mathcal{L}_{\text{A3C}} = \lambda_v \mathcal{L}_v + \lambda_\pi \mathcal{L}_\pi - \lambda_H \mathbb{E}_{s\sim\pi}[H(\pi(s, \cdot, \theta)]$ with λ_v, λ_π, and λ_H, being weighting terms on the individual loss components. Wang et al. [106] took A3C's framework but used off-policy learning

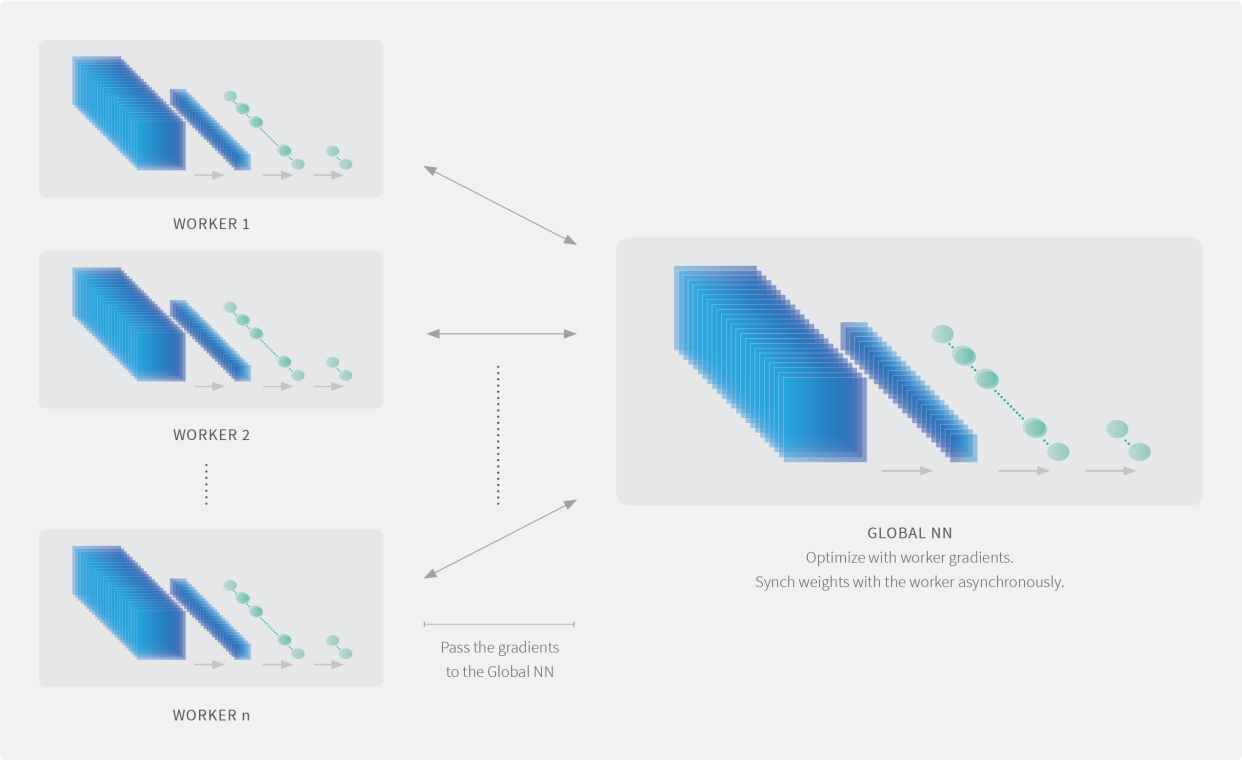


Figure 4: Asynchronous Advantage Actor-Critic (A3C) employs multiple (CPUs) workers without needing an ER buffer. Each worker has its own NN and independently interacts with the environment to compute the loss and gradients. Workers then pass computed gradients to the global NN that optimizes the parameters and synchronizes with the worker *asynchronously*. This distributed system is designed for single-agent deep RL. Compared to different DQN variants, A3C obtains better performance on a variety of Atari games using substantially less training time with multiple CPU cores of standard laptops without a GPU [93]. However, we note that more recent approaches use both multiple CPU cores for more efficient training data generation and GPUs for more efficient learning.

to create the Actor-critic with experience replay (ACER) algorithm. Gu et al. [107] introduced the Interpolated Policy Gradient (IPG) algorithm and showed a connection between ACER and DDPG: they are a pair of reparametrization terms (they are special cases of IPG) when they are put under the same stochastic policy setting, and when the policy is deterministic they collapse into DDPG.

Jaderberg et al. [84] built the Unsupervised Reinforcement and Auxiliary Learning (UNREAL) framework on top of A3C and introduced unsupervised *auxiliary tasks* (e.g., reward prediction) to speed up the learning process. Auxiliary tasks in general are not used for anything other than shaping the features of the agent, i.e., facilitating and regularizing the representation learning process [108, 109]; their formalization in RL is related to the concept of *general value functions* [20, 110]. The UNREAL framework optimizes a combined loss function $\mathcal{L}_{\text{UNREAL}} \approx \mathcal{L}_{\text{A3C}} + \sum_i \lambda_{AT_i} \mathcal{L}_{AT_i}$, that combines the A3C loss, $\mathcal{L}_{\text{A3C}}$, together with auxiliary task losses $\mathcal{L}_{AT_i}$, where λ_{AT_i} are weight terms (see Section 4.1 for use of auxiliary tasks in MDRL). In contrast to A3C, UNREAL uses a prioritized ER buffer, in which transitions with positive reward are given higher probability of being sampled. This approach can be viewed as a simple form of prioritized replay [111], which was in turn inspired by model-based RL algorithms like prioritized sweeping [112, 113].

Another distributed architecture is the Importance Weighted Actor-Learner Architecture (IM-

PALA) [114]. Unlike A3C or UNREAL, IMPALA actors communicate *trajectories of experience* (sequences of states, actions, and rewards) to a centralized learner, thus IMPALA decouples acting from learning.

Trust Region Policy Optimization (TRPO) [60] and Proximal Policy Optimization (PPO) [115] are recently proposed policy gradient algorithms where the latter represents the state-of-the art with advantages such as being simpler to implement and having better empirical sample complexity. Interestingly, a recent work [79] studying PPO and TRPO arrived at the surprising conclusion that these methods often deviate from what the theoretical framework would predict: gradient estimates are poorly correlated with the true gradient and value networks tend to produce inaccurate predictions for the true value function. Compared to vanilla policy gradient algorithms, PPO prevents abrupt changes in policies during training through the loss function, similar to early work by Kakade [116]. Another advantage of PPO is that it can be used in a distributed fashion, i.e, Distributed PPO (DPPO) [117]. Note that *distributed approaches* like DPPO or A3C use parallelization only to improve the learning by more efficient training data generation through multiple CPU cores for single agent DRL and they should not be considered multiagent approaches (except for recent work which tries to exploit this parallelization in a multiagent environment [118]).

Lastly, there's a connection between policy gradient algorithms and Q-learning [119] within the framework of entropy-regularized reinforcement learning [120] where the value and Q functions are slightly altered to consider the entropy of the policy. In this vein, Soft Actor-Critic (SAC) [121] is a recent algorithm that concurrently learns a stochastic policy, two Q-functions (taking inspiration from Double Q-learning) and a value function. SAC alternates between collecting experience with the current policy and updating from batches sampled from the ER buffer.

We have reviewed recent algorithms in DRL, while the list is not exhaustive, it provides an overview of the different state-of-art techniques and algorithms which will become useful while describing the MDRL techniques in the next section.

3. Multiagent Deep Reinforcement Learning (MDRL)

First, we briefly introduce the general framework on multiagent learning and then we dive into the categories and the research on MDRL.

3.1. Multiagent Learning

Learning in a multiagent environment is inherently more complex than in the single-agent case, as agents interact at the same time with environment and potentially with each other [5]. The *independent* learners, a.k.a. *decentralized* learners approach [122] directly uses single-agent algorithms in the multi-agent setting despite the underlying assumptions of these algorithms being violated (each agent independently learns its own policy, treating other agents as part of the environment). In particular the *Markov property* (the future dynamics, transitions, and rewards depend only on the current state) becomes invalid since the environment is no longer stationary [4, 6, 123]. This approach ignores the multiagent nature of the setting entirely and it can fail when an opponent adapts or learns, for example, based on the past history of interactions [2]. Despite the lack of guarantees, independent learners have been used in practice, providing advantages with regards to scalability while often achieving good results [8].

To understand why multiagent domains are non-stationary from agents' local perspectives, consider a simple stochastic (also known as Markov) game $(\mathcal{S}, \mathcal{N}, \mathcal{A}, \mathcal{T}, \mathcal{R})$, which can be seen as an extension of an MDP to multiple agents [124, 125]. One key distinction is that the transition,

$\mathcal{T}$, and reward function, $\mathcal{R}$, depend on the actions $\mathcal{A} = A_1 \times ... \times A_\mathcal{N}$ of all, $\mathcal{N}$, agents, this means, $\mathcal{R} = R_1 \times ... \times R_\mathcal{N}$ and $\mathcal{T} = \mathcal{S} \times A_1 \times ... \times A_\mathcal{N}$.

Given a learning agent i and using the common shorthand notation $-\boldsymbol{i} = \mathcal{N} \setminus \{i\}$ for the set of opponents, the value function now depends on the joint action $\boldsymbol{a} = (a_i, \boldsymbol{a_{-i}})$, and the joint policy $\boldsymbol{\pi}(s, \boldsymbol{a}) = \prod_j \pi_j(s, a_j)$:[6]

$$V_i^{\boldsymbol{\pi}}(s) = \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\pi}(s, \boldsymbol{a}) \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a_i, \boldsymbol{a_{-i}}, s')[R_i(s, a_i, \boldsymbol{a_{-i}}, s') + \gamma V_i(s')]. \tag{4}$$

Consequently, the optimal policy is dependent on the other agents' policies,

$$\begin{aligned} &\pi_i^*(s, a_i, \boldsymbol{\pi_{-i}}) = \arg\max_{\pi_i} V_i^{(\pi_i, \boldsymbol{\pi_{-i}})}(s) = \\ &\arg\max_{\pi_i} \sum_{\boldsymbol{a} \in \mathcal{A}} \pi_i(s, a_i) \boldsymbol{\pi_{-i}}(s, \boldsymbol{a_{-i}}) \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a_i, \boldsymbol{a_{-i}}, s')[R_i(s, a_i, \boldsymbol{a_{-i}}, s') + \gamma V_i^{(\pi_i, \boldsymbol{\pi_{-i}})}(s')]. \end{aligned} \tag{5}$$

Specifically, the opponents' joint policy $\boldsymbol{\pi_{-i}}(s, \boldsymbol{a_{-i}})$ can be non-stationary, i.e., changes as the opponents' policies change over time, for example with learning opponents.

Convergence results. Littman [125] studied convergence properties of reinforcement learning joint action agents [126] in Markov games with the following conclusions: in adversarial environments (zero-sum games) an optimal play can be guaranteed against an arbitrary opponent, i.e., Minimax Q-learning [124]. In coordination environments (e.g., in cooperative games all agents share the same reward function), strong assumptions need be made about other agents to guarantee convergence to optimal behavior [125], e.g., Nash Q-learning [127] and Friend-or-Foe Q-learning [128]. In other types of environments no value-based RL algorithms with guaranteed convergence properties are known [125].

Recent work on MDRL have addressed scalability and have focused significantly less on convergence guarantees, with few exceptions [129, 130, 131, 132]. One notable work has shown a connection between update rules for actor-critic algorithms for multiagent partially observable settings and (counterfactual) regret minimization:[7] the advantage values are scaled counterfactual regrets. This lead to new convergence properties of independent RL algorithms in zero-sum games with imperfect information [136]. The result is also used to support policy gradient optimization against worst-case opponents, in a new algorithm called Exploitability Descent [137].[8]

We refer the interested reader to seminal works about convergence in multiagent domains [139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149]. Note that instead of convergence, some MAL algorithms have proved learning a best response against classes of opponents [150, 151, 152].

There are other common problems in MAL, including action shadowing [34, 33], the curse of dimensionality [5], and multiagent credit assignment [32]. Describing each problem is out of the

[6] In this setting each agent independently executes a policy, however, there are other cases where this does not hold, for example when agents have a coordinated exploration strategy.

[7] Counterfactual regret minimization is a technique for solving large games based on regret minimization [133, 134] due to a well-known connection between regret and Nash equilibria [135]. It has been one of the reasons of successes in Poker [16, 17].

[8] This algorithm is similar to CFR-BR [138] and has the main advantage that the current policy convergences rather than the average policy, so there is no need to learn the average strategy, which requires large reservoir buffers or many past networks.

scope of this survey. However, we refer the interested reader to excellent resources on general MAL [4, 153, 154], as well as surveys in specific areas: game theory and multiagent reinforcement learning [5, 6], cooperative scenarios [7, 8], evolutionary dynamics of multiagent learning [9], learning in non-stationary environments [10], agents modeling agents [11], and transfer learning in multiagent RL [12].

3.2. MDRL categorization

In Section 2.2 we outlined some recent works in single-agent DRL since an exhaustive list is out of the scope of this article. This explosion of works has led DRL to be extended and combined with other techniques [23, 37, 29]. One natural extension to DRL is to test whether these approaches could be applied in a multiagent environment.

We analyzed the most recent works (that are not covered by previous MAL surveys [10, 11] and we do not consider genetic algorithms or swarm intelligence in this survey) that have a clear connection with MDRL. We propose 4 categories which take inspiration from previous surveys [1, 5, 7, 11] and that conveniently describe and represent current works. Note that some of these works fit into more than one category (they are not mutually exclusive), therefore their summaries are presented in all applicable Tables 1-4, however, for the ease of exposition when describing them in the text we only do so in one category. Additionally, for each work we present its learning type, either a value-based method (e.g., DQN) or a policy gradient method (e.g., actor-critic); also, we mention if the setting is evaluated in a fully cooperative, fully competitive or mixed environment (both cooperative and competitive).

- *Analysis of emergent behaviors.* These works, in general, do not propose learning algorithms — their main focus is to analyze and evaluate DRL algorithms, e.g., DQN [155, 156, 157], PPO [158, 157] and others [159, 157, 160], in a multiagent environment. In this category we found works which analyze behaviors in the three major settings: cooperative, competitive and mixed scenarios; see Section 3.3 and Table 1.

- *Learning communication* [161, 160, 162, 163, 164, 165]. These works explore a sub-area in which agents can share information with communication protocols, for example through direct messages [162] or via a shared memory [165]. This area is attracting attention and it had not been explored much in the MAL literature. See Section 3.4 and Table 2.

- *Learning cooperation.* While learning to communicate is an emerging area, fostering cooperation in learning agents has a long history of research in MAL [7, 8]. In this category the analyzed works are evaluated in either cooperative or mixed settings. Some works in this category take inspiration from MAL (e.g., leniency, hysteresis, and difference rewards concepts) and extend them to the MDRL setting [35, 166, 167]. A notable exception [168] takes a key component from RL (i.e., experience replay buffer) and adapts it for MDRL. See Section 3.5 and Table 3.

- *Agents modeling agents.* Albrecht and Stone [11] presented a thorough survey in this topic and we have found many works that fit into this category in the MDRL setting, some taking inspiration from DRL [169, 170, 171], and others from MAL [172, 173, 64, 174, 175]. Modeling agents is helpful not only to cooperate, but also for modeling opponents [172, 169, 171, 173], inferring goals [170], and accounting for the learning behavior of other agents [64]. In this category the analyzed algorithms present their results in either a competitive setting or a mixed one (cooperative and competitive). See Section 3.6 and Table 4.

In the rest of this section we describe each category along with the summaries of related works.

3.3. Emergent behaviors

Some recent works have analyzed the previously mentioned *independent* DRL agents (see Section 3.1) from the perspective of types of emerging behaviors (e.g., cooperative or competitive).

One of the earliest MDRL works is by Tampuu et al. [155], which had two independent DQN learning agents to play the Atari Pong game. Their focus was to adapt the reward function for the learning agents, which resulted in either cooperative or competitive emergent behaviors.

Leibo et al. [156] meanwhile studied independent DQNs in the context of *sequential social dilemmas*: a Markov game that satisfies certain inequalities [156]. The focus of this work was to highlight that cooperative or competitive behaviors exist not only as discrete (atomic) actions, but they are temporally extended (over policies). In the related setting of one shot Markov social dilemmas, Lerer and Peysakhovich [176] extended the famous Tit-for-Tat (TFT)[9] strategy [187] for DRL (using function approximators) and showed (theoretically and experimentally) that such agents can maintain cooperation. To construct the agents they used self-play and two reward schemes: selfish and cooperative. Previously, different MAL algorithms were designed to foster cooperation in social dilemmas with Q-learning agents [188, 189].

Self-play is a useful concept for learning algorithms (e.g., fictitious play [190]) since under certain classes of games it can guarantee convergence[10] and it has been used as a standard technique in previous RL and MAL works [192, 14, 193]. Despite its common usage self-play can be brittle to forgetting past knowledge [194, 172, 195] (see Section 4.5 for a note on the role of self-play as an open question in MDRL). To overcome this issue, Leibo et al. [159] proposed Malthusian reinforcement learning as an extension of self-play to population dynamics. The approach can be thought of as community coevolution and has been shown to produce better results (avoiding local optima) than independent agents with intrinsic motivation [196]. A limitation of this work is that it does not place itself within the state of the art in evolutionary and genetic algorithms. Evolutionary strategies have been employed for solving reinforcement learning problems [197] and for evolving function approximators [75]. Similarly, they have been used multiagent scenarios to compute approximate Nash equilibria [198] and as metaheuristic optimization algorithms [199, 200, 7, 201].

Bansal et al. [158] explored the emergent behaviors in competitive scenarios using the MuJoCo simulator [202]. They trained independent learning agents with PPO and incorporated two main modifications to deal with the MAL nature of the problem. First, they used *exploration rewards* [203] which are dense rewards that allow agents to learn basic (non-competitive) behaviors — this type of reward is annealed through time giving more weight to the environmental (competitive) reward. Exploration rewards come from early work in robotics [204] and single-agent RL [205], and their goal is to provide dense feedback for the learning algorithm to improve sample efficiency (Ng et al. [206] studied the theoretical conditions under which modifications of the reward function of an MDP preserve the optimal policy). For multiagent scenarios, these dense rewards help agents in the beginning phase of the training to learn basic non-competitive skills, increasing the probability of random actions from the agent yielding a positive reward. The second

[9] TFT originated in an iterated prisoner's dilemma tournament and later inspired different strategies in MAL [185], its generalization, Godfather, is a representative of *leader strategies* [186].

[10] The average strategy profile of fictitious players converges to a Nash equilibrium in certain classes of games, e.g., two-player zero-sum and potential games [191].

Table 1: These papers analyze *emergent behaviors* in MDRL. Learning type is either value-based (VB) or policy gradient (PG). Setting where experiments were performed: cooperative (CO), competitive (CMP) or mixed. A detailed description is given in Section 3.3.

Work	Summary	Learning	Setting
Tampuu et al. [155]	Train DQN agents to play Pong.	VB	CO&CMP
Leibo et al. [156]	Train DQN agents to play sequential social dilemmas.	VB	Mixed
Lerer and Peysakhovich [176]	Propose DRL agents able to cooperate in social dilemmas.	VB	Mixed
Leibo et al. [159]	Propose Malthusian reinforcement learning which extends self-play to population dynamics.	VB	Mixed
Bansal et al. [158]	Train PPO agents in competitive MuJoCo scenarios.	PG	CMP
Raghu et al. [157]	Train PPO, A3C, and DQN agents in attacker-defender games.	VB, PG	CMP
Lazaridou et al. [161]	Train agents represented with NN to learn a communication language.	PG	CO
Mordatch and Abbeel [160]	Learn communication with an end-to-end differentiable model to train with backpropagation.	PG	CO

Table 2: These papers propose algorithms for *learning communication*. Learning type is either value-based (VB) or policy gradient (PG). Setting were experiments were performed: cooperative (CO) or mixed. A more detailed description is given in Section 3.4.

Algorithm	Summary	Learning	Setting
Lazaridou et al. [161]	Train agents represented with NN to learn a communication language.	PG	CO
Mordatch and Abbeel [160]	Learn communication with an end-to-end differentiable model to train with backpropagation.	PG	CO
RIAL [162]	Use a single network (parameter sharing) to train agents that take environmental and communication actions.	VB	CO
DIAL [162]	Use gradient sharing during learning and communication actions during execution.	VB	CO
CommNet [163]	Use a continuous vector channel for communication on a single network.	PG	CO
BiCNet [164]	Use the actor-critic paradigm where communication occurs in the latent space.	PG	Mixed
MD-MADDPG [165]	Use of a shared memory as a means to multiagent communication.	PG	CO
MADDPG-MD [177]	Extend dropout technique to robustify communication when applied in multiagent scenarios with direct communication.	PG	CO

Table 3: These papers aim to *learn cooperation*. Learning type is either value-based (VB) or policy gradient (PG). Setting where experiments were performed: cooperative (CO), competitive (CMP) or mixed. A more detailed description is given in Section 3.5.

Algorithm	Summary	Learning	Setting
Lerer and Peysakhovich [176]	Propose DRL agents able to cooperate in social dilemmas.	VB	Mixed
MD-MADDPG [165]	Use of a shared memory as a means to multiagent communication.	PG	CO
MADDPG-MD [177]	Extend dropout technique to robustify communication when applied in multiagent scenarios with direct communication.	PG	CO
RIAL [162]	Use a single network (parameter sharing) to train agents that take environmental and communication actions.	VB	CO
DIAL [162]	Use gradient sharing during learning and communication actions during execution.	VB	CO
DCH/PSRO [172]	Policies can overfit to opponents: better compute approximate best responses to a mixture of policies.	VB	CO & CMP
Fingerprints [168]	Deal with ER problems in MDRL by conditioning the value function on a fingerprint that disambiguates the age of the sampled data.	VB	CO
Lenient-DQN [35]	Achieve cooperation by leniency, optimism in the value function by forgiving suboptimal (low-rewards) actions.	VB	CO
Hysteretic-DRQN [166]	Achieve cooperation by using two learning rates, depending on the updated values together with multitask learning via policy distillation.	VB	CO
WDDQN [178]	Achieve cooperation by leniency, weighted double estimators, and a modified prioritized experience replay buffer.	VB	CO
FTW [179]	Agents act in a mixed environment (composed of teammates and opponents), it proposes a two-level architecture and population-based learning.	PG	Mixed
VDN [180]	Decompose the team action-value function into pieces across agents, where the pieces can be easily added.	VB	Mixed
QMIX [181]	Decompose the team action-value function together with a mixing network that can recombine them.	VB	Mixed
COMA [167]	Use a centralized critic and a counter-factual advantage function based on solving the multiagent credit assignment.	PG	Mixed
PS-DQN, PS-TRPO, PS-A3C [182]	Propose parameter sharing for learning cooperative tasks.	VB, PG	CO
MADDPG [63]	Use an actor-critic approach where the critic is augmented with information from other agents, the actions of all agents.	PG	Mixed

Table 4: These papers consider *agents modeling agents*. Learning type is either value-based (VB) or policy gradient (PG). Setting where experiments were performed: cooperative (CO), competitive (CMP) or mixed. A more detailed description is given in Section 3.6.

Algorithm	Summary	Learning	Setting
MADDPG [63]	Use an actor-critic approach where the critic is augmented with information from other agents, the actions of all agents.	PG	Mixed
DRON [169]	Have a network to infer the opponent behavior together with the standard DQN architecture.	VB	Mixed
DPIQN, DPIRQN [171]	Learn policy features from raw observations that represent high-level opponent behaviors via auxiliary tasks.	VB	Mixed
SOM [170]	Assume the reward function depends on a hidden goal of both agents and then use an agent's own policy to infer the goal of the other agent.	PG	Mixed
NFSP [173]	Compute approximate Nash equilibria via self-play and two neural networks.	VB	CMP
PSRO/DCH [172]	Policies can overfit to opponents: better compute approximate best responses to a mixture of policies.	PG	CO & CMP
M3DDPG [183]	Extend MADDPG with minimax objective to robustify the learned policy.	PG	Mixed
LOLA [64]	Use a learning rule where the agent accounts for the parameter update of other agents to maximize its own reward.	PG	Mixed
ToMnet [174]	Use an architecture for end-to-end learning and inference of diverse opponent types.	PG	Mixed
Deep Bayes-ToMoP [175]	Best respond to opponents using Bayesian policy reuse, theory of mind, and deep networks.	VB	CMP
Deep BPR+[184]	Bayesian policy reuse and policy distillation to quickly best respond to opponents.	VB	CO & CMP

contribution was *opponent sampling* which maintains a pool of older versions of the opponent to sample from, in contrast to using the most recent version.

Raghu et al. [157] investigated how DRL algorithms (DQN, A2C, and PPO) performed in a family of two-player zero-sum games with tunable complexity, called Erdos-Selfridge-Spencer games [207, 208]. Their reasoning is threefold: (i) these games provide a parameterized family of environments where (ii) optimal behavior can be completely characterized, and (iii) support multiagent play. Their work showed that algorithms can exhibit wide variation in performance as the algorithms are tuned to the game's difficulty.

Lazaridou et al. [161] proposed a framework for language learning that relies on multiagent communication. The agents, represented by (feed-forward) neural networks, need to develop an *emergent language* to solve a task. The task is formalized as a *signaling game* [209] in which two agents, a sender and a receiver, obtain a pair of images. The sender is told one of them is the target and is allowed to send a message (from a fixed vocabulary) to the receiver. Only when the receiver identifies the target image do both agents receive a positive reward. The results show that agents can coordinate for the experimented visual-based domain. To analyze the semantic properties[11] of the learned communication protocol they looked whether symbol usage reflects the semantics of the visual space, and that despite some variation, many high level objects groups correspond to the same learned symbols using a t-SNE [210] based analysis (t-SNE is a visualization technique for high-dimensional data and it has also been used to better understand the behavior of trained DRL agents [211, 212]). A key objective of this work was to determine if the agent's language could be human-interpretable. To achieve this, learned symbols were grounded with natural language by extending the signaling game with a supervised image labelling task (the sender will be encouraged to use conventional names, making communication more transparent to humans). To measure the interpretability of the extended game, a crowdsourced survey was performed, and in essence, the trained agent receiver was replaced with a human. The results showed that 68% of the cases, human participants picked the correct image.

Similarly, Mordatch and Abbeel [160] investigated the emergence of language with the difference that in their setting there were no explicit roles for the agents (i.e., sender or receiver). To learn, they proposed an end-to-end differentiable model of all agent and environment state dynamics over time to calculate the gradient of the return with backpropagation.

3.4. Learning communication

As we discussed in the previous section, one of the desired emergent behaviors of multiagent interaction is the emergence of communication [161, 160]. This setting usually considers a set of *cooperative agents* in a *partially observable* environment (see Section 2.2) where agents need to maximize their shared utility by means of communicating information.

Reinforced Inter-Agent Learning (RIAL) and Differentiable Inter-Agent Learning (DIAL) are two methods using deep networks to learn to communicate [162]. Both methods use a neural net that outputs the agent's Q values (as done in standard DRL algorithms) and a message to communicate to other agents in the next timestep. RIAL is based on DRQN and also uses the concept of *parameter sharing*, i.e., using a single network whose parameters are shared among all agents. In contrast, DIAL directly passes gradients via the communication channel during learning, and messages are discretized and mapped to the set of communication actions during execution.

[11] The vocabulary that agents use was arbitrary and had no initial meaning. To understand its emerging semantics they looked at the relationship between symbols and the sets of images they referred to [161].

Memory-driven (MD) communication was proposed on top of the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [63] method. In MD-MADDPG [165], the agents use a shared memory as a communication channel: before taking an action, the agent first reads the memory, then writes a response. In this case the agent's policy becomes dependent on its private observation and its interpretation of the collective memory. Experiments were performed with two agents in cooperative scenarios. The results highlighted the fact that the communication channel was used differently in each environment, e.g., in simpler tasks agents significantly decrease their memory activity near the end of the task as there are no more changes in the environment; in more complex environments, the changes in memory usage appear at a much higher frequency due to the presence of many sub-tasks.

Dropout [213] is a technique to prevent overfitting (in supervised learning this happens when the learning algorithm achieves good performance only on a specific data set and fails to generalize) in neural networks which is based on randomly dropping units and their connections during training time. Inspired by dropout, Kim et al. [177] proposed a similar approach in multiagent environments where direct communication through messages is allowed. In this case, the messages of other agents are dropped out at training time, thus the authors proposed the Message-Dropout MADDPG algorithm [177]. This method is expected to work in fully or limited communication environments. The empirical results show that with properly chosen message dropout rate, the proposed method both significantly improves the training speed and the robustness of learned policies (by introducing communication errors) during execution time. This capability is important as MDRL agents trained in simulated or controlled environments will be less fragile when transferred to more realistic environments.

While RIAL and DIAL used a discrete communication channel, CommNet [163] used a continuous vector channel. Through this channel agents receive the summed transmissions of other agents. The authors assume full cooperation and train a single network for all the agents. There are two distinctive characteristics of CommNet from previous works: it allows multiple communication cycles at each timestep and a dynamic variation of agents at run time, i.e., agents come and go in the environment.

In contrast to previous approaches, in Multiagent Bidirectionally Coordinated Network (BiCNet) [164], communication takes place in the latent space (i.e., in the hidden layers). It also uses parameter sharing, however, it proposes bidirectional recurrent neural networks [214] to model the actor and critic networks of their model. Note that in BiCNet agents do not *explicitly* share a message and thus it can be considered a method for learning cooperation.

Learning communication is an active area in MDRL with many open questions, in this context, we refer the interested reader to a recent work by Lowe et al. [215] where it discusses common pitfalls (and recommendations to avoid those) while measuring communication in multiagent environments.

3.5. Learning cooperation

Although *explicit communication* is a new emerging trend in MDRL, there has already been a large amount of work in MAL for cooperative settings[12] that do not involve communication [7, 8]. Therefore, it was a natural starting point for many recent MDRL works.

Foerster et al. [168] studied the simple scenario of *cooperation with independent Q-learning agents* (see Section 3.1), where the agents use the standard DQN architecture of neural networks

[12]There is a large body of research on coordinating multiagent teams by specifying communication protocols [216, 217]: these expect agents to know the team's goal as well as the tasks required to accomplish the goal.

and an experience replay buffer (see Figure 3). However, for the ER to work, the data distribution needs to follow certain assumptions (see Section 2.2) which are no loger valid due to the multiagent nature of the world: the dynamics that generated the data in the ER no longer reflect the current dynamics, making the experience obsolete [168, 90]. Their solution is to add information to the experience tuple that can help to *disambiguate the age of the sampled data* from the replay memory. Two approaches were proposed. The first is Multiagent Importance Sampling which adds the probability of the joint action so an importance sampling correction [70, 218] can computed when the tuple is later sampled for training. This was similar to previous works in adaptive importance sampling [219, 220] and off-environment RL [221]. The second approach is Multiagent Fingerprints which adds the estimate (i.e., fingerprint) of other agents' policies (loosely inspired by Hyper-Q [150], see Section 4.1). For the practical implementation, good results were obtained by using the training iteration number and exploration rate as the fingerprint.

Gupta et al. [182] tackled cooperative environments in partially observable domains without explicit communication. They proposed *parameter sharing* (PS) as a way to improve learning in homogeneous multiagent environments (where agents have the same set of actions). The idea is to have one globally shared learning network that can still behave differently in execution time, i.e., because its inputs (individual agent observation and agent index) will be different. They tested three variations of this approach with parameter sharing: PS-DQN, PS-DDPG and PS-TRPO, which extended single-agent DQN, DDPG and TRPO algorithms, respectively. The results showed that PS-TRPO outperformed the other two. Note that Foerster et al. [162] concurrently proposed a similar concept, see Section 3.4.

Lenient-DQN (LDQN) [35] took the *leniency* concept [222] (originally presented in MAL) and extended their use to MDRL. The purpose of leniency is to overcome a pathology called relative overgeneralization [34, 223, 224]. Similar to other approaches designed to overcome relative overgeneralization (e.g., distributed Q-learning [225] and hysteretic Q-learning [8]) lenient learners initially maintain an optimistic disposition to mitigate the noise from transitions resulting in miscoordination, preventing agents from being drawn towards sub-optimal but wide peaks in the reward search space [97]. However, similar to other MDRL works [168], the LDQN authors experienced problems with the ER buffer and arrived at a similar solution: adding information to the experience tuple, in their case, the leniency value. When sampling from the ER buffer, this value is used to determine a leniency condition; if the condition is not met then the sample is ignored.

In a similar vein, Decentralized-Hysteretic Deep Recurrent Q-Networks (DEC-HDRQNs) [166] were proposed for fostering cooperation among independent learners. The motivation is similar to LDQN, making an optimistic value update, however, their solution is different. Here, the authors took inspiration from Hysteretic Q-learning [8], originally presented in MAL, where two learning rates were used. A difference between lenient agents and hysteretic Q-learning is that lenient agents are only *initially* forgiving towards teammates. Lenient learners over time apply less leniency towards updates that would lower utility values, taking into account how frequently observation-action pairs have been encountered. The idea being that the transition from optimistic to average reward learner will help make lenient learners more robust towards misleading stochastic rewards [222]. Additionally, in DEC-HDRQNs the ER buffer is also extended into *concurrent experience replay trajectories*, which are composed of three dimensions: agent index, the episode, and the timestep; when training, the sampled traces have the same starting timesteps. Moreover, to improve on generalization over different tasks, i.e., multi-task learning[226], DEC-HDRQNs make use of policy distillation [227, 228] (see Section 4.1). In contrast to other approaches, DEC-

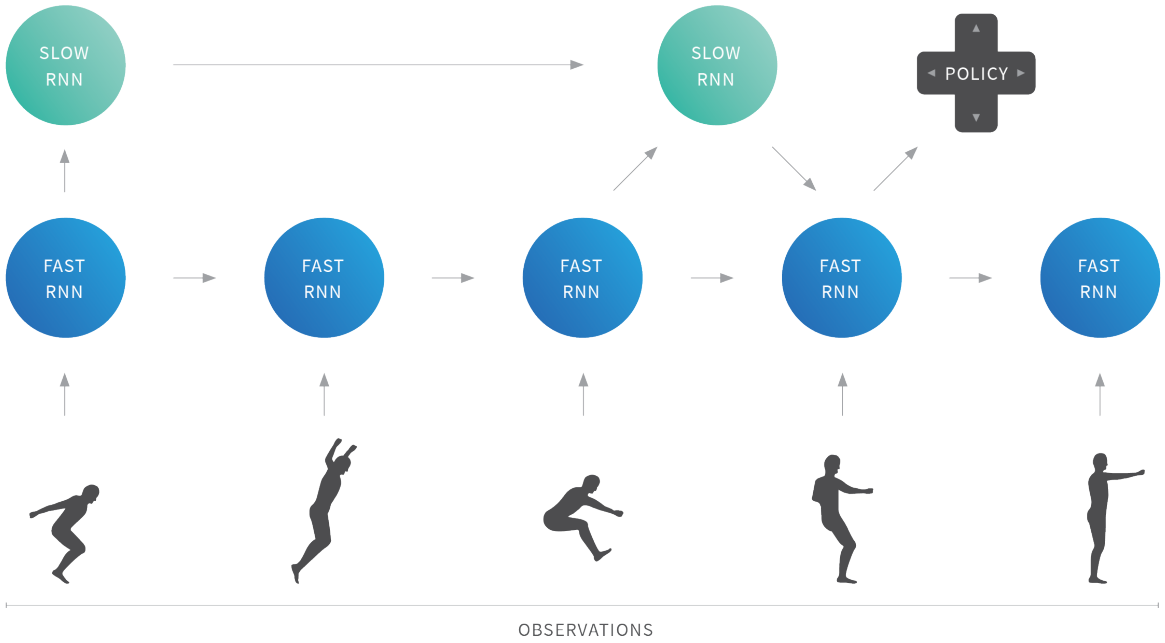


Figure 5: A schematic view of the architecture used in FTW (For the Win) [179]: two unrolled recurrent neural networks (RNNs) operate at different time-scales, the idea is that the *Slow RNN* helps with long term temporal correlations. Observations are latent space output of some convolutional neural network to learn non-linear features. Feudal Networks [229] is another work in single-agent DRL that also maintains a multi-time scale hierarchy where the slower network sets the goal, and the faster network tries to achieve them. Fedual Networks were in turn, inspired by early work in RL which proposed a hierarchy of Q-learners [230, 231].

HDRQNS are fully decentralized during learning and execution.

Weighted Double Deep Q-Network (WDDQN) [178] is based on having double estimators. This idea was originally introduced in Double Q-learning [91] and aims to remove the existing overestimation bias caused by using the maximum action value as an approximation for the maximum expected action value (see Section 4.1). It also uses a *lenient* reward [222] to be optimistic during initial phase of coordination and proposes a *scheduled* replay strategy in which samples closer to the terminal states are heuristically given higher priority; this strategy might not be applicable for any domain. For other works extending the ER to multiagent settings see MADDPG [63], Sections 4.1 and 4.2.

While previous approaches were mostly inspired by how MAL algorithms could be extended to MDRL, other works take as base the results by single-agent DRL. One example is the For The Win (FTW) [179] agent which is based on the actor-learner structure of IMPALA [114] (see Section 2.2). The authors test FTW in a game where two opposing teams compete to capture each other's flags [232]. To deal with the MAL problem they propose two main additions: a *hierarchical two-level representation* with recurrent neural networks operating at different timescales, as depicted in Figure 5, and a *population based training* [233, 234, 235] where 30 agents were trained in parallel together with a stochastic matchmaking scheme that biases agents to be of similar *skills*.

The Elo rating system [236] was originally devised to rate chess player skills,[13] TrueSkill [237] extended Elo by tracking uncertainty in skill rating, supporting draws, and matches beyond 1 vs 1; $\alpha-$Rank is a more recent alternative to ELO [238]. FTW did not use TrueSkill but a simpler extension of Elo for n vs n games (by adding individual agent ratings to compute the team skill). Hierarchical approaches were previously proposed in RL, e.g., Feudal RL [230, 231], and were later extended to DRL in Feudal networks [229]; population based training can be considered analogous to evolutionary strategies that employ self-adaptive hyperparameter tuning to modify how the genetic algorithm itself operates [234, 239, 240]. An interesting result from FTW is that the population-based training obtained better results than training via self-play [192], which was a standard concept in previous works [14, 193]. FTW used heavy compute resources, it used 30 agents (processes) in parallel where every training game lasted 4500 agent steps ($\approx$ five minutes) and agents were trained for two billion steps ($\approx$ 450K games).

Lowe et al. [63] noted that using standard policy gradient methods (see Section 2.1) on multiagent environments yields high variance and performs poorly. This occurs because the variance is further increased as all the agents' rewards depend on the rest of the agents, and it is formally shown that as the number of agents increase, the probability of taking a correct gradient direction decreases exponentially [63]. Therefore, to overcome this issue Lowe et al. proposed the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [63], building on DDPG [65] (see Section 2.2), to train a centralized critic per agent that is given all agents' policies during training to reduce the variance by removing the non-stationarity caused by the concurrently learning agents. Here, the actor only has local information (turning the method into a centralized training with decentralized execution) and the ER buffer records experiences of *all* agents. MADDPG was tested in both cooperative and competitive scenarios, experimental results show that it performs better than several decentralized methods (such as DQN, DDPG, and TRPO). The authors mention that traditional RL methods do not produce consistent gradient signals. This is exemplified in a challenging competitive scenarios where agents continuously adapt to each other causing the learned best-response policies oscillate — for such a domain, MADDPG is shown to learn more robustly than DDPG.

Another approach based on policy gradients is the Counterfactual Multi-Agent Policy Gradients (COMA) [167]. COMA was designed for the fully centralized setting and the *multiagent credit assignment problem* [241], i.e., how the agents should deduce their contributions when learning in a cooperative setting in the presence of only global rewards. Their proposal is to compute a *counterfactual baseline*, that is, marginalize out the action of the agent while keeping the rest of the other agents' actions fixed. Then, an advantage function can be computed comparing the current Q value to the counterfactual. This counterfactual baseline has its roots in *difference rewards*, which is a method for obtaining the individual contribution of an agent in a cooperative multiagent team [241]. In particular, the *aristocrat* utility aims to measure the difference between an agent's actual action and the average action [31]. The intention would be equivalent to sideline the agent by having the agent perform an action where the reward does not depend on the agent's actions, i.e., to consider the reward that would have arisen assuming a world without that agent having ever existed (see Section 4.2).

On the one hand, fully centralized approaches (e.g., COMA) do not suffer from non-stationarity

[13] Elo uses a normal distribution for each player skill, and after each match, both players' distributions are updated based on measure of *surprise*, i.e., if a user with previously lower (predicted) skill beats a high skilled one, the low-skilled player is significantly increased.

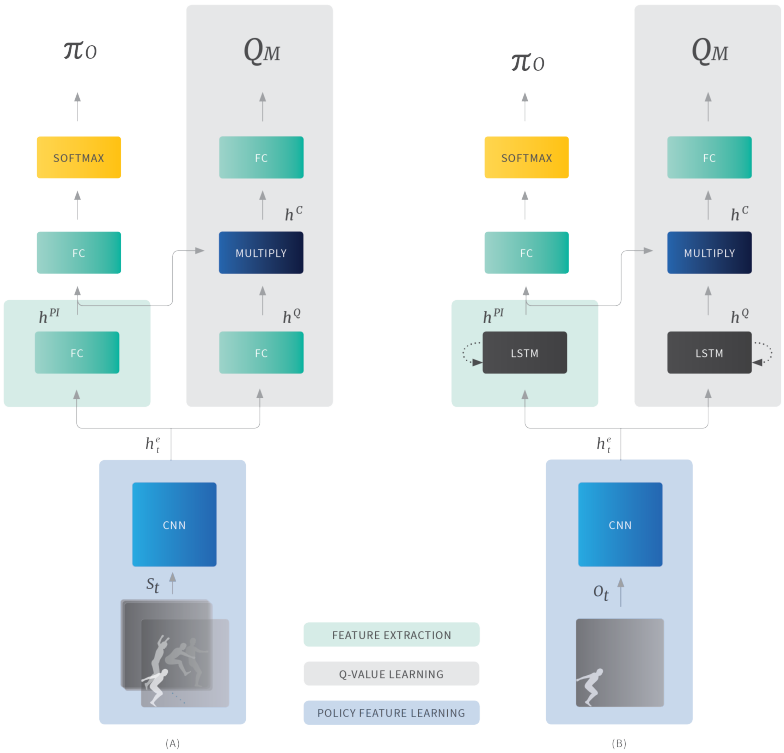


Figure 6: (a) Deep Policy Inference Q-Network: receives four stacked frames as input (similar to DQN, see Figure 2). (b) Deep Policy Inference Recurrent Q-Network: receives one frame as input and has an LSTM layer instead of a fully connected layer (FC). Both approaches [171] condition the Q_M value outputs on the policy features, h^{PI}, which are also used to learn the opponent policy π_o.

but have constrained scalability. On the other hand, independent learning agents are better suited to scale but suffer from non-stationarity issues. There are some hybrid approaches that learn *a centralized but factored* Q value function [242, 243]. Value Decomposition Networks (VDNs) [180] decompose a team value function into an additive decomposition of the individual value functions. Similarly, QMIX [181] relies on the idea of factorizing, however, instead of sum, QMIX assumes a *mixing network* that combines the local values in a non-linear way, which can represent monotonic action-value functions. While the mentioned approaches have obtained good empirical results, the factorization of value-functions in multiagent scenarios using function approximators (MDRL) is an ongoing research topic, with open questions such as how well factorizations capture complex coordination problems and how to learn those factorizations [244] (see Section 4.4).

3.6. Agents modeling agents

An important ability for agents to have is to reason about the behaviors of other agents by constructing models that make predictions about the modeled agents [11]. An early work for modeling agents while using deep neural networks was the Deep Reinforcement Opponent Network (DRON) [169]. The idea is to have two networks: one which evaluates Q-values and a second one that *learns a representation of the opponent's policy*. Moreover, the authors proposed to have several expert networks to combine their predictions to get the estimated Q value, the idea being that each expert network captures one type of opponent strategy [245]. This is related to previous

works in type-based reasoning from game theory [246, 139] later applied in AI [245, 11, 247]. The mixture of experts idea was presented in supervised learning where each expert handled a subset of the data (a subtask), and then a gating network decided which of the experts should be used [248].

DRON uses hand-crafted features to define the opponent network. In contrast, Deep Policy Inference Q-Network (DPIQN) and its recurrent version, DPIRQN [171] learn *policy features* directly from raw observations of the other agents. The way to learn these policy features is by means of auxiliary tasks [84, 110] (see Sections 2.2 and 4.1) that provide additional learning goals, in this case, the auxiliary task is to learn the opponents' policies. This auxiliary task modifies the loss function by computing an auxiliary loss: the cross entropy loss between the inferred opponent policy and the ground truth (one-hot action vector) of the opponent. Then, the Q value function of the learning agent is conditioned on the opponent's policy features (see Figure 6), which aims to reduce the non-stationarity of the environment. The authors used an adaptive training procedure to adjust the attention (a weight on the loss function) to either emphasize learning the policy features (of the opponent) or the respective Q values of the agent. An advantage of these approaches is that modeling the agents can work for both opponents and teammates [171].

In many previous works an opponent model is learned from observations. Self Other Modeling (SOM) [170] proposed a different approach, this is, using *the agent's own policy as a means to predict the opponent's actions*. SOM can be used in cooperative and competitive settings (with an arbitrary number of agents) and infers other agents' goals. This is important because in the evaluated domains, the reward function depends on the goal of the agents. SOM uses two networks, one used for computing the agents' own policy, and a second one used to infer the opponent's goal. The idea is that these networks have the same input parameters but with different values (the agent's or the opponent's). In contrast to previous approaches, SOM is not focused on learning the opponent policy, i.e., a probability distribution over next actions, but rather on estimating the opponent's goal. SOM is expected to work best when agents share a set of goals from which each agent gets assigned one at the beginning of the episode and the reward structure depends on both of their assigned goals. Despite its simplicity, training takes longer as an additional optimization step is performed given the other agent's observed actions.

There is a long-standing history of combining game theory and MAL [2, 6, 193]. From that context, some approaches were inspired by influential game theory approaches. Neural Fictitious Self-Play (NFSP) [173] builds on fictitious (self-) play [190, 249], together with two deep networks to find *approximate Nash equilibria*[14] in two-player imperfect information games [251] (for example, consider Poker: when it is an agent's turn to move it does not have access to all information about the world). One network learns an *approximate best response* ($\epsilon-$greedy over Q values) to the historical behavior of other agents and the second one (called the average network) learns to imitate its own past best response behaviour using supervised classification. The agent behaves using a mixture of the average and the best response networks depending on the probability of an anticipatory parameter [252]. Comparisons with DQN in Leduc Holdem Poker revealed that DQN's deterministic strategy is highly exploitable. Such strategies are sufficient to behave optimally in single-agent domains, i.e., MDPs for which DQN was designed. However, imperfect-information games generally require stochastic strategies to achieve optimal behaviour [173]. DQN learning experiences are both highly correlated over time, and highly focused on a narrow state distribution.

[14] Nash equilibrium [250] is a solution concept in game theory in which no agent would choose to deviate from its strategy (they are a best response to others' strategies). This concept has been explored in seminal MAL algorithms like Nash-Q learning [127] and Minimax-Q learning [124, 128].

In contrast to NFSP agents whose experience varies more smoothly, resulting in a more stable data distribution, more stable neural networks and better performance.

The (N)FSP concept was further generalized in Policy-Space Response Oracles (PSRO) [172], where it was shown that fictitious play is one specific meta-strategy distribution over a set of previous (approximate) best responses (summarized by a meta-game obtained by empirical game theoretic analysis [253]), but there are a wide variety to choose from. One reason to use mixed meta-strategies is that it prevents overfitting[15] the responses to one specific policy, and hence provides a form of opponent/teammate regularization. An approximate scalable version of the algorithm leads to a graph of agents best-responding independently called Deep Cognitive Hierarchies (DCHs) [172] due to its similarity to behavioral game-theoretic models [255, 256].

Minimax is a paramount concept in game theory that is roughly described as minimizing the worst case scenario (maximum loss) [251]. Li et al. [183] took the minimax idea as an approach to robustify learning in multiagent environments so that the learned robust policy should be able to behave well even with strategies not seen during training. They extended the MADDPG algorithm [63] to Minimax Multiagent Deep Deterministic Policy Gradients (M3DDPG), which updates policies considering a worst-case scenario: assuming that all other agents act adversarially. This yields a minimax learning objective which is computationally intractable to directly optimize. They address this issue by taking ideas from robust reinforcement learning [257] which implicitly adopts the minimax idea by using the *worst noise* concept [258]. In MAL different approaches were proposed to assess the robustness of an algorithm, e.g., guarantees of safety [152, 259], security [260] or exploitability [261, 262, 263].

Previous approaches usually learned a model of the other agents as a way to predict their behavior. However, they do not explicitly *account for anticipated learning of the other agents*, which is the objective of Learning with Opponent-Learning Awareness (LOLA) [64]. LOLA optimizes the expected return *after the opponent updates its policy one step*. Therefore, a LOLA agent directly shapes the policy updates of other agents to maximize its own reward. One of LOLA's assumptions is having access to opponents' policy parameters. LOLA builds on previous ideas by Zhang and Lesser [264] where the learning agent predicts the opponent's policy parameter update but only uses it to learn a best response (to the anticipated updated parameters).

Theory of mind is part of a group of *recursive reasoning* approaches[265, 245, 266, 267] in which agents have explicit beliefs about the mental states of other agents. The mental states of other agents may, in turn, also contain beliefs and mental states of other agents, leading to a nesting of beliefs [11]. Theory of Mind Network (ToMnet) [174] starts with a simple premise: when encountering a novel opponent, *the agent should already have a strong and rich prior about how the opponent should behave*. ToMnet has an architecture composed of three networks: (i) a character network that learns from historical information, (ii) a mental state network that takes the character output and the recent trajectory, and (iii) the prediction network that takes the current state as well as the outputs of the other networks as its input. The output of the architecture is open for different problems but in general its goal is to predict the opponent's next action. A main advantage of ToMnet is that it can predict general behavior, for all agents; or specific, for a particular agent.

Deep Bayesian Theory of Mind Policy (Bayes-ToMoP) [175] is another algorithm that takes inspiration from theory of mind [268]. The algorithm assumes the opponent has different stationary

[15] Johanson et al. [254] also found "overfitting" when solving large extensive games (e.g., poker) — the performance in an abstract game improved but it was worse in the full game.

strategies to act and changes among them over time [269]. Earlier work in MAL dealt with this setting, e.g., BPR+ [270] extends the Bayesian policy reuse[16] framework [271] to multiagent settings (BPR assumes a single-agent environment; BPR+ aims to best respond to the opponent in a multiagent game). A limitation of BPR+ is that it behaves poorly against itself (self-play), thus, Deep Bayes-ToMoP uses theory of mind to provide a higher-level reasoning strategy which provides an optimal behavior against BPR+ agents.

Deep BPR+ [184] is another work inspired by BPR+ which uses neural networks as value-function approximators. It not only uses the environment reward but also uses the online learned opponent model [272, 273] to construct a rectified belief over the opponent strategy. Additionally, it leverages ideas from policy distillation [227, 228] and extends them to the multiagent case to create a distilled policy network. In this case, whenever a new acting policy is learned, distillation is applied to consolidate the new updated library which improves in terms of storage and generalization (over opponents).

4. Bridging RL, MAL and MDRL

This section aims to provide directions to promote fruitful cooperations between sub-communities. First, we address the pitfall of *deep learning amnesia*, roughly described as missing citations to the original works and not exploiting the advancements that have been made in the past. We present examples on how ideas originated earlier, for example in RL and MAL, were successfully extended to MDRL (see Section 4.1). Second, we outline *lessons learned* from the works analyzed in this survey (see Section 4.2). Then we point the readers to recent benchmarks for MDRL (see Section 4.3) and we discuss the practical challenges that arise in MDRL like high computational demands and reproducibility (see Section 4.4). Lastly, we pose some open research challenges and reflect on their relation with previous open questions in MAL [11] (see Section 4.5).

4.1. Avoiding deep learning amnesia: examples in MDRL

This survey focuses on recent *deep* works, however, in previous sections, when describing recent algorithms, we also point to original works that inspired them. Schmidhuber said "Machine learning is the science of credit assignment. The machine learning community itself profits from proper credit assignment to its members" [274]. In this context, we want to avoid committing the pitfall of not giving credit to original ideas that were proposed earlier, a.k.a. *deep learning amnesia*. Here, we provide some specific examples of research milestones that were studied earlier, e.g., RL or MAL, and that now became highly relevant for MDRL. Our purpose is to highlight that existent literature contains pertinent ideas and algorithms that should not be ignored. On the contrary, they should be examined and cited [275, 276] to understand recent developments [277].

Dealing with non-stationarity in independent learners. It is well known that using independent learners makes the environment non-stationary from the agent's point of view [4, 123]. Many MAL algorithms tried to solve this problem in different ways [10]. One example is *Hyper-Q* [150] which accounts for the (values of mixed) strategies of other agents and includes that information in the state representation, which effectively turns the learning problem into a stationary one. Note that in

[16] Bayesian policy reuse assumes an agent with prior experience in the form of a library of policies. When a novel task instance occurs, the objective is to reuse a policy from its library based on observed signals which correlate to policy performance [271].

this way it is possible to even consider adaptive agents. Foerster et al. [162] make use of this insight to propose their *fingerprint* algorithm in an MDRL problem (see Section 3.5). Other examples include the leniency concept [222] and Hysteretic Q-learning [8] originally presented in MAL, which now have their "deep" counterparts, LDQNs [35] and DEC-HDRQNs[166], see Section 3.5.

Multiagent credit assignment. In cooperative multiagent scenarios, it is common to use either *local rewards*, unique for each agent, or *global rewards*, which represent the entire group's performance [278]. However, local rewards are usually harder to obtain, therefore, it is common to rely only on the global ones. This raises the problem of *credit assignment*: how does a single agent's actions contribute to a system that involves the actions of many agents [32]. A solution that came from MAL research that has proven successful in many scenarios is *difference rewards* [241, 278, 279], which aims to capture an agent's contribution to the system's global performance. In particular the *aristocrat* utility aims to measure the difference between an agents actual action and the average action [31], however, it has a self-consistency problem and in practice it is more common to compute the *wonderful life utility* [280, 31], which proposes to use a clamping operation that would be equivalent to removing that player from the team. COMA [167] builds on these concepts to propose an *advantage function* based on the contribution of the agent, which can be efficiently computed with deep neural networks (see Section 3.5).

Multitask learning. In the context of RL, multitask learning [226] is an area that develops agents that can act in *several related tasks* rather than just in a single one [281]. *Distillation*, roughly defined as transferring the knowledge from a large model to a small model, was a concept originally introduced for supervised learning and model compression [282, 228]. Inspired by those works, Policy distillation [227] was extended to the DRL realm. Policy distillation was used to train a much smaller network and to merge *several task-specific policies* into a single policy, i.e., for multitask learning. In the MDRL setting, Omidshafiei et al. [166] successfully adapted policy distillation within Dec-HDRQNs to obtain a more general multitask multiagent network (see Section 3.5). Another example is Deep BPR+ [184] which uses distillation to generalize over multiple opponents (see Section 3.6).

Auxiliary tasks. Jaderberg et al. [84] introduced the term auxiliary task with the insight that (single-agent) environments contain a variety of possible training signals (e.g., pixel changes). These tasks are naturally implemented in DRL in which the last layer is split into multiple parts (heads), each working on a different task. All heads propagate errors into the same shared preceding part of the network, which would then try to form representations, in its next-to-last layer, to support all the heads [20]. However, the idea of multiple predictions about arbitrary signals was originally suggested for RL, in the context of general value functions [110, 20] and there still open problems, for example, better theoretical understanding [109, 283]. In the context of neural networks, early work proposed *hints* that improved the network performance and learning time. Suddarth and Kergosien [284] presented a minimal example of a small neural network where it was shown that adding an auxiliary task effectively removed local minima. One could think of extending these auxiliary tasks to modeling other agents' behaviors [285, 160], which is one of the key ideas that DPIQN and DRPIQN [171] proposed in MDRL settings (see Section 3.6).

Experience replay. Lin [90, 89] proposed the concept of experience replay to speed up the credit assignment propagation process in single agent RL. This concept became central to many DRL works [72] (see Section 2.2). However, Lin stated that a condition for the ER to be useful is

that "the environment should not change over time because this makes past experiences irrelevant or even harmful" [90]. This is a problem in domains where many agents are learning since the environment becomes non-stationary from the point of view of each agent. Since DRL relies heavily on experience replay, this is an issue in MDRL: the non-stationarity introduced means that the dynamics that generated the data in the agent's replay memory no longer reflect the current dynamics in which it is learning [162]. To overcome this problem different methods have been proposed [168, 35, 166, 178], see Section 4.2.

Double estimators. Double Q-learning [91] proposed to reduce the overestimation of action values in Q-learning, this is caused by using the maximum action value as an approximation for the maximum expected action value. Double Q-learning works by keeping two Q functions and was proven to convergence to the optimal policy [91]. Later this idea was applied to arbitrary function approximators, including deep neural networks, i.e., Double DQN [92], which were naturally applied since two networks were already used in DQN (see Section 2.2). These ideas have also been recently applied to MDRL [178].

4.2. Lessons learned

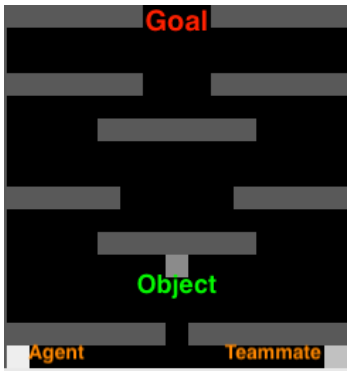
We have exemplified how RL and MAL can be extended for MDRL settings. Now, we outline general *best practices* learned from the works analyzed throughout this paper.

- *Experience replay buffer in MDRL.* While some works removed the ER buffer in MDRL [162] it is an important component in many DRL and MDRL algorithms. However, using the standard buffer (i.e., keeping $\langle s, a, r, s' \rangle$) will probably fail due to a lack of theoretical guarantees under this setting, see Sections 2.2 and 4.1. *Adding information in the experience tuple* that can help disambiguate the sample is the solution adopted in many works, whether a value based method [168, 35, 166, 178] or a policy gradient method [63]. In this regard, it is an open question to consider how new DRL ideas could be best integrated into the ER [286, 111, 287, 288, 96] and how those ideas would fare in a MDRL setting.

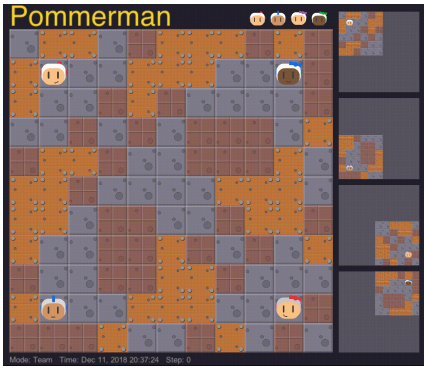
- *Centralized learning with decentralized execution.* Many MAL works were either fully centralized or fully decentralized approaches. However, inspired by *decentralized partially observable Markov decison processes* (DEC-POMDPs) [289, 290],[17] in MDRL this new mixed paradigm has been commonly used [168, 35, 181, 172, 167, 63] (a notable exception are DEC-HDRQNs [166] which perform learning and execution in a decentralized manner, see Section 3.5). Note that not all real-world problems fit into this paradigm and it is more common for robotics or games where a simulator is generally available [162]. The main benefit is that during learning *additional information can be used* (e.g., global state, action, or rewards) and during execution this information is removed.

- *Parameter sharing.* Another frequent component in many MDRL works is the idea of sharing parameters, i.e., training a single network in which agents share their weights. Note that, since agents could receive different observations (e.g., in partially observable scenarios), they can still behave differently. This method was proposed concurrently in different works [292, 162] and later it has been successfully applied in many others [163, 164, 168, 180, 181].

[17] Centralized planning and decentralized execution is also a standard paradigm for multiagent planning [291].



(a) Multiagent object transportation



(b) Pommerman

Figure 7: (a) A fully cooperative benchmark with two agents, Multiagent Object Trasportation. (b) A mixed cooperative-competitive domain with four agents, Pommerman. For more MDRL benchmarks see Section 4.3.

- *Recurrent networks.* Recurrent neural networks (RNNs) enhanced neural networks with a memory capability, however, they suffer from the vanishing gradient problem, which renders them inefficient for long-term dependencies [293]. However, RNN variants such as LSTMs [86, 294] and GRUs (Gated Recurrent Unit) [295] addressed this challenge. In single-agent DRL, DRQN [85] initially proposed idea of using recurrent networks in single-agent *partially observable* environments. Then, Feudal Networks [229] proposed a hierarchical approach [230], *multiple LSTM networks with different time-scales*, i.e., the observation input schedule is different for each LSTM network, to create a temporal hierarchy so that it can better address the long-term credit assignment challenge for RL problems. Recently, the use of recurrent networks has been extended to MDRL to address the challenge of partially observability [158, 162, 164, 166, 180, 181, 170, 171, 174] for example, in FTW [179], depicted in Figure 5 and DRPIRQN [171] depicted in Figure 6. See Section 4.4 for practical challenges (e.g., training issues) of recurrent networks in MDRL.

- *Overfitting in MAL.* In single-agent RL, agents can overfit to the environment [296]. A similar problem can occur in multiagent settings [254], agents can overfit, i.e., an agent's policy can easily get stuck in a local optima and the learned policy may be only locally optimal to other agents' current policies [183]. This has the effect of limiting the generalization of the learned policies [172]. To reduce this problem, a solution is to have a set of policies (an ensemble) and learn from them or best respond to the mixture of them [172, 63, 169]. Another solution has been to robustify algorithms — a robust policy should be able to behave well even with strategies different from its training (better generalization) [183].

4.3. Benchmarks for MDRL

Standardized environments such as the Arcade Learning Environment (ALE) [297, 298] and OpenAI Gym [299] have allowed single-agent RL to move beyond toy domains. For DRL there are open-source frameworks that provide compact and reliable implementations of some state-of-the-art DRL algorithms [300]. Even though MDRL is a recent area, there are now a number of open sourced simulators and benchmarks to use with different characteristics, which we describe below.

- Fully Cooperative Multiagent Object Transporation Problems (CMOTPs)[18] were originally presented by Busoniu et al. [36] as a simple two-agent coordination problem in MAL. Palmer et al. [35] proposed two pixel-based extensions to the original setting which include narrow passages that test the agents' ability to master fully-cooperative sub-tasks, stochastic rewards and noisy observations, see Figure 7a.

- The Apprentice Firemen Game[19] (inspired by the classic climb game [126]) is another two-agent pixel-based environment that simultaneously confronts learners with four pathologies in MAL: relative overgeneralization, stochasticity, the moving target problem, and alter exploration problem [97].

- Pommerman [301] is a multiagent benchmark useful for testing cooperative, competitive and mixed (cooperative and competitive) scenarios. It supports partial observability and communication among agents, see Figure 7b. Pommerman is a very challenging domain from the exploration perspective as the rewards are very sparse and delayed [302]. A recent competition was held during NeurIPS-2018[20] and the top agents from that competition are available for training purposes.

- Starcraft Multiagent Challenge [303] is based on the real-time strategy game StarCraft II and focuses on micromanagement challenges,[21] that is, fine-grained control of individual units, where each unit is controlled by an independent agent that must act based on local observations. It is accompanied by a MDRL framework including state-of-the-art algorithms (e.g., QMIX and COMA).[22]

- The Multi-Agent Reinforcement Learning in Malmö (MARLÖ) competition [304] is another multiagent challenge with multiple cooperative 3D games[23] within Minecraft. The scenarios were created with the open source Malmö platform [305], providing examples of how a wider range of multiagent cooperative, competitive and mixed scenarios can be experimented on within Minecraft.

- Hanabi is a cooperative multiplayer card game (two to five players). The main characteristic of the game is that players do not observe their own cards but other players can reveal information about them. This makes an interesting challenge for learning algorithms in particular in the context of self-play learning and ad-hoc teams [306, 307, 308]. The Hanabi Learning Environment [309] was recently released[24] and it is accompanied with a baseline (deep RL) agent [310].

- Arena [311] is platform for multiagent research[25] based on the Unity engine [312]. It has 35 multiagent games (e.g., social dilemmas) and supports communication among agents. It has basseline implementations of recent DRL algorithms such as independent PPO learners.

[18] https://github.com/gjp1203/nui_in_madrl
[19] https://github.com/gjp1203/nui_in_madrl
[20] https://www.pommerman.com/
[21] https://github.com/oxwhirl/smac
[22] https://github.com/oxwhirl/pymarl
[23] https://github.com/crowdAI/marlo-single-agent-starter-kit/
[24] https://github.com/deepmind/hanabi-learning-environment
[25] https://github.com/YuhangSong/Arena-BuildingToolkit

- MuJoCo Multiagent Soccer [313] uses the MuJoCo physics engine [202]. The environment simulates a 2 vs. 2 soccer game with agents having a 3-dimensional action space.[26]
- Neural MMO [314] is a research platform[27] inspired by the human game genre of Massively Multiplayer Online (MMO) Role-Playing Games. These games involve a large, variable number of players competing to survive.

4.4. Practical challenges in MDRL

In this section we take a more critical view with respect to MDRL and highlight different practical challenges that already happen in DRL and that are likely to occur in MDRL such as reproducibility, hyperparameter tuning, the need of computational resources and conflation of results. We provide pointers on how we think those challenges could be (partially) addressed.

Reproducibility, troubling trends and negative results. Reproducibility is a challenge in RL which is only aggravated in DRL due to different sources of stochasticity: baselines, hyperparameters, architectures [315, 316] and random seeds [317]. Moreover, DRL does not have common practices for statistical testing [318] which has led to bad practices such as only reporting the results when algorithms perform well, sometimes referred as *cherry picking* [319] (Azizzadenesheli also describes *cherry planting* as adapting an environment to a specific algorithm [319]). We believe that together with following the advice on how to design experiments and report results [320], the community would also benefit from reporting *negative results* [321, 322, 318, 323] for carefully designed hypothesis and experiments.[28] However, we found very few papers with this characteristic[324, 325, 326] — we note that this is not encouraged in the ML community; moreover, negative results reduce the chance of paper acceptance [320]. In this regard, we ask the community to reflect on these practices and find ways to remove these obstacles.

Implementation challenges and hyperparameter tuning. One problem is that canonical implementations of DRL algorithms often contain additional non-trivial optimizations — these are sometimes necessary for the algorithms to achieve good performance [79]. A recent study by Tucker et al. [59] found that several published works on action-dependant baselines contained bugs and errors — those were the real reason of the high performance in the experimental results, not the proposed method. Melis et al. [327] compared a series of works with increasing innovations in network architectures and the vanilla LSTMs [86] (originally proposed in 1997). The results showed that, when properly tuned, LSTMs outperformed the more recent models. In this context, Lipton and Steinhardt noted that the community may have benefited more by learning the details of the hyperparameter tuning [320]. A partial reason for this surprising result might be that this type of networks are known for being difficult to train [293] and there are recent works in DRL that report problems when using recurrent networks [182, 328, 329, 330]. Another known complication is catastrophic forgetting (see Section 2.2) with recent examples in DRL [157, 92] — we expect that

[26] `https://github.com/deepmind/dm_control/tree/master/dm_control/locomotion/soccer`

[27] `https://github.com/openai/neural-mmo`

[28] This idea was initially inspired by the Workshop "Critiquing and Correcting Trends in Machine Learning" at NeurIPS 2018 where it was possible to submit *Negative results* papers: "Papers which show failure modes of existing algorithms or suggest new approaches which one might expect to perform well but which do not. The aim is to provide a venue for work which might otherwise go unpublished but which is still of interest to the community." `https://ml-critique-correct.github.io/`

these issues would likely occur in MDRL. The effects of hyperparameter tuning were analyzed in more detail in DRL by Henderson et al. [315], who arrived at the conclusion that hyperparameters can have significantly different effects across algorithms (they tested TRPO, DDPG, PPO and ACKTR) and environments since there is an intricate interplay among them [315]. The authors urge the community to report *all* parameters used in the experimental evaluations for accurate comparison — we encourage a similar behavior for MDRL. Note that hyperparameter tuning is related to the troubling trend of cherry picking in that it can show a carefully picked set of parameters that make an algorithm work (see previous challenge). Lastly, note that hyperparameter tuning is computationally very expensive, which relates to the connection with the following challenge of computational demands.

Computational resources. Deep RL usually requires millions of interactions for an agent to learn [331], i.e., low sample efficiency [332], which highlights the need for large computational infrastructure in general. The original A3C implementation [93] uses 16 CPU workers for 4 days to learn to play an Atari game with a total of 200M training frames[29] (results are reported for 57 Atari games). Distributed PPO used 64 workers (presumably one CPU per worker, although this is not clearly stated in the paper) for 100 hours (more than 4 days) to learn locomotion tasks [117]. In MDRL, for example, the Atari Pong game, agents were trained for 50 epochs, 250k time steps each, for a total of 1.25M training frames [155]. The FTW agent [179] uses 30 agents (processes) in parallel and every training game lasts for five minues; FTW agents were trained for approximately 450K games ≈4.2 years. These examples highlight the computational demands sometimes needed within DRL and MDRL.

Recent works have reduced the learning of an Atari game to minutes (Stooke and Abbeel [334] trained DRL agents in less than one hour with hardware consisting of 8 GPUs and 40 cores). However, this is (for now) the exception and computational infrastructure is a major bottleneck for doing DRL and MDRL, especially for those who do not have such large compute power (e.g., most companies and most academic research groups) [212, 322].[30] Within this context we propose two ways to address this problem. (1) Raising awareness: For DRL we found few works that study the computational demands of recent algorithms [335, 331]. For MDRL most published works do not provide information regarding computational resources used such as CPU/GPU usage, memory demands, and wall-clock computation. Therefore, the first way to tackle this issue is by raising awareness and encouraging authors to report metrics about computational demands for accurately comparison and evaluation. (2) Delve into algorithmic contributions. Another way to address these issues is to prioritize the algorithmic contribution for the new MDRL algorithms rather than the computational resources spent. Indeed, for this to work, it needs to be accompanied with high-quality reviewers.

We have argued to raise awareness on the computational demands and report results, however, there is still the open question on *how* and *what* to measure/report. There are several dimensions to measure efficiency: sample efficiency is commonly measured by counting state-action pairs used for training; computational efficiency could be measured by number of CPUs/GPUs and days used for training. How do we measure the impact of other resources, such as external data sources or

[29] It is sometimes unclear in the literature what is the meaning of frame due to the "frame skip" technique. It is therefore suggested to refer to "game frames" and "training frames" [333].

[30] One recent effort by Beeching et al. [212] proposes to use only "mid-range hardware" (8 CPUs and 1 GPU) to train deep RL agents.

annotations?[31] Similarly, do we need to differentiate the computational needs of the algorithm itself versus the environment it is run in? We do not have the answers, however, we point out that current standard metrics might not be entirely comprehensive.

In the end, we believe that high compute based methods act as a frontier to showcase benchmarks [19, 18], i.e., they show what results are possible as data and compute is scaled up (e.g., OpenAI Five generates 180 years of gameplay data each day using 128,000 CPU cores and 256 GPUs [18]; AlphaStar uses 200 years of Starcraft II gameplay [19]); however, lighter compute based algorithmic methods can also yield significant contributions to better tackle real-world problems.

Occam's razor and ablative analysis. Finding the simplest context that exposes the innovative research idea remains challenging, and if ignored leads to a conflation of fundamental research (working principles in the most abstract setting) and applied research (working systems as complete as possible). In particular, some deep learning papers are presented as learning from pixels without further explanation, while object-level representations would have already exposed the algorithmic contribution. This still makes sense to remain comparable with established benchmarks (e.g., OpenAI Gym [299]), but less so if custom simulations are written without open source access, as it introduces unnecessary variance in pixel-level representations and artificially inflates computational resources (see previous point about *computational resources*).[32] In this context there are some notable exceptions where the algorithmic contribution is presented in a minimal setting and then results are scaled into complex settings: LOLA [64] first presented a minimalist setting with a two-player two-action game and then with a more complex variant; similarly, QMIX [181] presented its results in a two-step (matrix) game and then in the more involved Starcraft II micromanagement domain [303].

4.5. Open questions

Finally, here we present some open questions for MDRL and point to suggestions on how to approach them. We believe that there are solid ideas in earlier literature and we refer the reader to Section 4.1 to avoid deep learning amnesia.

- On the challenge of sparse and delayed rewards.

  Recent MDRL competitions and environments have complex scenarios where many actions are taken before a reward signal is available (see Section 4.3). This sparseness is already a challenge for RL [20, 338] where approaches such as count-based exploration/intrinsic motivation [196, 339, 340, 341, 342] and hierarchical learning [343, 344, 111] have been proposed to address it — in MDRL this is even more problematic since the agents not only need to learn basic behaviors (like in DRL), but also to learn the strategic element (e.g., competitive/collaborative) embedded in the multiagent setting. To address this issue, recent MDRL approaches applied *dense* rewards [206, 205, 204] (a concept originated in RL) at each step to allow the agents to learn basic motor skills and then decrease these *dense* rewards over

[31] NeurIPS 2019 hosts the "MineRL Competition on Sample Efficient Reinforcement Learning using Human Priors" where the primary goal of the competition is to foster the development of algorithms which can efficiently leverage human demonstrations to drastically reduce the number of samples needed to solve complex, hierarchical, and sparse environments [336].

[32] Cuccu, Togelius and Cudré-Mauroux achieved state-of-the-art policy learning in Atari games with only 6 to 18 neurons [337]. The main idea was to decouple image processing from decision-making.

time in favor of the environmental reward [158], see Section 3.3. Recent works like OpenAI Five [18] uses hand-crafted intermediate rewards to accelerate the learning and FTW [179] lets the agents learn their internal rewards by a hierarchical two-tier optimization. In *single agent* domains, RUDDER [345] has been recently proposed for such delayed sparse reward problems. RUDDER generates a new MDP with *more intermediate rewards* whose optimal solution is still an optimal solution to the original MDP. This is achieved by using LSTM networks to redistribute the original sparse reward to earlier state-action pairs and automatically provide reward shaping. How to best extend RUDDER to multiagent domains is an open avenue of research.

- On the role of self-play.

  Self-play is a cornerstone in MAL with impressive results [147, 127, 145, 346, 143]. While notable results had also been shown in MDRL [173, 193], recent works have also shown that *plain* self-play does not yield the best results. However, adding diversity, i.e., evolutionary methods [239, 240, 233, 234] or sampling-based methods, have shown good results [158, 179, 159]. A drawback of these solutions is the additional computational requirements since they need either parallel training (more CPU computation) or memory requirements. Then, it is still an open problem to improve the computational efficiency of these previously proposed successful methods, i.e., achieving similar training stability with smaller population sizes that uses fewer CPU workers in MAL and MDRL (see Section 4.4 and Albrecht et al. [11, Section 5.5]).

- On the challenge of the combinatorial nature of MDRL.

  Monte Carlo tree search (MCTS) [347] has been the backbone of the major breakthroughs behind AlphaGo [14] and AlphaGo Zero [15] that combined search and DRL. A recent work [348] has outlined how search and RL can be better combined for potentially new methods. However, for multiagent scenarios, there is an additional challenge of the exponential growth of all the agents' action spaces for centralized methods [349]. One way to tackle this challenge within multiagent scenarios is the use of search parallelization [350, 351]. Given more scalable planners, there is room for research in combining these techniques in MDRL settings.

  To learn complex multiagent interactions some type of abstraction [352] is often needed, for example, factored value functions [353, 354, 242, 243, 355, 356] (see QMIX and VDN in Section 3.5 for recent work in MDRL) try to exploit independence among agents through (factored) structure; however, in MDRL there are still open questions such as understanding their representational power [244] (e.g., the accuracy of the learned Q-function approximations) and how to learn those factorizations, where ideas from transfer planning techniques could be useful [357, 103]. In transfer planning the idea is to define a simpler "source problem" (e.g., with fewer agents), in which the agent(s) can plan [357] or learn [103]; since it is less complex than the real multiagent problem, issues such as the non-stationarity of the environment can be reduced/removed. Lastly, another related idea are *influence* abstractions [358, 359, 10], where instead of learning a complex multiagent model, these methods try to build smaller models based on the influence agents can exert on one another. While this has not been sufficiently explored in actual multiagent settings, there is some evidence that these ideas can lead to effective inductive biases, improving effectiveness of DRL in such local abstractions [360].

5. Conclusions

Deep reinforcement learning has shown recent success on many fronts [13, 14, 16] and a natural next step is to test multiagent scenarios. However, learning in multiagent environments is fundamentally more difficult due to non-stationarity, the increase of dimensionality, and the credit-assignment problem, among other factors [1, 5, 10, 147, 241, 361, 97].

This survey provides broad overview of recent works in the emerging area of Multiagent Deep Reinforcement Learning (MDRL). First, we categorized recent works into four different topics: emergent behaviors, learning communication, learning cooperation, and agents modeling agents. Then, we exemplified how key components (e.g., experience replay and difference rewards) originated in RL and MAL need to be adapted to work in MDRL. We provided general lessons learned applicable to MDRL, pointed to recent multiagent benchmarks and highlighted some open research problems. Finally, we also reflected on the practical challenges such as computational demands and reproducibility in MDRL.

Our conclusions of this work are that while the number of works in DRL and MDRL are notable and represent important milestones for AI, at the same time we acknowledge there are also open questions in both (deep) single-agent learning [38, 298, 362, 79] and multiagent learning [363, 364, 365, 366, 367, 368]. Our view is that there are practical issues within MDRL that hinder its scientific progress: the necessity of high compute power, complicated reproducibility (e.g., hyperparameter tuning), and the lack of sufficient encouragement for publishing negative results. However, we remain highly optimistic of the multiagent community and hope this work serves to raise those issues, encounter good solutions, and ultimately take advantage of the existing literature and resources available to move the area in the right direction.

Acknowledgements

We would like to thank Chao Gao, Nidhi Hegde, Gregory Palmer, Felipe Leno Da Silva and Craig Sherstan for reading earlier versions of this work and providing feedback, to April Cooper for her visual designs for the figures in the article, to Frans Oliehoek, Sam Devlin, Marc Lanctot, Nolan Bard, Roberta Raileanu, Angeliki Lazaridou, and Yuhang Song for clarifications in their areas of expertise, to Baoxiang Wang for his suggestions on recent deep RL works, to Michael Kaisers, Daan Bloembergen, and Katja Hofmann for their comments about the practical challenges of MDRL, and to the editor and three anonymous reviewers whose comments and suggestions increased the quality of this work.

References

[1] P. Stone, M. M. Veloso, Multiagent Systems - A Survey from a Machine Learning Perspective., Autonomous Robots 8 (3) (2000) 345–383.

[2] Y. Shoham, R. Powers, T. Grenager, If multi-agent learning is the answer, what is the question?, Artificial Intelligence 171 (7) (2007) 365–377.

[3] E. Alonso, M. D'inverno, D. Kudenko, M. Luck, J. Noble, Learning in multi-agent systems, Knowledge Engineering Review 16 (03) (2002) 1–8.

[4] K. Tuyls, G. Weiss, Multiagent learning: Basics, challenges, and prospects, AI Magazine 33 (3) (2012) 41–52.

[5] L. Busoniu, R. Babuska, B. De Schutter, A Comprehensive Survey of Multiagent Reinforcement Learning, IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 38 (2) (2008) 156–172.

[6] A. Nowé, P. Vrancx, Y.-M. De Hauwere, Game theory and multi-agent reinforcement learning, in: Reinforcement Learning, Springer, 2012, pp. 441–470.

[7] L. Panait, S. Luke, Cooperative Multi-Agent Learning: The State of the Art, Autonomous Agents and Multi-Agent Systems 11 (3).
[8] L. Matignon, G. J. Laurent, N. Le Fort-Piat, Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems, Knowledge Engineering Review 27 (1) (2012) 1–31.
[9] D. Bloembergen, K. Tuyls, D. Hennes, M. Kaisers, Evolutionary Dynamics of Multi-Agent Learning: A Survey., Journal of Artificial Intelligence Research 53 (2015) 659–697.
[10] P. Hernandez-Leal, M. Kaisers, T. Baarslag, E. Munoz de Cote, A Survey of Learning in Multiagent Environments - Dealing with Non-Stationarity.
URL http://arxiv.org/abs/1707.09183
[11] S. V. Albrecht, P. Stone, Autonomous agents modelling other agents: A comprehensive survey and open problems, Artificial Intelligence 258 (2018) 66–95.
[12] F. L. Silva, A. H. R. Costa, A survey on transfer learning for multiagent reinforcement learning systems, Journal of Artificial Intelligence Research 64 (2019) 645–703.
[13] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.
[14] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.
[15] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of Go without human knowledge, Nature 550 (7676) (2017) 354.
[16] M. Moravčík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, M. Bowling, DeepStack: Expert-level artificial intelligence in heads-up no-limit poker, Science 356 (6337) (2017) 508–513.
[17] N. Brown, T. Sandholm, Superhuman AI for heads-up no-limit poker: Libratus beats top professionals, Science 359 (6374) (2018) 418–424.
[18] Open AI Five, https://blog.openai.com/openai-five, [Online; accessed 7-September-2018] (2018).
[19] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, T. Ewalds, D. Horgan, M. Kroiss, I. Danihelka, J. Agapiou, J. Oh, V. Dalibard, D. Choi, L. Sifre, Y. Sulsky, S. Vezhnevets, J. Molloy, T. Cai, D. Budden, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, T. Pohlen, Y. Wu, D. Yogatama, J. Cohen, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, C. Apps, K. Kavukcuoglu, D. Hassabis, D. Silver, AlphaStar: Mastering the Real-Time Strategy Game StarCraft II, https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/ (2019).
[20] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, 2nd Edition, MIT Press, 2018.
[21] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[22] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks 61 (2015) 85–117.
[23] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, A Brief Survey of Deep Reinforcement Learning .
URL http://arXiv.org/abs/1708.05866v2
[24] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau, et al., An introduction to deep reinforcement learning, Foundations and Trends® in Machine Learning 11 (3-4) (2018) 219–354.
[25] Y. Yang, J. Hao, M. Sun, Z. Wang, C. Fan, G. Strbac, Recurrent Deep Multiagent Q-Learning for Autonomous Brokers in Smart Grid, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018.
[26] J. Zhao, G. Qiu, Z. Guan, W. Zhao, X. He, Deep reinforcement learning for sponsored search real-time bidding, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 1021–1030.
[27] B. M. Lake, T. D. Ullman, J. Tenenbaum, S. Gershman, Building machines that learn and think like people, Behavioral and Brain Sciences 40.
[28] A. Tamar, S. Levine, P. Abbeel, Y. Wu, G. Thomas, Value Iteration Networks., NIPS (2016) 2154–2162.
[29] G. Papoudakis, F. Christianos, A. Rahman, S. V. Albrecht, Dealing with non-stationarity in multi-agent deep reinforcement learning, arXiv preprint arXiv:1906.04737.
[30] T. T. Nguyen, N. D. Nguyen, S. Nahavandi, Deep reinforcement learning for multi-agent systems: A review of challenges, solutions and applications, arXiv preprint arXiv:1812.11794.

[31] D. H. Wolpert, K. Tumer, Optimal payoff functions for members of collectives, in: Modeling complexity in economic and social systems, 2002, pp. 355–369.
[32] A. K. Agogino, K. Tumer, Unifying Temporal and Structural Credit Assignment Problems., in: Proceedings of 17th International Conference on Autonomous Agents and Multiagent Systems, 2004.
[33] N. Fulda, D. Ventura, Predicting and Preventing Coordination Problems in Cooperative Q-learning Systems, in: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence, Hyderabad, India, 2007, pp. 780–785.
[34] E. Wei, S. Luke, Lenient Learning in Independent-Learner Stochastic Cooperative Games., Journal of Machine Learning Research.
[35] G. Palmer, K. Tuyls, D. Bloembergen, R. Savani, Lenient Multi-Agent Deep Reinforcement Learning., in: International Conference on Autonomous Agents and Multiagent Systems, 2018.
[36] L. Busoniu, R. Babuska, B. De Schutter, Multi-agent reinforcement learning: An overview, in: D. Srinivasan, L. C. Jain (Eds.), Innovations in Multi-Agent Systems and Applications - 1, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 183–221.
[37] Y. Li, Deep reinforcement learning: An overview, CoRR abs/1701.07274. arXiv:1701.07274.
URL http://arxiv.org/abs/1701.07274
[38] A. Darwiche, Human-level intelligence or animal-like abilities?, Commun. ACM 61 (10) (2018) 56–67.
[39] M. Wiering, M. Van Otterlo, Reinforcement learning, Adaptation, learning, and optimization 12.
[40] L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: A survey, Journal of artificial intelligence research 4 (1996) 237–285.
[41] M. L. Puterman, Markov decision processes: Discrete stochastic dynamic programming, John Wiley & Sons, Inc., 1994.
[42] A. R. Cassandra, Exact and approximate algorithms for partially observable Markov decision processes, Ph.D. thesis, Computer Science Department, Brown University (May 1998).
[43] K. J. Astrom, Optimal control of Markov processes with incomplete state information, Journal of mathematical analysis and applications 10 (1) (1965) 174–205.
[44] R. Bellman, A Markovian decision process, Journal of Mathematics and Mechanics 6 (5) (1957) 679–684.
[45] J. Watkins, Learning from delayed rewards, Ph.D. thesis, King's College, Cambridge, UK (Apr. 1989).
[46] T. Kamihigashi, C. Le Van, Necessary and Sufficient Conditions for a Solution of the Bellman Equation to be the Value Function: A General Principle.
URL https://halshs.archives-ouvertes.fr/halshs-01159177
[47] J. Tsitsiklis, Asynchronous stochastic approximation and Q-learning, Machine Learning 16 (3) (1994) 185–202.
[48] T. Jaakkola, M. I. Jordan, S. P. Singh, Convergence of stochastic iterative dynamic programming algorithms, in: Advances in neural information processing systems, 1994, pp. 703–710.
[49] C. Szepesvári, M. L. Littman, A unified analysis of value-function-based reinforcement-learning algorithms, Neural computation 11 (8) (1999) 2017–2060.
[50] E. Even-Dar, Y. Mansour, Learning rates for Q-learning, Journal of Machine Learning Research 5 (Dec) (2003) 1–25.
[51] C. Szepesvári, Algorithms for reinforcement learning, Synthesis lectures on artificial intelligence and machine learning 4 (1) (2010) 1–103.
[52] S. Singh, T. Jaakkola, M. L. Littman, C. Szepesvári, Convergence results for single-step on-policy reinforcement-learning algorithms, Machine learning 38 (3) (2000) 287–308.
[53] H. Van Seijen, H. Van Hasselt, S. Whiteson, M. Wiering, A theoretical and empirical analysis of Expected Sarsa, in: IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, Nashville, TN, USA, 2009, pp. 177–184.
[54] V. R. Konda, J. Tsitsiklis, Actor-critic algorithms, in: Advances in Neural Information Processing Systems, 2000.
[55] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, Policy Gradient Methods for Reinforcement Learning with Function Approximation., in: Advances in Neural Information Processing Systems, 2000.
[56] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine learning 8 (3-4) (1992) 229–256.
[57] H. Liu, Y. Feng, Y. Mao, D. Zhou, J. Peng, Q. Liu, Action-depedent control variates for policy optimization via stein's identity, in: International Conference on Learning Representations, 2018.
[58] S. Gu, T. Lillicrap, Z. Ghahramani, R. E. Turner, S. Levine, Q-prop: Sample-efficient policy gradient with an off-policy critic, in: International Conference on Learning Representations, 2017.
[59] G. Tucker, S. Bhupatiraju, S. Gu, R. E. Turner, Z. Ghahramani, S. Levine, The mirage of action-dependent

baselines in reinforcement learning, in: International Conference on Machine Learning, 2018.
[60] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, P. Moritz, Trust Region Policy Optimization., in: 31st International Conference on Machine Learning, Lille, France, 2015.
[61] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, M. Riedmiller, Deterministic policy gradient algorithms, in: ICML, 2014.
[62] R. Hafner, M. Riedmiller, Reinforcement learning in feedback control, Machine learning 84 (1-2) (2011) 137–169.
[63] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, I. Mordatch, Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments., in: Advances in Neural Information Processing Systems, 2017, pp. 6379–6390.
[64] J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, I. Mordatch, Learning with Opponent-Learning Awareness., in: Proceedings of 17th International Conference on Autonomous Agents and Multiagent Systems, Stockholm, Sweden, 2018.
[65] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, in: International Conference on Learning Representations, 2016.
[66] L. D. Pyeatt, A. E. Howe, et al., Decision tree function approximation in reinforcement learning, in: Proceedings of the third international symposium on adaptive systems: evolutionary computation and probabilistic graphical models, Vol. 2, Cuba, 2001, pp. 70–77.
[67] R. S. Sutton, Generalization in reinforcement learning: Successful examples using sparse coarse coding, in: Advances in neural information processing systems, 1996, pp. 1038–1044.
[68] R. M. Kretchmar, C. W. Anderson, Comparison of CMACs and radial basis functions for local function approximators in reinforcement learning, in: Proceedings of International Conference on Neural Networks (ICNN'97), Vol. 2, IEEE, 1997, pp. 834–837.
[69] J. A. Boyan, A. W. Moore, Generalization in reinforcement learning: Safely approximating the value function, in: Advances in neural information processing systems, 1995, pp. 369–376.
[70] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.
[71] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, G. Tesauro, Learning to learn without forgetting by maximizing transfer and minimizing interference, CoRR abs/1810.11910. arXiv:1810.11910.
URL http://arxiv.org/abs/1810.11910
[72] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing Atari with Deep Reinforcement Learning.
URL http://arxiv.org/abs/1312.5602v1
[73] G. J. Gordon, Approximate solutions to Markov decision processes, Tech. rep., Carnegie-Mellon University (1999).
[74] L. Baird, Residual algorithms: Reinforcement learning with function approximation, in: Machine Learning Proceedings 1995, 1995, pp. 30–37.
[75] S. Whiteson, P. Stone, Evolutionary function approximation for reinforcement learning, Journal of Machine Learning Research 7 (May) (2006) 877–917.
[76] J. Achiam, E. Knight, P. Abbeel, Towards Characterizing Divergence in Deep Q-Learning, CoRR abs/1903.08894. arXiv:1903.08894.
URL http://arxiv.org/abs/1903.08894
[77] H. van Hasselt, Y. Doron, F. Strub, M. Hessel, N. Sonnerat, J. Modayil, Deep reinforcement learning and the deadly triad, CoRR abs/1812.02648. arXiv:1812.02648.
URL http://arxiv.org/abs/1812.02648
[78] S. Fujimoto, H. van Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: International Conference on Machine Learning, 2018.
[79] A. Ilyas, L. Engstrom, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, A. Madry, Are deep policy gradient algorithms truly policy gradient algorithms?, CoRR abs/1811.02553. arXiv:1811.02553.
URL http://arxiv.org/abs/1811.02553
[80] T. Lu, D. Schuurmans, C. Boutilier, Non-delusional Q-learning and value-iteration, in: Advances in Neural Information Processing Systems, 2018, pp. 9949–9959.
[81] J. N. Tsitsiklis, B. Van Roy, Analysis of temporal-diffference learning with function approximation, in: Advances in neural information processing systems, 1997, pp. 1075–1081.
[82] F. S. Melo, S. P. Meyn, M. I. Ribeiro, An analysis of reinforcement learning with function approximation, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 664–671.
[83] D. Ernst, P. Geurts, L. Wehenkel, Tree-based batch mode reinforcement learning, Journal of Machine Learning Research 6 (Apr) (2005) 503–556.

[84] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, K. Kavukcuoglu, Reinforcement Learning with Unsupervised Auxiliary Tasks., in: International Conference on Learning Representations, 2017.
[85] M. Hausknecht, P. Stone, Deep Recurrent Q-Learning for Partially Observable MDPs, in: International Conference on Learning Representations, 2015.
[86] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
[87] M. Riedmiller, Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method, in: European Conference on Machine Learning, Springer, 2005, pp. 317–328.
[88] R. H. Crites, A. G. Barto, Elevator group control using multiple reinforcement learning agents, Machine learning 33 (2-3) (1998) 235–262.
[89] L. J. Lin, Programming robots using reinforcement learning and teaching., in: AAAI, 1991, pp. 781–786.
[90] L.-J. Lin, Self-improving reactive agents based on reinforcement learning, planning and teaching, Machine learning 8 (3-4) (1992) 293–321.
[91] H. V. Hasselt, Double Q-learning, in: Advances in Neural Information Processing Systems, 2010, pp. 2613–2621.
[92] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double Q-learning, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[93] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: International Conference on Machine Learning, 2016, pp. 1928–1937.
[94] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of learning and motivation, Vol. 24, Elsevier, 1989, pp. 109–165.
[95] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, Y. Bengio, An empirical investigation of catastrophic forgetting in gradient-based neural networks.
URL `https://arxiv.org/abs/1312.6211`
[96] D. Isele, A. Cosgun, Selective experience replay for lifelong learning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[97] G. Palmer, R. Savani, K. Tuyls, Negative update intervals in deep multi-agent reinforcement learning, in: 18th International Conference on Autonomous Agents and Multiagent Systems, 2019.
[98] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, N. De Freitas, Dueling network architectures for deep reinforcement learning, in: International Conference on Machine Learning, 2016.
[99] M. Hauskrecht, Value-function approximations for partially observable Markov decision processes, Journal of Artificial Intelligence Research 13 (1).
[100] N. Meuleau, L. Peshkin, K.-E. Kim, L. P. Kaelbling, Learning finite-state controllers for partially observable environments, in: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, 1999, pp. 427–436.
[101] D. Steckelmacher, D. M. Roijers, A. Harutyunyan, P. Vrancx, H. Plisnier, A. Nowé, Reinforcement learning in pomdps with memoryless options and option-observation initiation sets, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[102] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, Proceedings of the 32nd International Conference on Machine Learning (2015) 448–456.
[103] E. Van der Pol, F. A. Oliehoek, Coordinated deep reinforcement learners for traffic light control, in: Proceedings of Learning, Inference and Control of Multi-Agent Systems at NIPS, 2016.
[104] T. Salimans, D. P. Kingma, Weight normalization: A simple reparameterization to accelerate training of deep neural networks, in: Advances in Neural Information Processing Systems, 2016, pp. 901–909.
[105] OpenAI Baselines: ACKTR & A2C, `https://openai.com/blog/baselines-acktr-a2c/`, [Online; accessed 29-April-2019] (2017).
[106] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, N. de Freitas, Sample efficient actor-critic with experience replay, arXiv preprint arXiv:1611.01224.
[107] S. S. Gu, T. Lillicrap, R. E. Turner, Z. Ghahramani, B. Schölkopf, S. Levine, Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning, in: Advances in Neural Information Processing Systems, 2017, pp. 3846–3855.
[108] E. Shelhamer, P. Mahmoudieh, M. Argus, T. Darrell, Loss is its own reward: Self-supervision for reinforcement learning, ICLR workshops.
[109] M. G. Bellemare, W. Dabney, R. Dadashi, A. A. Taïga, P. S. Castro, N. L. Roux, D. Schuurmans, T. Lattimore, C. Lyle, A Geometric Perspective on Optimal Representations for Reinforcement Learning, CoRR abs/1901.11530. `arXiv:1901.11530`.

URL http://arxiv.org/abs/1901.11530
[110] R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, D. Precup, Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction, in: The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2, International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 761–768.
[111] T. Schaul, J. Quan, I. Antonoglou, D. Silver, Prioritized Experience Replay, in: International Conference on Learning Representations, 2016.
[112] A. W. Moore, C. G. Atkeson, Prioritized sweeping: Reinforcement learning with less data and less time, Machine learning 13 (1) (1993) 103–130.
[113] D. Andre, N. Friedman, R. Parr, Generalized prioritized sweeping, in: Advances in Neural Information Processing Systems, 1998, pp. 1001–1007.
[114] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al., IMPALA: Scalable distributed Deep-RL with importance weighted actor-learner architectures, in: International Conference on Machine Learning, 2018.
[115] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal Policy Optimization Algorithms.
URL http://arxiv.org/abs/1707.06347
[116] S. M. Kakade, A natural policy gradient, in: Advances in neural information processing systems, 2002, pp. 1531–1538.
[117] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. M. A. Eslami, M. A. Riedmiller, D. Silver, Emergence of Locomotion Behaviours in Rich Environments.
URL http://arXiv.org/abs/1707.02286v2
[118] G. Bacchiani, D. Molinari, M. Patander, Microscopic traffic simulation by cooperative multi-agent deep reinforcement learning, in: AAMAS, 2019.
[119] J. Schulman, P. Abbeel, X. Chen, Equivalence Between Policy Gradients and Soft Q-Learning, CoRR abs/1704.06440. arXiv:1704.06440.
URL http://arxiv.org/abs/1704.06440
[120] T. Haarnoja, H. Tang, P. Abbeel, S. Levine, Reinforcement learning with deep energy-based policies, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017, pp. 1352–1361.
[121] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: International Conference on Machine Learning, 2018.
[122] M. Tan, Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents, in: Machine Learning Proceedings 1993 Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, June 27–29, 1993, 1993, pp. 330–337.
[123] G. J. Laurent, L. Matignon, L. Fort-Piat, et al., The world of independent learners is not Markovian, International Journal of Knowledge-based and Intelligent Engineering Systems 15 (1) (2011) 55–64.
[124] M. L. Littman, Markov games as a framework for multi-agent reinforcement learning, in: Proceedings of the 11th International Conference on Machine Learning, New Brunswick, NJ, USA, 1994, pp. 157–163.
[125] M. L. Littman, Value-function reinforcement learning in Markov games, Cognitive Systems Research 2 (1) (2001) 55–66.
[126] C. Claus, C. Boutilier, The dynamics of reinforcement learning in cooperative multiagent systems, in: Proceedings of the 15th National Conference on Artificial Intelligence, Madison, Wisconsin, USA, 1998, pp. 746–752.
[127] J. Hu, M. P. Wellman, Nash Q-learning for general-sum stochastic games, The Journal of Machine Learning Research 4 (2003) 1039–1069.
[128] M. L. Littman, Friend-or-foe Q-learning in general-sum games, in: Proceedings of 17th International Conference on Autonomous Agents and Multiagent Systems, Williamstown, MA, USA, 2001, pp. 322–328.
[129] X. Song, T. Wang, C. Zhang, Convergence of multi-agent learning with a finite step size in general-sum games, in: 18th International Conference on Autonomous Agents and Multiagent Systems, 2019.
[130] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, T. Graepel, The mechanics of n-player differentiable games, in: Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, Stockholm, Sweden, 2018, pp. 354–363.
[131] J. Pérolat, B. Piot, O. Pietquin, Actor-critic fictitious play in simultaneous move multistage games, in: 21st International Conference on Artificial Intelligence and Statistics, 2018.
[132] G. Bono, J. S. Dibangoye, L. Matignon, F. Pereyron, O. Simonin, Cooperative multi-agent policy gradient, in: European Conference on Machine Learning, 2018.
[133] T. W. Neller, M. Lanctot, An introduction to counterfactual regret minimization, in: Proceedings of Model AI Assignments, The Fourth Symposium on Educational Advances in Artificial Intelligence (EAAI-2013), 2013.

[134] M. Zinkevich, M. Johanson, M. Bowling, C. Piccione, Regret minimization in games with incomplete information, in: Advances in neural information processing systems, 2008, pp. 1729–1736.
[135] A. Blum, Y. Monsour, Learning, regret minimization, and equilibria, in: Algorithmic Game Theory, Cambridge University Press, 2007, Ch. 4.
[136] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Pérolat, K. Tuyls, R. Munos, M. Bowling, Actor-critic policy optimization in partially observable multiagent environments, in: Advances in Neural Information Processing Systems, 2018, pp. 3422–3435.
[137] E. Lockhart, M. Lanctot, J. Pérolat, J. Lespiau, D. Morrill, F. Timbers, K. Tuyls, Computing approximate equilibria in sequential adversarial games by exploitability descent, CoRR abs/1903.05614. arXiv:1903.05614. URL http://arxiv.org/abs/1903.05614
[138] M. Johanson, N. Bard, N. Burch, M. Bowling, Finding optimal abstract strategies in extensive-form games, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[139] E. Kalai, E. Lehrer, Rational learning leads to Nash equilibrium, Econometrica: Journal of the Econometric Society (1993) 1019–1045.
[140] T. W. Sandholm, R. H. Crites, Multiagent reinforcement learning in the iterated prisoner's dilemma, Biosystems 37 (1-2) (1996) 147–166.
[141] S. Singh, M. Kearns, Y. Mansour, Nash convergence of gradient dynamics in general-sum games, in: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 2000, pp. 541–548.
[142] B. Banerjee, J. Peng, Adaptive policy gradient in multiagent learning, in: Proceedings of the second international joint conference on Autonomous agents and multiagent systems, ACM, 2003, pp. 686–692.
[143] A. Greenwald, K. Hall, Correlated Q-learning, in: Proceedings of 17th International Conference on Autonomous Agents and Multiagent Systems, Washington, DC, USA, 2003, pp. 242–249.
[144] M. Bowling, Convergence problems of general-sum multiagent reinforcement learning, in: International Conference on Machine Learning, 2000, pp. 89–94.
[145] M. Bowling, Convergence and no-regret in multiagent learning, in: Advances in Neural Information Processing Systems, Vancouver, Canada, 2004, pp. 209–216.
[146] M. Zinkevich, A. Greenwald, M. L. Littman, Cyclic equilibria in Markov games, in: Advances in Neural Information Processing Systems, 2006, pp. 1641–1648.
[147] M. Bowling, M. Veloso, Multiagent learning using a variable learning rate, Artificial Intelligence 136 (2) (2002) 215–250.
[148] M. Kaisers, K. Tuyls, FAQ-learning in matrix games: demonstrating convergence near Nash equilibria, and bifurcation of attractors in the battle of sexes, in: AAAI Workshop on Interactive Decision Theory and Game Theory, San Francisco, CA, USA, 2011, pp. 309–316.
[149] M. Wunder, M. L. Littman, M. Babes, Classes of Multiagent Q-learning Dynamics with epsilon-greedy Exploration, in: Proceedings of the 35th International Conference on Machine Learning, Haifa, Israel, 2010, pp. 1167–1174.
[150] G. Tesauro, Extending Q-learning to general adaptive multi-agent systems, in: Advances in Neural Information Processing Systems, Vancouver, Canada, 2003, pp. 871–878.
[151] M. Weinberg, J. S. Rosenschein, Best-response multiagent learning in non-stationary environments, in: Proceedings of the 3rd International Conference on Autonomous Agents and Multiagent Systems, New York, NY, USA, 2004, pp. 506–513.
[152] D. Chakraborty, P. Stone, Multiagent learning in the presence of memory-bounded agents, Autonomous Agents and Multi-Agent Systems 28 (2) (2013) 182–213.
[153] G. Weiss (Ed.), Multiagent Systems, 2nd Edition, (Intelligent Robotics and Autonomous Agents series), MIT Press, Cambridge, MA, USA, 2013.
[154] Multiagent Learning, Foundations and Recent Trends, https://www.cs.utexas.edu/~larg/ijcai17_tutorial/multiagent_learning.pdf, [Online; accessed 7-September-2018] (2017).
[155] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, R. Vicente, Multiagent cooperation and competition with deep reinforcement learning, PLOS ONE 12 (4) (2017) e0172395.
[156] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, Multi-agent Reinforcement Learning in Sequential Social Dilemmas, in: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, Sao Paulo, 2017.
[157] M. Raghu, A. Irpan, J. Andreas, R. Kleinberg, Q. Le, J. Kleinberg, Can Deep Reinforcement Learning solve Erdos-Selfridge-Spencer Games?, in: Proceedings of the 35th International Conference on Machine Learning, 2018.

[158] T. Bansal, J. Pachocki, S. Sidor, I. Sutskever, I. Mordatch, Emergent Complexity via Multi-Agent Competition., in: International Conference on Machine Learning, 2018.
[159] J. Z. Leibo, J. Perolat, E. Hughes, S. Wheelwright, A. H. Marblestone, E. Duéñez-Guzmán, P. Sunehag, I. Dunning, T. Graepel, Malthusian reinforcement learning, in: 18th International Conference on Autonomous Agents and Multiagent Systems, 2019.
[160] I. Mordatch, P. Abbeel, Emergence of grounded compositional language in multi-agent populations, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[161] A. Lazaridou, A. Peysakhovich, M. Baroni, Multi-Agent Cooperation and the Emergence of (Natural) Language, in: International Conference on Learning Representations, 2017.
[162] J. N. Foerster, Y. M. Assael, N. De Freitas, S. Whiteson, Learning to communicate with deep multi-agent reinforcement learning, in: Advances in Neural Information Processing Systems, 2016, pp. 2145–2153.
[163] S. Sukhbaatar, A. Szlam, R. Fergus, Learning Multiagent Communication with Backpropagation, in: Advances in Neural Information Processing Systems, 2016, pp. 2244–2252.
[164] P. Peng, Q. Yuan, Y. Wen, Y. Yang, Z. Tang, H. Long, J. Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games.
URL http://arxiv.org/abs/1703.10069
[165] E. Pesce, G. Montana, Improving coordination in multi-agent deep reinforcement learning through memory-driven communication, CoRR abs/1901.03887. arXiv:1901.03887.
URL http://arxiv.org/abs/1901.03887
[166] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, J. Vian, Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability, in: Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017.
[167] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, S. Whiteson, Counterfactual Multi-Agent Policy Gradients., in: 32nd AAAI Conference on Artificial Intelligence, 2017.
[168] J. N. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. S. Torr, P. Kohli, S. Whiteson, Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning., in: International Conference on Machine Learning, 2017.
[169] H. He, J. Boyd-Graber, K. Kwok, H. Daume, Opponent modeling in deep reinforcement learning, in: 33rd International Conference on Machine Learning, 2016, pp. 2675–2684.
[170] R. Raileanu, E. Denton, A. Szlam, R. Fergus, Modeling Others using Oneself in Multi-Agent Reinforcement Learning., in: International Conference on Machine Learning, 2018.
[171] Z.-W. Hong, S.-Y. Su, T.-Y. Shann, Y.-H. Chang, C.-Y. Lee, A Deep Policy Inference Q-Network for Multi-Agent Systems, in: International Conference on Autonomous Agents and Multiagent Systems, 2018.
[172] M. Lanctot, V. F. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, T. Graepel, A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning., in: Advances in Neural Information Processing Systems, 2017.
[173] J. Heinrich, D. Silver, Deep Reinforcement Learning from Self-Play in Imperfect-Information Games.
URL http://arxiv.org/abs/1603.01121
[174] N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. M. A. Eslami, M. Botvinick, Machine Theory of Mind., in: International Conference on Machine Learning, Stockholm, Sweden, 2018.
[175] T. Yang, J. Hao, Z. Meng, C. Zhang, Y. Z. Z. Zheng, Towards Efficient Detection and Optimal Response against Sophisticated Opponents, in: IJCAI, 2019.
[176] A. Lerer, A. Peysakhovich, Maintaining cooperation in complex social dilemmas using deep reinforcement learning, CoRR abs/1707.01068. arXiv:1707.01068.
URL http://arxiv.org/abs/1707.01068
[177] W. Kim, M. Cho, Y. Sung, Message-Dropout: An Efficient Training Method for Multi-Agent Deep Reinforcement Learning, in: 33rd AAAI Conference on Artificial Intelligence, 2019.
[178] Y. Zheng, J. Hao, Z. Zhang, Weighted double deep multiagent reinforcement learning in stochastic cooperative environments.
URL http://arXiv.org/abs/1802.08534
[179] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castañeda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, N. Sonnerat, T. Green, L. Deason, J. Z. Leibo, D. Silver, D. Hassabis, K. Kavukcuoglu, T. Graepel, Human-level performance in 3d multiplayer games with population-based reinforcement learning, Science 364 (6443) (2019) 859–865. arXiv:https://science.sciencemag.org/content/364/6443/859.full.pdf, doi:10.1126/science.aau6249.
URL https://science.sciencemag.org/content/364/6443/859

[180] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. F. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, T. Graepel, Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward., in: Proceedings of 17th International Conference on Autonomous Agents and Multiagent Systems, Stockholm, Sweden, 2018.
[181] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. N. Foerster, S. Whiteson, QMIX - Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning., in: International Conference on Machine Learning, 2018.
[182] J. K. Gupta, M. Egorov, M. Kochenderfer, Cooperative multi-agent control using deep reinforcement learning, in: G. Sukthankar, J. A. Rodriguez-Aguilar (Eds.), Autonomous Agents and Multiagent Systems, Springer International Publishing, Cham, 2017, pp. 66–83.
[183] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, S. Russell, Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient, in: AAAI Conference on Artificial Intelligence, 2019.
[184] Y. Zheng, Z. Meng, J. Hao, Z. Zhang, T. Yang, C. Fan, A Deep Bayesian Policy Reuse Approach Against Non-Stationary Agents, in: Advances in Neural Information Processing Systems, 2018, pp. 962–972.
[185] R. Powers, Y. Shoham, Learning against opponents with bounded memory, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburg, Scotland, UK, 2005, pp. 817–822.
[186] M. L. Littman, P. Stone, Implicit Negotiation in Repeated Games, ATAL '01: Revised Papers from the 8th International Workshop on Intelligent Agents VIII.
[187] R. Axelrod, W. D. Hamilton, The evolution of cooperation, Science 211 (27) (1981) 1390–1396.
[188] E. Munoz de Cote, A. Lazaric, M. Restelli, Learning to cooperate in multi-agent social dilemmas, in: Proceedings of the 5th International Conference on Autonomous Agents and Multiagent Systems, Hakodate, Hokkaido, Japan, 2006, pp. 783–785.
[189] J. L. Stimpson, M. A. Goodrich, Learning to cooperate in a social dilemma: A satisficing approach to bargaining, in: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 728–735.
[190] G. W. Brown, Iterative solution of games by fictitious play, Activity analysis of production and allocation 13 (1) (1951) 374–376.
[191] D. Monderer, L. S. Shapley, Fictitious play property for games with identical interests, Journal of economic theory 68 (1) (1996) 258–265.
[192] G. Tesauro, Temporal difference learning and TD-Gammon, Communications of the ACM 38 (3) (1995) 58–68.
[193] M. Bowling, N. Burch, M. Johanson, O. Tammelin, Heads-up limit hold'em poker is solved, Science 347 (6218) (2015) 145–149.
[194] J. Z. Leibo, E. Hughes, M. Lanctot, T. Graepel, Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research, CoRR abs/1903.00742. arXiv:1903.00742. URL http://arxiv.org/abs/1903.00742
[195] S. Samothrakis, S. Lucas, T. Runarsson, D. Robles, Coevolving game-playing agents: Measuring performance and intransitivities, IEEE Transactions on Evolutionary Computation 17 (2) (2013) 213–226.
[196] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, R. Munos, Unifying count-based exploration and intrinsic motivation, in: Advances in Neural Information Processing Systems, 2016, pp. 1471–1479.
[197] D. E. Moriarty, A. C. Schultz, J. J. Grefenstette, Evolutionary algorithms for reinforcement learning, Journal of Artificial Intelligence Research 11 (1999) 241–276.
[198] F. A. Oliehoek, E. D. De Jong, N. Vlassis, The parallel Nash memory for asymmetric games, in: Proceedings of the 8th annual conference on Genetic and evolutionary computation, ACM, 2006, pp. 337–344.
[199] L. Bull, T. C. Fogarty, M. Snaith, Evolution in multi-agent systems: Evolving communicating classifier systems for gait in a quadrupedal robot, in: Proceedings of the 6th International Conference on Genetic Algorithms, Morgan Kaufmann Publishers Inc., 1995, pp. 382–388.
[200] L. Bull, Evolutionary computing in multi-agent environments: Operators, in: International Conference on Evolutionary Programming, Springer, 1998, pp. 43–52.
[201] H. Iba, Emergent cooperation for multiple agents using genetic programming, in: International Conference on Parallel Problem Solving from Nature, Springer, 1996, pp. 32–41.
[202] E. Todorov, T. Erez, Y. Tassa, MuJoCo - A physics engine for model-based control, Intelligent Robots and Systems (2012) 5026–5033.
[203] V. Gullapalli, A. G. Barto, Shaping as a method for accelerating reinforcement learning, in: Proceedings of the 1992 IEEE international symposium on intelligent control, IEEE, 1992, pp. 554–559.
[204] S. Mahadevan, J. Connell, Automatic programming of behavior-based robots using reinforcement learning, Artificial intelligence 55 (2-3) (1992) 311–365.

[205] G. Konidaris, A. Barto, Autonomous shaping: Knowledge transfer in reinforcement learning, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 489–496.
[206] A. Y. Ng, D. Harada, S. J. Russell, Policy invariance under reward transformations: Theory and application to reward shaping, in: Proceedings of the Sixteenth International Conference on Machine Learning, 1999, pp. 278–287.
[207] P. Erdös, J. L. Selfridge, On a combinatorial game, Journal of Combinatorial Theory, Series A 14 (3) (1973) 298–301.
[208] J. Spencer, Randomization, derandomization and antirandomization: three games, Theoretical Computer Science 131 (2) (1994) 415–429.
[209] D. Fudenberg, J. Tirole, Game Theory, The MIT Press, 1991.
[210] L. v. d. Maaten, G. Hinton, Visualizing data using t-SNE, Journal of machine learning research 9 (Nov) (2008) 2579–2605.
[211] T. Zahavy, N. Ben-Zrihem, S. Mannor, Graying the black box: Understanding DQNs, in: International Conference on Machine Learning, 2016, pp. 1899–1908.
[212] E. Beeching, C. Wolf, J. Dibangoye, O. Simonin, Deep Reinforcement Learning on a Budget: 3D Control and Reasoning Without a Supercomputer, CoRR abs/1904.01806. arXiv:1904.01806.
URL http://arxiv.org/abs/1904.01806
[213] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research 15 (1) (2014) 1929–1958.
[214] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (11) (1997) 2673–2681.
[215] R. Lowe, J. Foerster, Y.-L. Boureau, J. Pineau, Y. Dauphin, On the pitfalls of measuring emergent communication, in: 18th International Conference on Autonomous Agents and Multiagent Systems, 2019.
[216] M. Tambe, Towards flexible teamwork, Journal of artificial intelligence research 7 (1997) 83–124.
[217] B. J. Grosz, S. Kraus, Collaborative plans for complex group action, Artificial Intelligence 86 (2) (1996) 269–357.
[218] D. Precup, R. S. Sutton, S. Singh, Eligibility traces for off-policy policy evaluation, in: Proceedings of the Seventeenth International Conference on Machine Learning., 2000.
[219] J. Frank, S. Mannor, D. Precup, Reinforcement learning in the presence of rare events, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 336–343.
[220] T. I. Ahamed, V. S. Borkar, S. Juneja, Adaptive importance sampling technique for markov chains using stochastic approximation, Operations Research 54 (3) (2006) 489–504.
[221] K. A. Ciosek, S. Whiteson, Offer: Off-environment reinforcement learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[222] D. Bloembergen, M. Kaisers, K. Tuyls, Lenient frequency adjusted Q-learning, in: Proceedings of the 22nd Belgian/Netherlands Artificial Intelligence Conference, 2010.
[223] L. Panait, K. Sullivan, S. Luke, Lenience towards teammates helps in cooperative multiagent learning, in: Proceedings of the 5th International Conference on Autonomous Agents and Multiagent Systems, Hakodate, Japan, 2006.
[224] L. Panait, K. Tuyls, S. Luke, Theoretical advantages of lenient learners: An evolutionary game theoretic perspective, JMLR 9 (Mar) (2008) 423–457.
[225] M. Lauer, M. Riedmiller, An algorithm for distributed reinforcement learning in cooperative multi-agent systems, in: In Proceedings of the Seventeenth International Conference on Machine Learning, 2000.
[226] R. Caruana, Multitask learning, Machine learning 28 (1) (1997) 41–75.
[227] A. A. Rusu, S. G. Colmenarejo, C. Gulcehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, R. Hadsell, Policy Distillation, in: International Conference on Learning Representations, 2016.
[228] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: NIPS Deep Learning Workshop, 2014.
[229] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, K. Kavukcuoglu, FeUdal Networks for Hierarchical Reinforcement Learning., International Conference On Machine Learning.
[230] P. Dayan, G. E. Hinton, Feudal reinforcement learning, in: Advances in neural information processing systems, 1993, pp. 271–278.
[231] S. P. Singh, Transfer of learning by composing solutions of elemental sequential tasks, Machine Learning 8 (3-4) (1992) 323–339.
[232] Capture the Flag: the emergence of complex cooperative agents, https://deepmind.com/blog/

capture-the-flag/, [Online; accessed 7-September-2018] (2018).
[233] C. D. Rosin, R. K. Belew, New methods for competitive coevolution, Evolutionary computation 5 (1) (1997) 1–29.
[234] J. Lehman, K. O. Stanley, Exploiting open-endedness to solve problems through the search for novelty., in: ALIFE, 2008, pp. 329–336.
[235] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, et al., Population based training of neural networks.
URL http://arxiv.org/abs/1711.09846
[236] A. E. Elo, The rating of chessplayers, past and present, Arco Pub., 1978.
[237] R. Herbrich, T. Minka, T. Graepel, TrueSkill: a Bayesian skill rating system, in: Advances in neural information processing systems, 2007, pp. 569–576.
[238] S. Omidshafiei, C. Papadimitriou, G. Piliouras, K. Tuyls, M. Rowland, J.-B. Lespiau, W. M. Czarnecki, M. Lanctot, J. Perolat, R. Munos, α-Rank: Multi-Agent Evaluation by Evolution, Scientific Reports 9.
[239] T. Back, Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms, Oxford university press, 1996.
[240] K. A. De Jong, Evolutionary computation: a unified approach, MIT press, 2006.
[241] K. Tumer, A. Agogino, Distributed agent-based air traffic flow management, in: Proceedings of the 6th International Conference on Autonomous Agents and Multiagent Systems, Honolulu, Hawaii, 2007.
[242] C. Guestrin, D. Koller, R. Parr, Multiagent planning with factored MDPs, in: Advances in neural information processing systems, 2002, pp. 1523–1530.
[243] J. R. Kok, N. Vlassis, Sparse cooperative Q-learning, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 61.
[244] J. Castellini, F. A. Oliehoek, R. Savani, S. Whiteson, The Representational Capacity of Action-Value Networks for Multi-Agent Reinforcement Learning, in: 18th International Conference on Autonomous Agents and Multiagent Systems, 2019.
[245] P. J. Gmytrasiewicz, P. Doshi, A framework for sequential planning in multiagent settings, Journal of Artificial Intelligence Research 24 (1) (2005) 49–79.
[246] J. C. Harsanyi, Games with incomplete information played by Bayesian players, I–III Part I. The basic model, Management science 14 (3) (1967) 159–182.
[247] S. Barrett, P. Stone, S. Kraus, A. Rosenfeld, Teamwork with Limited Knowledge of Teammates., in: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, WS, USA, 2013, pp. 102–108.
[248] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, et al., Adaptive mixtures of local experts., Neural computation 3 (1) (1991) 79–87.
[249] J. Heinrich, M. Lanctot, D. Silver, Fictitious self-play in extensive-form games, in: International Conference on Machine Learning, 2015, pp. 805–813.
[250] J. F. Nash, Equilibrium points in n-person games, Proceedings of the National Academy of Sciences 36 (1) (1950) 48–49.
[251] J. Von Neumann, O. Morgenstern, Theory of games and economic behavior, Vol. 51, Bull. Amer. Math. Soc, 1945.
[252] J. S. Shamma, G. Arslan, Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria, IEEE Transactions on Automatic Control 50 (3) (2005) 312–327.
[253] W. E. Walsh, R. Das, G. Tesauro, J. O. Kephart, Analyzing complex strategic interactions in multi-agent systems, AAAI-02 Workshop on Game-Theoretic and Decision-Theoretic Agents (2002) 109–118.
[254] M. Johanson, K. Waugh, M. Bowling, M. Zinkevich, Accelerating best response calculation in large extensive games, in: Twenty-Second International Joint Conference on Artificial Intelligence, 2011.
[255] C. F. Camerer, T.-H. Ho, J.-K. Chong, A cognitive hierarchy model of games, The Quarterly Journal of Economics 119 (3) (2004) 861.
[256] M. Costa Gomes, V. P. Crawford, B. Broseta, Cognition and Behavior in Normal–Form Games: An Experimental Study, Econometrica 69 (5) (2001) 1193–1235.
[257] J. Morimoto, K. Doya, Robust reinforcement learning, Neural computation 17 (2) (2005) 335–359.
[258] L. Pinto, J. Davidson, R. Sukthankar, A. Gupta, Robust adversarial reinforcement learning, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2817–2826.
[259] R. Powers, Y. Shoham, T. Vu, A general criterion and an algorithmic framework for learning in multi-agent systems, Machine Learning 67 (1-2) (2007) 45–76.
[260] J. W. Crandall, M. A. Goodrich, Learning to compete, coordinate, and cooperate in repeated games using

reinforcement learning, Machine Learning 82 (3) (2011) 281–314.
[261] M. Johanson, M. A. Zinkevich, M. Bowling, Computing Robust Counter-Strategies., in: Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 2007, pp. 721–728.
[262] P. McCracken, M. Bowling, Safe strategies for agent modelling in games, in: AAAI Fall Symposium, 2004, pp. 103–110.
[263] S. Damer, M. Gini, Safely using predictions in general-sum normal form games, in: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, Sao Paulo, 2017.
[264] C. Zhang, V. Lesser, Multi-agent learning with policy prediction, in: Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.
[265] P. J. Gmytrasiewicz, E. H. Durfee, Rational Coordination in Multi-Agent Environments, Autonomous Agents and Multi-Agent Systems 3 (4) (2000) 319–350.
[266] C. F. Camerer, T.-H. Ho, J.-K. Chong, Behavioural Game Theory: Thinking, Learning and Teaching, in: Advances in Understanding Strategic Behavior, New York, 2004, pp. 120–180.
[267] D. Carmel, S. Markovitch, Incorporating opponent models into adversary search, in: AAAI/IAAI, Vol. 1, 1996, pp. 120–125.
[268] H. de Weerd, R. Verbrugge, B. Verheij, How much does it help to know what she knows you know? An agent-based simulation study, Artificial Intelligence 199-200 (C) (2013) 67–92.
[269] P. Hernandez-Leal, M. Kaisers, Towards a Fast Detection of Opponents in Repeated Stochastic Games, in: G. Sukthankar, J. A. Rodriguez-Aguilar (Eds.), Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Best Papers, Sao Paulo, Brazil, May 8-12, 2017, Revised Selected Papers, 2017, pp. 239–257.
[270] P. Hernandez-Leal, M. E. Taylor, B. Rosman, L. E. Sucar, E. Munoz de Cote, Identifying and Tracking Switching, Non-stationary Opponents: a Bayesian Approach, in: Multiagent Interaction without Prior Coordination Workshop at AAAI, Phoenix, AZ, USA, 2016.
[271] B. Rosman, M. Hawasly, S. Ramamoorthy, Bayesian Policy Reuse, Machine Learning 104 (1) (2016) 99–127.
[272] P. Hernandez-Leal, Y. Zhan, M. E. Taylor, L. E. Sucar, E. Munoz de Cote, Efficiently detecting switches against non-stationary opponents, Autonomous Agents and Multi-Agent Systems 31 (4) (2017) 767–789.
[273] P. Hernandez-Leal, M. Kaisers, Learning against sequential opponents in repeated stochastic games, in: The 3rd Multi-disciplinary Conference on Reinforcement Learning and Decision Making, Ann Arbor, 2017.
[274] J. Schmidhuber, Critique of Paper by "Deep Learning Conspiracy" (Nature 521 p 436), http://people.idsia.ch/~juergen/deep-learning-conspiracy.html (2015).
[275] Do I really have to cite an arXiv paper?, http://approximatelycorrect.com/2017/08/01/do-i-have-to-cite-arxiv-paper/, [Online; accessed 21-May-2019] (2017).
[276] Collaboration & Credit Principles, How can we be good stewards of collaborative trust?, http://colah.github.io/posts/2019-05-Collaboration/index.html, [Online; accessed 31-May-2019] (2019).
[277] H. Wang, B. Raj, E. P. Xing, On the origin of deep learning, CoRR abs/1702.07800. arXiv:1702.07800. URL http://arxiv.org/abs/1702.07800
[278] A. K. Agogino, K. Tumer, Analyzing and visualizing multiagent rewards in dynamic and stochastic domains, Autonomous Agents and Multi-Agent Systems 17 (2) (2008) 320–338.
[279] S. Devlin, L. M. Yliniemi, D. Kudenko, K. Tumer, Potential-based difference rewards for multiagent reinforcement learning., in: 13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014, Paris, France, 2014.
[280] D. H. Wolpert, K. R. Wheeler, K. Tumer, General principles of learning-based multi-agent systems, in: Proceedings of the Third International Conference on Autonomous Agents, 1999.
[281] M. E. Taylor, P. Stone, Transfer learning for reinforcement learning domains: A survey, The Journal of Machine Learning Research 10 (2009) 1633–1685.
[282] C. Bucilua, R. Caruana, A. Niculescu-Mizil, Model compression, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 535–541.
[283] Y. Du, W. M. Czarnecki, S. M. Jayakumar, R. Pascanu, B. Lakshminarayanan, Adapting auxiliary losses using gradient similarity, arXiv preprint arXiv:1812.02224.
[284] S. C. Suddarth, Y. Kergosien, Rule-injection hints as a means of improving network performance and learning time, in: Neural Networks, Springer, 1990, pp. 120–129.
[285] P. Hernandez-Leal, B. Kartal, M. E. Taylor, Agent Modeling as Auxiliary Task for Deep Reinforcement Learning, in: AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2019.
[286] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, W. Zaremba, Hindsight experience replay, in: Advances in Neural Information Processing Systems, 2017.
[287] Z. C. Lipton, K. Azizzadenesheli, A. Kumar, L. Li, J. Gao, L. Deng, Combating Reinforcement Learning's

Sisyphean Curse with Intrinsic Fear.
URL http://arxiv.org/abs/1611.01211v8

[288] T. De Bruin, J. Kober, K. Tuyls, R. Babuška, Experience selection in deep reinforcement learning for control, The Journal of Machine Learning Research 19 (1) (2018) 347–402.

[289] D. S. Bernstein, R. Givan, N. Immerman, S. Zilberstein, The complexity of decentralized control of Markov decision processes, Mathematics of operations research 27 (4) (2002) 819–840.

[290] F. A. Oliehoek, C. Amato, et al., A concise introduction to decentralized POMDPs, Springer, 2016.

[291] F. A. Oliehoek, M. T. Spaan, N. Vlassis, Optimal and approximate Q-value functions for decentralized POMDPs, Journal of Artificial Intelligence Research 32 (2008) 289–353.

[292] J. K. Gupta, M. Egorov, M. J. Kochenderfer, Cooperative Multi-agent Control using deep reinforcement learning, in: Adaptive Learning Agents at AAMAS, Sao Paulo, 2017.

[293] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: International conference on machine learning, 2013, pp. 1310–1318.

[294] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, J. Schmidhuber, LSTM: A Search Space Odyssey, IEEE Transactions on Neural Networks and Learning Systems 28 (10) (2017) 2222–2232.

[295] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: Deep Learning and Representation Learning Workshop, 2014.

[296] S. Whiteson, B. Tanner, M. E. Taylor, P. Stone, Protecting against evaluation overfitting in empirical reinforcement learning, in: 2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), IEEE, 2011, pp. 120–127.

[297] M. G. Bellemare, Y. Naddaf, J. Veness, M. Bowling, The arcade learning environment: An evaluation platform for general agents, Journal of Artificial Intelligence Research 47 (2013) 253–279.

[298] M. C. Machado, M. G. Bellemare, E. Talvitie, J. Veness, M. Hausknecht, M. Bowling, Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents, Journal of Artificial Intelligence Research 61 (2018) 523–562.

[299] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba, OpenAI Gym, arXiv preprint arXiv:1606.01540.

[300] P. S. Castro, S. Moitra, C. Gelada, S. Kumar, M. G. Bellemare, Dopamine: A Research Framework for Deep Reinforcement Learning.
URL http://arxiv.org/abs/1812.06110

[301] C. Resnick, W. Eldridge, D. Ha, D. Britz, J. Foerster, J. Togelius, K. Cho, J. Bruna, Pommerman: A Multi-Agent Playground.
URL http://arxiv.org/abs/1809.07124

[302] C. Gao, B. Kartal, P. Hernandez-Leal, M. E. Taylor, On Hard Exploration for Reinforcement Learning: a Case Study in Pommerman, in: AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2019.

[303] M. Samvelyan, T. Rashid, C. S. de Witt, G. Farquhar, N. Nardelli, T. G. J. Rudner, C. Hung, P. H. S. Torr, J. N. Foerster, S. Whiteson, The StarCraft Multi-Agent Challenge, CoRR abs/1902.04043. arXiv:1902.04043.
URL http://arxiv.org/abs/1902.04043

[304] D. Pérez-Liébana, K. Hofmann, S. P. Mohanty, N. Kuno, A. Kramer, S. Devlin, R. D. Gaina, D. Ionita, The multi-agent reinforcement learning in Malmö (MARLÖ) competition, CoRR abs/1901.08129. arXiv:1901.08129.
URL http://arxiv.org/abs/1901.08129

[305] M. Johnson, K. Hofmann, T. Hutton, D. Bignell, The Malmo platform for artificial intelligence experimentation., in: IJCAI, 2016, pp. 4246–4247.

[306] P. Stone, G. Kaminka, S. Kraus, J. S. Rosenschein, Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination., in: 32nd AAAI Conference on Artificial Intelligence, Atlanta, Georgia, USA, 2010, pp. 1504–1509.

[307] M. Bowling, P. McCracken, Coordination and adaptation in impromptu teams, in: Proceedings of the Nineteenth Conference on Artificial Intelligence, Vol. 5, 2005, pp. 53–58.

[308] S. V. Albrecht, S. Ramamoorthy, A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems, in: Proceedings of the 12th International Conference on Autonomous Agents and Multi-agent Systems, Saint Paul, MN, USA, 2013.

[309] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, I. Dunning, S. Mourad, H. Larochelle, M. G. Bellemare, M. Bowling, The Hanabi Challenge: A New Frontier for AI Research.

URL https://arxiv.org/abs/1902.00506
[310] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, D. Silver, Rainbow: Combining improvements in deep reinforcement learning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[311] Y. Song, J. Wang, T. Lukasiewicz, Z. Xu, M. Xu, Z. Ding, L. Wu, Arena: A general evaluation platform and building toolkit for multi-agent intelligence, CoRR abs/1905.08085. arXiv:1905.08085.
URL http://arxiv.org/abs/1905.08085
[312] A. Juliani, V. Berges, E. Vckay, Y. Gao, H. Henry, M. Mattar, D. Lange, Unity: A general platform for intelligent agents, CoRR abs/1809.02627. arXiv:1809.02627.
URL http://arxiv.org/abs/1809.02627
[313] S. Liu, G. Lever, J. Merel, S. Tunyasuvunakool, N. Heess, T. Graepel, Emergent coordination through competition, in: International Conference on Learning Representations, 2019.
[314] J. Suarez, Y. Du, P. Isola, I. Mordatch, Neural MMO: A massively multiagent game environment for training and evaluating intelligent agents, CoRR abs/1903.00784. arXiv:1903.00784.
URL http://arxiv.org/abs/1903.00784
[315] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep Reinforcement Learning That Matters., in: 32nd AAAI Conference on Artificial Intelligence, 2018.
[316] P. Nagarajan, G. Warnell, P. Stone, Deterministic implementations for reproducibility in deep reinforcement learning.
URL http://arxiv.org/abs/1809.05676
[317] K. Clary, E. Tosch, J. Foley, D. Jensen, Let's play again: Variability of deep reinforcement learning agents in Atari environments, in: NeurIPS Critiquing and Correcting Trends Workshop, 2018.
[318] J. Z. Forde, M. Paganini, The scientific method in the science of machine learning, in: ICLR Debugging Machine Learning Models workshop, 2019.
[319] K. Azizzadenesheli, Maybe a few considerations in reinforcement learning research?, in: Reinforcement Learning for Real Life Workshop, 2019.
[320] Z. C. Lipton, J. Steinhardt, Troubling trends in machine learning scholarship, in: ICML Machine Learning Debates workshop, 2018.
[321] R. Rosenthal, The file drawer problem and tolerance for null results., Psychological bulletin 86 (3) (1979) 638.
[322] D. Sculley, J. Snoek, A. Wiltschko, A. Rahimi, Winner's curse? on pace, progress, and empirical rigor, in: ICLR Workshop, 2018.
[323] O. Gencoglu, M. van Gils, E. Guldogan, C. Morikawa, M. Süzen, M. Gruber, J. Leinonen, H. Huttunen, Hark side of deep learning–from grad student descent to automated machine learning, arXiv preprint arXiv:1904.07633.
[324] K. Azizzadenesheli, B. Yang, W. Liu, E. Brunskill, Z. Lipton, A. Anandkumar, Surprising negative results for generative adversarial tree search, in: Critiquing and Correcting Trends in Machine Learning Workshop, 2018.
[325] C. Lyle, P. S. Castro, M. G. Bellemare, A comparative analysis of expected and distributional reinforcement learning, in: Thirty-Third AAAI Conference on Artificial Intelligence, 2019.
[326] B. Kartal, P. Hernandez-Leal, M. E. Taylor, Using Monte Carlo tree search as a demonstrator within asynchronous deep RL, in: AAAI Workshop on Reinforcement Learning in Games, 2019.
[327] G. Melis, C. Dyer, P. Blunsom, On the state of the art of evaluation in neural language models, in: International Conference on Learning Representations, 2018.
[328] Deep Reinforcement Learning: Pong from Pixels, https://karpathy.github.io/2016/05/31/rl/, [Online; accessed 7-May-2019] (2016).
[329] V. Firoiu, W. F. Whitney, J. B. Tenenbaum, Beating the World's Best at Super Smash Bros. with Deep Reinforcement Learning, CoRR abs/1702.06230. arXiv:1702.06230.
URL http://arxiv.org/abs/1702.06230
[330] C. Gao, P. Hernandez-Leal, B. Kartal, M. E. Taylor, Skynet: A Top Deep RL Agent in the Inaugural Pommerman Team Competition, in: 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making, 2019.
[331] D. Amodei, D. Hernandez, AI and Compute (2018).
URL https://blog.openai.com/ai-and-compute
[332] Y. Yu, Towards sample efficient reinforcement learning., in: IJCAI, 2018, pp. 5739–5743.
[333] F. P. Such, V. Madhavan, E. Conti, J. Lehman, K. O. Stanley, J. Clune, Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning, CoRR abs/1712.06567.

URL http://arxiv.org/abs/1712.06567
[334] A. Stooke, P. Abbeel, Accelerated methods for deep reinforcement learning, CoRR abs/1803.02811. arXiv:1803.02811.
URL http://arxiv.org/abs/1803.02811
[335] M. Babaeizadeh, I. Frosio, S. Tyree, J. Clemons, J. Kautz, Reinforcement learning through asynchronous advantage actor-critic on a GPU, in: International Conference on Learning Representations, 2017.
[336] W. H. Guss, C. Codel, K. Hofmann, B. Houghton, N. Kuno, S. Milani, S. P. Mohanty, D. P. Liebana, R. Salakhutdinov, N. Topin, M. Veloso, P. Wang, The MineRL Competition on Sample Efficient Reinforcement Learning using Human Priors, CoRR abs/1904.10079. arXiv:1904.10079.
URL http://arxiv.org/abs/1904.10079
[337] G. Cuccu, J. Togelius, P. Cudré-Mauroux, Playing Atari with six neurons, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 998–1006.
[338] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, J. Clune, Go-explore: a new approach for hard-exploration problems, arXiv preprint arXiv:1901.10995.
[339] R. I. Brafman, M. Tennenholtz, R-max-a general polynomial time algorithm for near-optimal reinforcement learning, Journal of Machine Learning Research 3 (Oct) (2002) 213–231.
[340] A. L. Strehl, M. L. Littman, An analysis of model-based interval estimation for Markov decision processes, Journal of Computer and System Sciences 74 (8) (2008) 1309–1331.
[341] J. Schmidhuber, A possibility for implementing curiosity and boredom in model-building neural controllers, in: Proc. of the international conference on simulation of adaptive behavior: From animals to animats, 1991, pp. 222–227.
[342] A. G. Barto, Intrinsic motivation and reinforcement learning, in: Intrinsically motivated learning in natural and artificial systems, Springer, 2013, pp. 17–47.
[343] T. D. Kulkarni, K. Narasimhan, A. Saeedi, J. Tenenbaum, Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation, in: Advances in neural information processing systems, 2016, pp. 3675–3683.
[344] T. G. Dietterich, Ensemble Methods in Machine Learning, in: MCS Proceedings of the First International Workshop on Multiple Classifier Systems, Springer Berlin Heidelberg, Cagliari, Italy, 2000, pp. 1–15.
[345] J. A. Arjona-Medina, M. Gillhofer, M. Widrich, T. Unterthiner, S. Hochreiter, RUDDER: Return Decomposition for Delayed Rewards.
URL http://arxiv.org/abs/1806.07857
[346] V. Conitzer, T. Sandholm, AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents, Machine Learning 67 (1-2) (2006) 23–43.
[347] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, S. Colton, A survey of Monte Carlo tree search methods, IEEE Transactions on Computational Intelligence and AI in games 4 (1) (2012) 1–43.
[348] T. Vodopivec, S. Samothrakis, B. Ster, On Monte Carlo tree search and reinforcement learning, Journal of Artificial Intelligence Research 60 (2017) 881–936.
[349] B. Kartal, J. Godoy, I. Karamouzas, S. J. Guy, Stochastic tree search with useful cycles for patrolling problems, in: Robotics and Automation (ICRA), 2015 IEEE International Conference on, IEEE, 2015, pp. 1289–1294.
[350] B. Kartal, E. Nunes, J. Godoy, M. Gini, Monte Carlo tree search with branch and bound for multi-robot task allocation, in: The IJCAI-16 Workshop on Autonomous Mobile Service Robots, 2016.
[351] G. Best, O. M. Cliff, T. Patten, R. R. Mettu, R. Fitch, Dec-MCTS: Decentralized planning for multi-robot active perception, The International Journal of Robotics Research 38 (2-3) (2019) 316–337.
[352] Y.-M. De Hauwere, P. Vrancx, A. Nowe, Learning multi-agent state space representations, in: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada, 2010, pp. 715–722.
[353] C. Guestrin, M. Lagoudakis, R. Parr, Coordinated reinforcement learning, in: ICML, Vol. 2, 2002, pp. 227–234.
[354] C. Guestrin, D. Koller, R. Parr, S. Venkataraman, Efficient solution algorithms for factored MDPs, Journal of Artificial Intelligence Research 19 (2003) 399–468.
[355] C. Amato, F. A. Oliehoek, Scalable Planning and Learning for Multiagent POMDPs, in: AAAI, 2015, pp. 1995–2002.
[356] F. A. Oliehoek, Interactive Learning and Decision Making - Foundations, Insights & Challenges., International Joint Conference on Artificial Intelligence.
[357] F. A. Oliehoek, S. Whiteson, M. T. Spaan, Approximate solutions for factored Dec-POMDPs with many

agents, in: Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 563–570.
[358] R. Becker, S. Zilberstein, V. Lesser, C. V. Goldman, Solving transition independent decentralized Markov decision processes, Journal of Artificial Intelligence Research 22 (2004) 423–455.
[359] F. A. Oliehoek, S. J. Witwicki, L. P. Kaelbling, Influence-based abstraction for multiagent systems, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[360] M. Suau de Castro, E. Congeduti, R. A. Starre, A. Czechowski, F. A. Oliehoek, Influence-based abstraction in deep reinforcement learning, in: Adaptive, Learning Agents workshop, 2019.
[361] E. Wei, D. Wicke, D. Freelan, S. Luke, Multiagent Soft Q-Learning.
URL http://arXiv.org/abs/1804.09817
[362] R. R. Torrado, P. Bontrager, J. Togelius, J. Liu, D. Perez-Liebana, Deep Reinforcement Learning for General Video Game AI.
URL http://arxiv.org/abs/1806.02448
[363] P. A. Ortega, S. Legg, Modeling friends and foes.
URL http://arxiv.org/abs/1807.00196
[364] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, J. Wang, Mean field multi-agent reinforcement learning, in: Proceedings of the 35th International Conference on Machine Learning, Stockholm Sweden, 2018.
[365] A. Grover, M. Al-Shedivat, J. K. Gupta, Y. Burda, H. Edwards, Learning Policy Representations in Multiagent Systems., in: International Conference on Machine Learning, 2018.
[366] C. K. Ling, F. Fang, J. Z. Kolter, What game are we playing? end-to-end learning in normal and extensive form games, in: Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.
[367] S. Omidshafiei, D. Hennes, D. Morrill, R. Munos, J. Perolat, M. Lanctot, A. Gruslys, J.-B. Lespiau, K. Tuyls, Neural Replicator Dynamics, arXiv e-prints (2019) arXiv:1906.00190arXiv:1906.00190.
[368] S. Khadka, S. Majumdar, K. Tumer, Evolutionary Reinforcement Learning for Sample-Efficient Multiagent Coordination, arXiv e-prints (2019) arXiv:1906.07315arXiv:1906.07315.