Annual Reviews in Control 53 (2022) 296–314

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

Review

The role of information structures in game-theoretic multi-agent learning

Tao Li *, Yuhan Zhao, Quanyan Zhu

Department of Electrical and Computer Engineering, New York University, NY, 11201, United States of America

ARTICLE INFO

Keywords:
Multi-agent learning
Information structures
Reinforcement learning
Belief generation
Game theory
Value of Information

ABSTRACT

Multi-agent learning (MAL) studies how agents learn to behave optimally and adaptively from their experience when interacting with other agents in dynamic environments. The outcome of a MAL process is jointly determined by all agents' decision-making. Hence, each agent needs to think strategically about others' sequential moves, when planning future actions. The strategic interactions among agents makes MAL go beyond the direct extension of single-agent learning to multiple agents. With the strategic thinking, each agent aims to build a subjective model of others decision-making using its observations. Such modeling is directly influenced by agents' perception during the learning process, which is called the information structure of the agent's learning. As it determines the input to MAL processes, information structures play a significant role in the learning mechanisms of the agents. This review creates a taxonomy of MAL and establishes a unified and systematic way to understand MAL from the perspective of information structures. We define three fundamental components of MAL: the information structure (i.e., what the agent can observe), the belief generation (i.e., how the agent forms a belief about others based on the observations), as well as the policy generation (i.e., how the agent generates its policy based on its belief). In addition, this taxonomy enables the classification of a wide range of state-of-the-art algorithms into four categories based on the belief-generation mechanisms of the opponents, including *stationary, conjectured, calibrated*, and *sophisticated opponents*. We introduce *Value of Information* (VoI) as a metric to quantify the impact of different information structures on MAL. Finally, we discuss the strengths and limitations of algorithms from different categories and point to promising avenues of future research.

1. Introduction

Multi-agent systems (MAS) are distributed systems involving a group of intelligent and autonomous entities called agents (Wooldridge, 2009). Agents perceive the environment[1] to understand the context and make decisions to achieve certain tasks and objectives. The history of MAS can be traced back to the 1980s, when the research in Distributed Artificial Intelligence (DAI) (Bond & Gasser, 2014; O'Hare & Jennings, 1996) prevailed, which focuses on modeling systems with multiple intelligent agents as well as coordinating their behaviors and complex interactions with the environment (Dorri, Kanhere, & Jurdak, 2018; Stone & Veloso, 2000).

Stemmed from the studies of MAS, Multi-Agent Learning (MAL) mainly focuses on applying learning-based methods to MAS problems. More formally, MAL studies how an intelligent agent learns to behave optimally and adaptively from its experience with the presence of other agents in dynamic environments (Tuyls & Weiss, 2012). Unlike techniques from distributed optimization and control, the learning-based methods aim to equip MAS with distributed intelligence that responds to uncertainties, anomalies, and disruptions to achieve the desired coordination of the agents within the system.

1.1. A brief history of multi-agent learning

In the early stage of MAL, machine learning was used to address challenges in MAS. Many techniques from different fields were studied and developed in this stage, such as distributed sensing and fusion (Luo & Kay, 1992; Mataric, 1998), herd behavior (Banerjee, 1992; Colorni, Dorigo, Maniezzo, et al., 1991), social learning (Coussi-Korbel & Fragaszy, 1995), evolutionary computation and games (Beer & Gallagher, 1992; Fogel, 1995; Weibull, 1997), and artificial neural networks (Hertz, Krogh, & Palmer, 1991; Yao, 1999). Meanwhile, in addition to the endeavors from the machine learning community, game theorists, economists, and biologists were also keen on the research of learning in games, which brought new interpretations of the equilibrium concepts and related results in evolutionary biology (Bowling & Veloso, 2002; Fudenberg, Drew, Levine, & Levine, 1998; Hofbauer &

* Corresponding author.
E-mail addresses: tl2636@nyu.edu (T. Li), yz5718@nyu.edu (Y. Zhao), qz494@nyu.edu (Q. Zhu).

[1] For a single agent, all the counterpart agents are perceived as part of the environment.

https://doi.org/10.1016/j.arcontrol.2022.03.003
Received 1 July 2021; Accepted 7 March 2022
Available online 26 March 2022

Sigmund, 2003). This booming stage ranging from the late 1980s to 2000 has greatly enriched the research scope and topics in MAL, and it is named as the *startup period* by Tuyls and Weiss (2012).

With the advances in single-agent reinforcement learning (RL) (Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 2018), MAL has ushered in another prosperity. More work such as Bowling and Veloso (2002), Hu and Wellman (2003) and Littman (2001) began to focus on the intersection of RL and MAL, which pushes MAL to a new stage called the *consolidation period* (Tuyls & Weiss, 2012). Traditional RL studies the single-agent learning scenario, where an agent learns the optimal policy for maximizing long-term return under the framework of the Markov Decision Processes (MDP) (Puterman, 2014). The optimal policy is learned from the agent's past interaction history through either value-based approaches, e.g., Q-learning (Watkins & Dayan, 1992), or policy-based ones, e.g., policy gradient (Silver et al., 2014).

The success of single-agent RL motivates the technique to be extended from single-agent RL to multi-agent cases. Naturally, this extension has to take the interactions among multiple agents and the environment into consideration. Most of the existing literature adopts game-theoretic models as formal frameworks (Tuyls & Weiss, 2012), which quantitatively depict how the interactions of independent agents with different information lead to coordinated behaviors at a system level. Compared with the first stage, the research in the consolidation period is more like a depth-first exploration characterized by a focus on RL theory in a game-theoretic context, which dominates the current MAL field (Tuyls & Weiss, 2012). Moreover, with the recent advances in deep learning (LeCun, Bengio, & Hinton, 2015), the combination of MAL algorithms and deep-learning-based function approximators provides practical solutions to many long-standing problems (Jaderberg et al., 2019; Mnih et al., 2015).

However, MAL challenges go beyond the direct application of single-agent learning. From a game-theoretic viewpoint, the outcome of a MAL process is jointly determined by every agent's sequential moves, and hence, when planning future actions, each agent needs to take into account others' decision-making as well. The strategic interactions among agents make MAL a research direction in its own right, rather than an extension of single-agent learning to strategically interacting agents.

This strategic thinking fundamentally differentiates MAL from single-agent learning. With the presence of strategic interactions, each agent in MAL is required to model others' decision-making and predict their future moves. Since agents' decision-making processes are unknown to each other, each individual can only build a subjective model or an estimate of others' strategies, using its own observations. Therefore, such modeling and estimation are directly influenced by agents' perception during the learning process, which further influences strategies used by agents, and hence, the learning outcome. To sum up, what information an agent receives, as the input to its decision-making, plays an important part in its learning process, and its role in MAL is even more significant, as it helps the agent quantify others' concurrent decision-making, leading to a strategic learning. In the following subsection, we take a closer look at the role of information played in MAL, and argue that a theoretical underpinning of information structure is necessary to the future development of MAL.

1.2. The role of information structures in MAL

For each agent in MAL, information refers to a set of random variables whose realizations can be observed by the agent. For example, for agents with full state observations in RL, the realization of state variable can be observed, and accordingly, the state variable belongs to the information received by these agents. Because of the possible spatial and temporal structures of the information an agent receive at each time instance, we refer to this set of observable variables as the information structure of the agent. A more mathematical characterization of information structures is provided in Definition 2, and the associated spatial and temporal structures are discussed in Section 3.

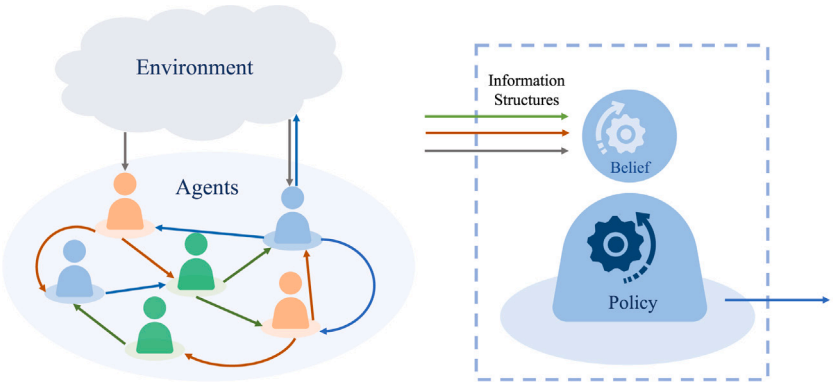
Following the discussion in Section 1.1, there are three stages within each agent's decision-making at each round of interactions: (1) the agent receives observations according to its information structure; (2) the agent forms a belief about others' decision-making using its observations; (3) based on its belief, the agent implements an action, which leads to new observations for the next round. A schematic illustration of this MAL process is provided in Fig. 1. The high performance of each learning agent requires beliefs as consistent as possible with the ground truth. The measure of the consistency depends on the information structure of the agent. A thorough investigation into the role of information structures in learning is indispensable for studying and designing MAL algorithms. In the following, we take three major challenges in MAL to elaborate how a deeper understanding of information structures can contribute to the future development of MAL both in theory and application. Other related challenges and future research directions are discussed in Section 5.3.

Heterogeneity. The heterogeneity of a MAS refers to the fact that agents within the same system may possess distinct learning capabilities, and operate under different information structures. In contrast with MAS with homogeneous agents, the analysis of agents' limiting behavior and the stabilized system outcome is more involved, since the dynamical system corresponding to the heterogeneous MAL becomes highly nonlinear, coupled and possibly time-varying. Existing techniques, such as Lyapunov methods, are not directly applicable.

One way to deal with this heterogeneity is to classify agents according to the information structure. For example, agents sharing the same information structure or reward structure can be labeled as one type (Sunehag et al., 2018; Tang, Tavafoghi, Subramanian, Nayyar, & Teneketzis, 2021), and the original system reduces to a much simplified MAS where each type is treated as a decision-maker, leading to a population-based MAL (Tembine, Zhu, & Baar, 2014). By examining heterogeneous MAS at a coarse scale, theoretical analysis of MAL becomes tractable, and the adaptability of heterogeneous agents under various information structures leads to system-level resiliency (Zhu, Tembine, & Baar, 2010). A detailed discussion on this topic is present in Section 5.3.2.

Non-stationarity. Non-stationarity often arises from the dynamically changing environment, and the concurrent learning of agents. Each agent constantly adjusts its strategy to adapt to other agents, and hence, from each agent's perspective, the transition probability from one state to another is no longer stationary, and is affected by other agents. If the agent simply ignores this non-stationarity and fails to adapt to the changing environment, it can be exploited by its opponents during the strategic interactions (Conitzer & Sandholm, 2007).

The key to combat this non-stationarity relies on the agent forming proper beliefs on others' decision-making. Since other agents' decision-making is hidden, the discrepancy between beliefs and true strategies employed by others cannot be directly observed by the agent. Therefore, the best one can do is to ensure that one's belief is consistent with one's observations. As the information structure dictates the environmental feedback received by the agent, it influences the measure of consistency, leading to various belief generation and calibration processes in MAL. Therefore, a thorough understanding of information structures allows for consistent conjectures on the non-stationary dynamics of external environment experienced by each agent, and bring up distributed intelligence in MAS that is responsive to uncertainties, anomalies and disruptions. Section 4 provides a comprehensive elaboration on the belief generation and calibration under different information structures.



(a) In multi-agent learning, agents adapt to the dynamic environment based on their perceptions of surroundings.

(b) The interconnection between information structures, the belief generation, and the policy generation.

Fig. 1. A schematic illustration of multi-agent learning (MAL). At each round of interactions, the agent first forms a belief about others' strategies using its observations about the environment and other agents. The observations, represented by arrows, are subject to information structures of MAL. Then, based on its belief, the agent generates a policy to be implemented.

Scalability. In MAL, one agent's learning depends on the decision-making processes of others. In this case, forming beliefs may require information regarding joint actions and rewards of all agents (Littman, 1994). The dimension of the joint action space increases exponentially with the number of agents, resulting in prohibitive sample and computation complexity (Zhang, Yang, & Baar, 2019). The study of information structures in MAL can create possible solutions to address this long standing challenge. With a proper design of information structures, agents may not need global information regarding the whole system to correctly form beliefs. For example, communications with neighboring agents in MAS creates information diffusion over networks, yielding efficient coordination (Liu, Li, & Zhu, 2020; Zhang, Yang, Liu, Zhang, & Baar, 2018). In addition to information sharing, a more structured design information structures can facilitate knowledge reuse and transfer among agents (Bannon, Windsor, Song, & Li, 2020), where latent variables are taken from direct observations as the learned knowledge, and then are shared across the MAS.

There are a few initial attempts to evaluate its impact on MAL in the literature. In Lowe, Foerster, Boureau, Pineau, and Dauphin (2019) and Naghizadeh, Gorlatova, Lan, and Chiang (2019), the effects of message passing or communication among agents are evaluated, identifying benefits and drawbacks of communication in MAL. Ouyang, Tavafoghi, and Teneketzis (2016) investigates the signaling effect of each agent's action under asymmetric information structures, where agents' actions reflect its hidden information. There is also a growing body of works addressing MAL with partial state observations, and we refer the reader to a recent review by Hernandez-Leal, Kartal, and Taylor (2019) on this topic. MAL under each information structure is a subfield of studies, and each of these initial endeavors contributed to a subfield of their own. However, efforts crossing the subfields and creating a unifying approach is missing. Furthermore, there is no mathematical formalism that can lay a theoretical foundation for future discussion. The holistic and systematic treatment of information structures in MAL has remained largely unexplored, and there is a need to better understanding the role played by information structures in MAL both qualitatively and quantitatively.

By proposing a mathematical characterization of information structures in the context of MAL, this paper takes the first step toward a theoretical underpinning of information structures, and unifying many subfields, including partially observable MDP, networked control theory, MAS under asymmetric information, and other related studies. Such mathematical formulation can facilitate future studies on this topic, leading to a holistic viewpoint on the impact of information structures on MAL. Based on this characterization, we examine existing MAL algorithms, and categorize these state-of-the-art works according to their different treatments on information structures and belief generation processes. On the other hand, similarities of learning algorithms across many subfields are also summarized, highlighting their strengths and limitations on handling various information structures.

To qualitatively compare different information structures, we introduce *Information Superiority*, a relation that establishes a partial order among different information structures, indicating how much information an information structure contains during the learning. A subsequent thought experiment provides an interesting finding that the agent's learning performance does not increase monotonically with respect to information superiority: more information does not necessarily lead to better learning performance. We refer to this non-monotonicity as *Information Paradox*. This paradox further necessitate the introduction of another metric called *Value of Information*, which quantitatively measures the impact of information structures on the agent's learning with respect to its average rewards. The information perspective proposed in this paper provides a unified view of recent MAL advancements, and together with the introduced metrics, it creates a stepping stone to address long-standing MAL challenges, including heterogeneity, non-stationarity, scalability as well as other emerging challenges presented in Section 5.3.

1.3. Our contributions

In this paper, we create a taxonomy of MAL based on information structures and establish a unified framework to capture MAL in structurally diverse settings, including Markov games, repeated games, extensive-form games, and multi-armed bandits. This review provides a systematic overview of the literature from the perspective of information structures, aiming to contribute to MAL in the following directions.

1. We show that an MAL algorithm comprises three components: the information structure, the belief generation, and the policy generation. This unified perspective provides a coherent view of state-of-the-art MAL algorithms.
2. We categorize the MAL algorithms into four categories, i.e., stationary opponents, conjectured opponents, calibrated opponents, and sophisticated opponents, depending on how the belief about the opponent is generated. For different belief generation processes, we provide concrete examples to illustrate the interconnection between belief generation and the information structure.

3. We formally define the information structure in MAL, which quantitatively specifies the influence of the information on the learning process via belief generation. The proposed framework paves the way for a systematic discussion of the strengths and limitations of algorithms within the four categories under various information structures.
4. We use the information structure as a theoretic underpinning to facilitate the description and presentation of MAL algorithms and discuss open questions in MAL, such as the heterogeneity of MAS, the scalability issues, and novel learning objectives.

1.4. Related taxonomies

Along with its development, many taxonomies have been proposed to understand MAL from different perspectives. In the following, we briefly review four of them. Stone and Veloso (2000) have categorized the MAS algorithms into four classes based on the degree of homogeneity and degree of communication. Potential learning methods and opportunities are discussed within each category. Although the taxonomy is based on works published before 2000, it provides the first systematic view of MAL.

As the research is growing in MAL in recent years, many new learning-oriented algorithms are being proposed. To summarize the contribution and challenges, Busoniu, Babuska, and De Schutter (2008) have proposed two taxonomies to classify MAL. One is based on the learning task type, and the other is based on the degree of agent awareness. For the task-based taxonomy, MAL is categorized into fully cooperative cases, fully competitive cases, and mixed cases. All agents cooperate to achieve the same objective in fully cooperative cases, while all agents optimize their objectives respectively in fully competitive cases. Mixed cases lie in the middle. For awareness-based taxonomy, agents in MAL ranges from fully unaware of the environment to fully aware of the environment. These two taxonomies build the foundation for modern MAL research. Many recent works such as Da Silva and Costa (2019) and Zhang et al. (2019) have acknowledged and used these taxonomies to categorize research works in the area.

Another recent taxonomy based on the agent's reaction pattern to the environment has been proposed in Hernandez-Leal, Kaisers, Baarslag, and Cote (2017). In a joint learning task, an agent can react to the environments through five patterns: ignore, forget, respond to target opponents, learn opponent models, and theory of mind. In each pattern, an agent chooses different actions to respond to the environment. For example, *ignoring* means a static environment while *theory of mind* refers to the recursive reasoning in the learning.

The rest of the paper is organized as follows. We formally introduce our proposed definition for MAL in Section 2, where an MAL algorithm is said to be comprised of the information structure, the belief generation as well as the policy generation. The detail of the information structure is discussed in Section 3 and Section 4 discusses the belief and policy generation in MAL. Based on different approaches in generating beliefs, we propose a categorization of MAL algorithms in Section 4. Following this categorization, a systematic discussion on the strengths and limitations of MAL algorithms from the proposed categories is provided in Section 5. Furthermore, we propose a metric called the value of information (VoI) in Section 5 in order to quantitatively describe the importance of the information structure in MAL. Some related applications in the security domain and MAL are reviewed in Section 6, and Section 7 concludes the paper.

2. A mathematical framework for multi-agent learning

This section introduces a mathematical model of MAL, which is a unified framework capable of describing, analyzing, and comparing various MAL algorithms. To facilitate our discussions on the framework, we begin with a formal definition of the multi-agent sequential decision-making problem (MASDM). Then, we proceed to introduce our unified MAL framework, which incorporates other prominent MAL frameworks, such as repeated games (Mertens, Sorin, & Zamir, 2015), Markov games (Shapley, 1953) and extensive-form games (Kuhn, 2016). We elaborate on three important components within this framework: the information structure, the belief generation, and the policy generation. The three components are closely related: the information structure determines what can be observed by the agent at each round of play, which further influences the agent on updating its belief about its opponent's policy.[2] Finally, the generation of future policy depends on the agent's belief.

Following the prescriptive viewpoint in Sandholm (2007), a MAL problem is essentially a sequential decision-making process in MAS under uncertainty. In general, the decision-making problem can be defined as follows.

Definition 1. A multi-agent decision-making process (MASDM) is defined by a tuple $\mathcal{G} = \langle S, \mathcal{N} \cup \{c\}, \{\mathcal{A}_i\}_{i\in\mathcal{N}\cup c}, \mu_c, \mathcal{H}, \tau, \mathcal{T}, \{R_i\}_{i\in\mathcal{N}}\rangle$, where

1. S is the state space;
2. $\mathcal{N} := \{1, 2, \dots, N\}$ denotes the set of N agents, and c is a special agent called chance or nature, which employs a fixed stochastic policy that specifies the randomness of the environment;
3. $\mathcal{A}_i$ is the set of all possible actions that agent i can take;
4. μ_c is nature's fixed policy, which is a probability measure over $\mathcal{A}_c$;
5. $\mathcal{H}$ is the set of all possible histories, where each history h is a sequence of states and actions;
6. $\tau : \mathcal{H} \to \mathcal{N} \cup \{c\}$ is the agent selection function that determines which agent takes the action after a sequence of plays;
7. $\mathcal{T} : \mathcal{H} \to \Delta(S)$ is a transition dynamics that specifies the probability of a certain state is chosen as the next state based on the history;
8. $R_i : \mathcal{H} \to \mathbb{R}$ is the utility function or reward function that determines the payoff or cost agent i receives when the historical play is h.

The length of admissible histories determines the horizon of the decision-making problem. If all possible histories are of finite length, i.e., the finite number of states and actions, then the problem is said to have a finite horizon. Otherwise, it has an infinite horizon.

Several remarks are in order. First, the introduction of the special agent c accounts for multi-agent decision making with incomplete information regarding the agents or, more broadly speaking, risk-related factors involved in decision-making (Park & Shapira, 2017). In other words, the actions of c correspond to random events in the environment subject to a prior distribution. These events are agent-independent in the sense that the occurrence does not depend on agents' actions. In Bayesian games literature (Zamir, 2009), the action space of chance $\mathcal{A}_c$ is often referred to as the type space of agents, elements of which specify the "type" of each agent according to the fixed policy μ_c, also called prior. These types from $\mathcal{A}_c$ accounts for hidden information which is privately revealed to agents, and the corresponding realization is unknown to others.

Another remark is about the history set $\mathcal{H}$. Figuratively speaking, any element $h \in \mathcal{H}$ is a system log that tracks everything that happened within the system, and it may not be observable to the agent. The observability issue will be discussed in the information structure

[2] We define the opponent of an agent as the set of other agents within the system of interest. The opponent-relevant quantities and mappings are denoted with the subscript $-i$.

section. Under some circumstances, a history h can be summarized by a state variable as argued in Lanctot et al. (2019), and hence the history set $\mathcal{H}$ can be suppressed. However, we find it necessary to include agents' actions in addition to the state variable when facing complicated systems such as non-Markovian environments (Sutton, Precup, & Singh, 1999).

Based on the MASDM model in Definition 1, MAL can be viewed as an online decision-making process with incomplete domain knowledge about $\mathcal{G}$ and certain observability conditions. We provide the mathematical definition of MAL as follows.

Definition 2 (*Multi-agent Learning*)**.** A multi-agent learning process is defined by a sequence of tuples $\{\langle \mathcal{I}_i^t, \Gamma_i^t, \Pi_i^t \rangle_{t=0}^T\}_{i\in\mathcal{N}}$, involving the following elements.

1. $T \in \mathbb{N}_+ \cup \{\infty\}$ is the horizon of the learning process, and $t \in \{0, 1, \dots, T\}$ is the time index, where $t = 0$ indicates the pre-learning stage.
2. $\mathcal{I}_i^t$ is the information structure of agent i at time t, a set of variables whose realizations can be observed by the agent at time t. When $t = 0$, $\mathcal{I}_i^0 \subset \mathcal{G}$ corresponds to the subset of the domain knowledge available to the agent. When $t \geq 1$, $\mathcal{I}_i^t \subset \{S^t, \{M_j^t\}_{j\in\mathcal{N}}, \{A_j^{t-1}\}_{j\in\mathcal{N}}, \{R_j^{t-1}\}_{j\in\mathcal{N}}\}$, where S^t is the state variable, M_j^t is the message variable of agent j from its message space $\mathcal{M}_j$, and A_j^t, R_j^t corresponds to the action and reward of agent j, respectively.
3. $\Gamma_i^t : \mathcal{I}_i^{0:t} \to \Delta(\mathcal{A}_{-i})$ is the belief mapping, generating a belief $\gamma_i^t \in \Delta(\mathcal{A}_{-i})$ about the opponent's policy at time t. For completeness, it is assumed that Γ_i^0 maps from $\mathcal{I}_i^0$ to an arbitrary point in $\Delta(\mathcal{A}_{-i})$
4. $\Pi_i^t : \mathcal{I}_i^{0:t} \to \Delta(\mathcal{A}_i)$ is the policy mapping, generating a policy $\pi_i^t \in \Delta(\mathcal{A}_i)$ to be implemented at time t.

Note that in MASDM, there is no message element. This is because when solving the decision-making problem, the message can be viewed as a costless action (Myerson, 1991) or can be interpreted as an action recommendation in mechanism/information design problems according to revelation principle (Myerson, 1979). However, in the learning paradigm, messages have a broader usage (Foerster, Assael, de Freitas, & Whiteson, 2016a), and hence, we incorporate the message variable into the definition of MAL in Definition 2. A detailed discussion on the information structure is provided in Section 3, where the included variables are further elaborated on.

Our MAL framework comprises three components: the information structure, the belief mapping, and the policy mapping. These components determine what information the agent can acquire, how the information is processed, and what is the best policy given the acquired information. The proposed framework in Definition 2 provides a coherent view of various MAL problems. We provide three examples to show that the mainstream MAL models are in fact special cases of our MAL framework.

Learning in Markov games. Markov games (MG), also stochastic games, first proposed by Shapley (1953) in the 1950s, have long been used to model multi-agent strategic interactions in a dynamic environment. In the early ages, advances in Markov games were mainly contributed by game theorists and economists (Solan & Vieille, 2015). It is not until the seminal work by Littman (1994) was Markov games widely accepted as a framework of multi-agent reinforcement learning by the community.

In a Markov game, after observing the current state, all agents make their decisions simultaneously, and are rewarded by the environment accordingly. Then, agents move to the next state following the transition dynamics. The most notable feature of Markov games is that its transition dynamics is Markovian, meaning that the selection of the next state only depends on the current state and the joint actions implemented by agents. The formal definition is given below.

Definition 3 (*Markov Games*)**.** A Markov game is defined by a tuple

$$\langle \mathcal{S}, \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{T}, \{R_i\}_{i\in\mathcal{N}} \rangle,$$

where $\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}$ are all defined in the same way as in Definition 1. The transition dynamics and the reward functions follow the Markovian property, that is

1. the transition dynamics $\mathcal{T} : \mathcal{S} \times \prod_{i\in\mathcal{N}} \mathcal{A}_i \to \Delta(\mathcal{S})$ determines the probability from any state $s \in \mathcal{S}$ to any state $s' \in \mathcal{S}$ for any joint action $a \in \prod_{i\in\mathcal{N}} \mathcal{A}_i$, irrelevant of historical plays;
2. the reward is determined by the function $R_i : \mathcal{S} \times \prod_{i\in\mathcal{N}} \mathcal{A}_i \to \mathbb{R}$, which is also irrelevant of historical plays.

Research on learning in Markov games, generally called multi-agent reinforcement learning (MARL), revolves around multi-agent sequential decision-making with unknown dynamics and reward functions. Using the language in Definition 2, in most MARL studies, each agent is assumed to acquire $\mathcal{I}_i^0 = \{\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}\}$ as the domain knowledge. Through interactions with other agents and the environment, each agent may observe additional information about others' decision-making and the environment dynamics. For example, one common assumption in MARL is that each agent can observe other agents' actions, the realized rewards of the joint actions as well as the state transitions, and in this case $\mathcal{I}_i^t = \{S^t, \{A_j^t\}_{j\in\mathcal{N}}, R_i^t\}$. Based on this information, the agent reason about the opponent's behaviors and plan its moves accordingly. A detailed discussion regarding the observable information and the reasoning process will be included in the following sections.

As a simplification of MG, repeated games (RG) is a special case of Markov games, where there is only one state, and agents play the same game repeatedly. The same game played at each time is called a stage game or base game, denoted by a tuple $\langle \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{R_i\}_{i\in\mathcal{N}} \rangle$. With a simpler structure, repeated games are often used as a testbed for MAL algorithms. Recent advances in best response dynamics (Leslie, Perkins, & Xu, 2020) and gradient-based learning (Bu, Ratliff, & Mesbahi, 2019; Mazumdar, Ratliff, & Sastry, 2020) in Markov games are inspired by existing results in repeated games.

On the other hand, repeated games are an essential research topic in their own right, which models long-term strategic interactions among agents within a stable system or institution (Mertens et al., 2015). From learning in repeated games, there arise many exciting concepts such as reputation (Fudenberg & Levine, 1989) and trigger strategy (Mertens et al., 2015). Moreover, learning in repeated games provides a new interpretation of equilibrium concepts: a certain equilibrium of the base game is a stable outcome of multi-agent learning processes, which is resilient under slight disturbance. This interpretation is the driving force of the development in evolutionary game theory, and for more details, we refer the reader to Hofbauer and Sigmund (2003) and Li, Peng, Zhu, and Basar (2021).

Learning in extensive-form games. Even though they constitute a classical formalism for MARL, Markov games defined in Definition 3 can only handle the fully observable case, that is, the agent has perfect information on the system state S^t and the executed action A^t. Nonetheless, many MAL applications involve imperfect information, where, for example, agents can only observe actions implemented by their neighbors in a network setting.

In contrast, another framework for multi-agent decision making named extensive-form games (EFG), introduced by Kuhn (2016), can handily model these imperfect information cases. We briefly introduce the EFG framework in the following.

Definition 4. An extensive-form game is defined by

$$\langle \mathcal{S}, \mathcal{N} \cup \{c\}, \{\mathcal{A}_i\}_{i\in\mathcal{N}\cup c}, \mu_c, \mathcal{H}, \tau, \{R_i\}_{i\in\mathcal{N}} \rangle$$

and all components are defined the same as in Definition 1 except that

1. each element in $\mathcal{H}$ is a sequence of actions taken from the beginning of the game;
2. the state space $\mathcal{S}$ in fact specifies a partition such that for any $s \in \mathcal{S}$ and any $h, h' \in s$, we have $\tau(h) = \tau(h')$, meaning that histories h and h' in the same partition are indistinguishable to the agent ($\tau(h)$) that is about to take action.

Rooted in game-theoretic and economic studies, EFG provides a rigorous treatment of imperfect information in the sequential decision-making process, and there are many impressive results built on this particular framework (Brown & Sandholm, 2017, 2019). It should be noted that many advances in learning in EFG are following the computational agenda (Sandholm, 2007): learning algorithms serve as an optimizer for computing the equilibrium. Recent years have witnessed a series of thrilling successes in applying learning algorithms in EFG, such as regret minimization (Brown, Lerer, Gross, & Sandholm, 2019; Zinkevich, Johanson, Bowling, & Piccione, 2007), fictitious self-play (Heinrich, Lanctot, & Silver, 2015). Some of them have led to strong poker AI that defeated human players (Brown & Sandholm, 2017, 2019; Moravík et al., 2017).

Learning in multi-armed bandits. Multi-armed bandits (MAB) problem (Lattimore & Szepesvári, 2020) is probably the simplest RL problem: the agent is in a room with multiple gambling machines (called "one-armed bandits"). At each time step, the agent pulls the arm of one of the machines and receives a reward. The agent is allowed a fixed number of pulls, and the goal is to maximize its total reward over a sequence of trials. Each arm is assumed to have a different distribution of rewards. Therefore, the goal is to find the arm with the best-expected return as early as possible and then to keep gambling using that arm. Mathematically speaking, a K-arm stochastic MAB can be defined by $\langle \mathcal{N} \cup \{c\}, \mu_c, \{\mathcal{A}_i\}_{i\in\mathcal{N}\cup\{c\}}, \{R_i\}_{i\in\mathcal{N}}\rangle$, where $\mathcal{N} = \{1\}$. In a MAB problem, nature first determines the realized rewards $r = (r_k)_{k=1}^K$ of all arms, in which $r_k \in \mathbb{R}, 1 \le k \le K$ denotes the actual reward of pulling the kth arm drew from the corresponding distribution $\mathcal{D}_k$. Hence, nature's action set is $\mathcal{A}_c = \mathbb{R}^K$ and its fixed policy follows $\mu_c = \prod_{1\le k\le K} \mathcal{D}_k$. While for the agent, its action set is $A_1 := \{1, 2, \dots, K\}$ with each element representing an arm. At each time, after nature complete its move which is unknown to the agent, the agent chooses an action k and pull kth arm. The reward is $R_1(r, k) = r_k$.

Even though the MAB problem can hardly be seen as a MAL problem since there is only one decision-maker, it is useful to analyze the decision-making in such a stationary settings, which provides a simple yet fundamental framework for studying the exploitation–exploration trade-off in online learning. Some key ideas and methodologies developed in MAB research, such as upper confidence bounds (Auer, Cesa-Bianchi, & Fischer, 2002) and Thompson sampling (Kaufmann, Korda, & Munos, 2012) shed further light on similar studies in more complicated scenarios, such as reinforcement learning (Jin, Allen-Zhu, Bubeck, & Jordan, 2018).

Definition 2 provides a theoretical underpinning of MAL, allowing for a systematic investigation into various MAL approaches with respect to the key components: information structures, belief generation and policy generation processes. In the subsequent, detailed discussions regarding these components are presented.

3. Information structure

The importance of the information structure is self-evident: it is the input argument of both the belief mapping and the policy mapping. In this section, we carry out a detailed discussion on the information structure $\mathcal{I}^t$ and its variants in MAL problems.

We begin our discussion with $\mathcal{I}_i^0$. From Definition 2, $\mathcal{I}_i^0$ indicates how much the agent knows about the current multi-agent environment, which qualitatively depicts the uncertainty faced by the agent. Current MAL research focuses on building distributed AI which allows agents to adapt to the unknown environment and versatile tasks. Hence, it is common in the literature to assume that the transition dynamics and the reward function are unknown. For $\mathcal{I}_i^t, t \ge 1$, we discuss each element in $\{S^t, O^t, \{M_j^t\}_{j\in\mathcal{N}}, \{A_j^{t-1}\}_{j\in\mathcal{N}}, \{R_j^{t-1}\}_{j\in\mathcal{N}}\}$ in the following paragraphs.

State variables. The state S^t is a summary of the current system status, which is payoff-relevant in most cases. The agent needs to observe the realizations of state variables and adapt its behavior to the dynamic environment. Without access to S^t or state-related information, the agent cannot quantify the dynamics of the environment, let alone quantify the impact of other agents' move on the changing environment.

Since the state variable summarizes the system status, which includes other agents' status, for example, their current locations, full observation of the environment state may not always be available to everyone within the system at all times. In modern network applications with large and complex network topologies, it is neither computationally feasible nor desirable to distribute system information to every entity in the network. Meanwhile, in practice, due to noised sensing and communication processes, it is not possible to recover the accurate state information from the collected data, and only some estimates are available, which is quite common in control applications (Marden & Shamma, 2018). Therefore, it is more practical to have partial observability in a real-world application, although it can significantly complicate the analysis (Kaelbling, Littman, & Cassandra, 1998).

There are two kinds of modeling elements in the literature that account for this partial observability: public partial observations and private partial observations. Public partial observation is relatively simple. Every agent enjoys the same observation O^t of the underlying state S^t. While in the private partial observation, each agent may have different private partial observations O_i^t, which may incur signaling effect in the learning process (Ouyang et al., 2016). This means that agents' implemented actions may reveal some information regarding their private observations.

Action variables. The observability of action variables $\{A_j\}_{j\in\mathcal{N}}$ is also of vital importance in tackling non-stationarity issue. By observing other agents' implemented actions, the agent can reason or estimate the policy employed by others, which corresponds to the belief generation to be discussed in the next section. In other words, observing the state variable makes the agent aware of how the environment is changing, and observing the action variables makes it clear to the agent why the environment is changing.

It is natural to assume that each agent knows their own actions. However, the observability of other agents' actions $\{A_j\}_{j\in\mathcal{N}_{-i}}$ is subject to a case-by-case discussion. Similar to the observability issue of environment state in large and complex networks, it is not practical to assume that each agent can observe distant agents over networks. This spatial constraint also applies to the observability of reward variables and message variables, which will further be discussed in *Spatial Structures* in the subsequent.

Reward variables. The realizations of reward variables at each time serve as the evaluation of the implemented actions, according to which agents adjust their policies. When the reward function is not accessible to the agent, the reward realization r_i^t can be obtained by trial and error, based on which the agent can estimate the reward function. For a brief discussion on the estimation process, we refer readers to a recent survey (Li et al., 2021).

Despite the observability of its own reward variables, the observation of other agents' reward realization is also allowed under some circumstances. For example, in team problems, all agents share the same reward function. Meanwhile, due to the increasing popularity of the idea *centralized-learning-decentralized-execution* (Lowe et al., 2017), in recent works on deep MARL, it is pretty common, even in competitive settings, to assume that agents can access others' realized payoffs (Zhang et al., 2019). Unlike the case where the agent utilizes action

observations to estimate the opponent's policy, the agent can even take a step further with the observation of the opponent's reward: it can reason how the opponent generates the policy. The details will be discussed in Section 4.

Message variables. As we have discussed in Definition 2, due to revelation principle (Myerson, 1979), messages can be interpreted as action recommendations when solving for the multi-agent decision-making problem defined in Definition 1. Naturally, in MAL, a message can still serve as a special action or an action recommendation, as it does in offline planning. Yet, messages can also be agents' beliefs about their opponent' play (Eksin & Ribeiro, 2017; Swenson, Eksin, Kar, & Ribeiro, 2017), agents' estimated value functions (Kar, Moura, & Poor, 2013), or parameters of function approximators employed by agents (Zhang et al., 2018). In our later discussion, we will elaborate on the important role messages variables or communication in general plays when dealing with spatial structures of information.

Spatial structures. As we have already mentioned in the above discussions, for MAS with complex topologies, the observability of different variables is subject to a certain spatial constraint. For ease of exposition, we first explicitly describe the underlying topology using a graph. Although there is no graph-theoretic component in Definition 1 or Definition 2, we claim that our proposed model is still able to capture the topological structure. As argued in Jackson and Zenou (2015), most of the network topologies can be characterized by the structure of reward functions.

Consider a graph $\langle \mathcal{N}, \mathcal{E} \rangle$, where $\mathcal{N} = \{1, 2, \dots, N\}$ is the node set representing the agents in the system, and $\mathcal{E} = \{(i, j) | i, j \text{ are connected}\}$ is the edge set. Agents in the system are connected via the edges in $\mathcal{E}$. The edges may have many different interpretations for different applications. For example, in a multi-agent robotic network, edges can represent two-way communication channels through which agents can share information, resulting in an undirected graph. There are also problems requiring a directed graph if the information flow is directed. For simplicity, we assume that the graph is undirected, and our characterization of information structures still applies to directed ones.

For agents connected via the undirected graph, if they are able to observe their neighbors' actions and further allowed to exchange realized rewards and messages, then the information structure at time t can be defined as

$$\mathcal{I}_i^t = \{S^t, \{A_j^t\}_{j\in\mathcal{N}(i)}, \{R_j^t\}_{j\in\mathcal{N}(i)}, \{M_j^t\}_{j\in\mathcal{N}(i)}\},$$
$$\mathcal{N}(i) := \{j | (i, j) \in \mathcal{E}\} \cup \{i\}.$$

4. Belief and policy generation

As we have pointed out earlier, one of the challenges in MAL is the non-stationarity issue. Each agent faces a moving target learning problem because other agents' time-varying strategies have an impact on its own reward. The key to tackle the non-stationarity issue is to identify the opponent's play based on the acquired information, including domain knowledge and online observations. In this section, we elaborate on the other two components in Definition 2: the belief mapping Γ_i^t and the policy mapping Π_i^t.

4.1. Belief generation

The belief indicates the agent's understanding of the environment and the opponent, and it is the key to dealing with the non-stationarity in MAL. In the following, we characterize how an agent generates the belief about other agents' policies in the learning process. Then, four categories with increasing order of sophistication of the belief generation are proposed. For each category, we start with an illustrative example to concretely describe how the information helps produce a proper belief and a good policy. Following the specific examples, we comment on different approaches in the belief generation under various information structures and eventually provide an extensive list of algorithms categorized by how they deal with the non-stationarity issue.

For ease of exposition, we introduce these categories and discuss related algorithms from the perspective of a single agent, called the learner. The rest of the agents within the system are called the opponent as before. The notations of opponent-relevant quantities and mappings remain the same as in Definition 2.

4.1.1. Stationary opponent

In this category, the opponent is assumed to use a fixed mixed strategy, and the learner's goal is to identify the stationary strategy used by the opponent. Even though this assumption may not fit the ground truth, it simplifies the learner's belief generation process. For example, it suffices for the learner to compute the empirical frequency of the opponent's play only by observing the immediate actions.

One typical early work of this kind is fictitious play (Fudenberg et al., 1998), a simple learning algorithm used in repeated games for Nash equilibrium seeking. Consider a two-player repeated game $\langle \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{R_i\}_{i\in\mathcal{N}} \rangle$, where $\mathcal{N} = \{1, 2\}$. The information structure for agent i is as

$$\mathcal{I}_i^0 = \{\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, R_i\}, \quad \mathcal{I}_i^t = \{\{A_j^{t-1}\}_{j\in\mathcal{N}}\}.$$

Each player knows its own utility function and can observe the actions of the opponent. In fictitious play, from player 1's viewpoint, player 2 is following a fixed policy and its actions are independent and identically distributed samples drew from this fixed policy. Therefore, one simple way to estimate the policy employed by player 2 is to maintain an empirical frequency of the plays by the opponent in the past. Mathematically, player i's belief $\gamma_i^t \in \Delta(\mathcal{A}_{-i})$ about the other's policy at time t is given by

$$\gamma_i^t(a) = \Gamma_i^t(\mathcal{I}_i^{0:t}) = \frac{1}{t-1} \sum_{k=1}^{t-1} \mathbb{1}_{\{a_{-i}^k = a\}},$$

which is the empirical frequency of the opponent's actions up to time $t-1$. Using this belief, the learner chooses the best action that maximizes the expected payoff, and the learner's policy for the tth round is given by

$$\pi_i^t = \Pi_i^t(\mathcal{I}_i^{0:t}, \gamma_i^t) = \arg\max_{x\in\Delta(\mathcal{A}_i)} R_i(x, \gamma_i^t), \tag{1}$$

where $R_i(x, \gamma_i^t)$ is the expected utility under x, γ_i^t, defined as

$$R_i(x, \gamma_i^t) := \mathbb{E}_{a_i\sim x, a_{-i}\sim\gamma_i^t}[R_i(a_i, a_{-i})].$$

To sum up, the belief and the policy generation in fictitious play can be written in the following recursive form:

$$\gamma_i^{t+1} = \left(1 - \frac{1}{t}\right)\gamma_i^{t-1} + \frac{1}{t} e_{a_{-i}^t},$$
$$\pi_i^{t+1} = \arg\max_{x\in\Delta(\mathcal{A}_i)} R_i(x, \gamma_i^{t+1}), \tag{2}$$

where $e_a \in \Delta(\mathcal{A}_{-i})$ is the unit vector in the simplex, with its ath entry being 1 and 0 for the rest. It has been shown that (2) is indeed the discretized version of the best response dynamics (Hofbauer & Sigmund, 2003). When both players believe that their opponent is stationary and adopt (2), the collection of their beliefs about the opponent's play (γ_1^t, γ_2^t) converges to Nash equilibrium under certain conditions. For more details on the convergence analysis, we refer readers to Li et al. (2021) and Swenson, Murray, and Kar (2018).

Information structure. Since the opponent is assumed to be stationary, the information structure for this type of algorithm is relatively simple. The learning agent needs two kinds of information. One is opponent actions for estimating the opponent's policy. The other one is related to its own reward structure. In the example of fictitious play, it is assumed that the agent knows its own utility function for making decisions,

and this assumption about acquiring domain knowledge can be further lessened: it suffices for the learner to observe the realized payoffs, from which the learner can construct an estimate of the reward function. In this case, the information structure follows

$$\mathcal{I}_i^0 = \{\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}\}, \quad \mathcal{I}_i^t = \{\{A_j^{t-1}\}_{j\in\mathcal{N}}, R_i^{t-1}\}.$$

Examples of MAL algorithms under the above information structure include joint-action learners (Claus & Boutilier, 1998), individual Q-learning (Leslie & Collins, 2003, 2005) and other MAL works using best response dynamics. For more related works, we refer readers to a recent survey on learning in games (Li et al., 2021).

The above example articulates how a learner can adapt its policy to its belief about the opponent's play in a two-player repeated game under the stationary opponent assumption. However, when carrying the similar idea to network games where agents are connected via edges, it may not be possible for an agent to observe the actions of all others, especially in large and complex network systems. In this case, the spatial constraints in the information structure shall be considered, which means that agents can only observe the actions of their neighbors.

One possible approach to estimate the policies employed by those distant players is to resort to communication. As investigated in Eksin and Ribeiro (2017) and Swenson et al. (2017), each player can pass and receive messages from its neighbors to acquire information of the distant players' actions. The information structure in this situation is

$$\mathcal{I}_i^0 = \{\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, R_i\}, \quad \mathcal{I}_i^t = \{\{A_j^{t-1}\}_{j\in\mathcal{N}(i)}, \{M_j^{t-1}\}_{j\in\mathcal{N}(i)}\},$$

where M_j^{t-1} is agent j's beliefs about other agents' policies. The idea behind this communication-assisted belief generation is simple but effective. Players first construct beliefs about their neighbors' policies using the same way in (2) and then share their beliefs with their neighbors. By doing the process repeatedly, the learner can hold beliefs about everyone's policy without directly observing the opponent's actions. Then the learner simply performs the best response with this belief shown in (1). Based on the information passed by neighbors, all players best respond to the estimated strategies. The entire process can be viewed as a gossip-based fictitious play, which is proved to converge to Nash equilibrium in weakly cyclic games (Marden, Young, Arslan, & Shamma, 2009).

4.1.2. Conjectured opponent

In this class of MAL algorithms, the learner conjectures that the opponent follows a specific behavioral model to generate policies. It is noted that the exact strategy or the behavioral model of the opponent is not known to the leaner, and its conjecture of the opponent's play might be far from reality. Clearly, the previous "Stationary Model" is a special case of "Conjectured Opponent", where the behavioral model reduces to a simple fixed policy.

For this type of learning algorithms, we use temporal-difference learning as an example to illustrate how the learner construct the belief in the learning process. Known as the Bellman's heritage (Shoham, Powers, & Grenager, 2007), temporal-difference learning serves as the theoretical foundation for a plethora of MAL research works, including various extensions of Q-learning (Watkins & Dayan, 1992). Consider a two-player zero-sum Markov game

$$\langle S, \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{T}, \{R_i\}_{i\in\mathcal{N}}, \beta\rangle,$$

where $\mathcal{N} = \{1,2\}$ and $R_1(s,a_1,a_2) + R_2(s,a_1,a_2) = 0$, for all $s \in S$, $a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2$. Note that due to the zero-sum nature of the game, it suffices for the learner to observe its own reward R_i^t. When dealing with general-sum cases (Hu & Wellman, 2003), $\mathcal{I}_i^t$ shall also include other agents' rewards. Following the notation in Littman (1994), we define a new reward function $R : S\times\mathcal{A}_1\times\mathcal{A}_2 \to \mathbb{R}$ so that $R_1 = R, R_2 = -R$. With the new definition of the reward, the goal of agent 1 is to maximize the discounted cumulative reward $\mathbb{E}[\sum_{k=1}^{\infty} R(s^k, a_1^k, a_2^k)]$, while agent 2 tries to minimize it.

When the opponent (player 2) is assumed to perform a temporal-difference learning, its decision-making is based on the Q function, which is updated as

$$Q^t(s,a_1,a_2) = \begin{cases} Q^{t-1}(s,a_1,a_2), & \text{if } (s,a_1,a_2) \neq (s^t,a_1^t,a_2^t), \\ Q^{t-1}(s,a_1,a_2) + \alpha^t[R^t + \beta V^{t-1}(s^{t+1}) \\ \quad - Q^{t-1}(s,a_1,a_2)], \text{otherwise} \end{cases}, \tag{3}$$

where $V^{t-1}(s)$ is the maxmin value of $Q(s,\cdot,\cdot) \in \mathbb{R}^{|\mathcal{A}_1|\times|\mathcal{A}_2|}$, defined as

$$V^{t-1}(s) := \max_{x\in\Delta(\mathcal{A}_1)} \min_{y\in\Delta(\mathcal{A}_2)} x^{\top} Q^{t-1}(s,\cdot,\cdot) y,$$

and Q^{t-1} serves as the action evaluation for the play in the $t-$th round. The corresponding updating rule requires the following information structure

$$\mathcal{I}_i^0 = \{S, \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}\}, \quad \mathcal{I}_i^t = \{S^t, \{A_j^{t-1}\}_{j\in\mathcal{N}}, R_i^{t-1}\}.$$

Since the opponent is a minimizer and assumed to rely on temporal-difference learning in a fully competitive setting, we have

$$\gamma_1^t = \Gamma_1^t(\mathcal{I}_1^{0:t}) = \arg\min_{y\in\Delta(\mathcal{A}_2)} \max_{x\in\Delta(\mathcal{A}_1)} x^{\top} Q^{t-1}(s^t,\cdot,\cdot) y.$$

With this belief, the corresponding policy of player 1 at state s, is simply the best response

$$\pi_1^t = \Pi_i^t(\mathcal{I}_i^{0:t}) = \arg\max_{x\in\Delta(\mathcal{A}_1)} x^{\top} Q^{t-1}(s^t,\cdot,\cdot)\gamma_i^t. \tag{4}$$

Information structure. As we have mentioned, when dealing with general-sum cases, such as Nash Q-learning (Hu & Wellman, 2003) and Correlated Q-learning (Greenwald & Hall, 2003), $\mathcal{I}_i^t$ shall also include other agents' rewards:

$$\mathcal{I}_i^0 = \{S, \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}\}, \quad \mathcal{I}_i^t = \{S^t, \{A_j^{t-1}\}_{j\in\mathcal{N}}, \{R_j^{t-1}\}_{j\in\mathcal{N}}\},$$

which helps the learner to construct other agents' Q tables. Naturally, having access to other agents' reward realizations is not a trivial assumption. In practice, each agent may only acquire local and neighboring information, especially in network applications. Similar to the belief-sharing process discussed in the previous subsection, when communication is allowed in the learning process, agents are able to exchange information regarding their Q tables, in order to have a conjectured model for their opponent. This idea has been investigated in distributed Q-learning (Kar et al., 2013), where neighboring agents try to reach a consensus on their Q tables by communication, and the corresponding information structure is

$$\mathcal{I}_i^0 = \{S, \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}\}, \quad \mathcal{I}_i^t = \{S^t, \{A_j^{t-1}\}_{j\in\mathcal{N}}, \{M_j^t\}_{j\in\mathcal{N}(i)}, R_i^t\},$$

where the message M_j^t is agent j's Q table.

In addition to the value-based approaches above, the policy-based methods in MARL such as policy gradient (Bu et al., 2019; Mazumdar et al., 2020) and actor–critic (Foerster, Farquhar, Afouras, Nardelli, & Whiteson, 2018; Lowe et al., 2017; Zhang et al., 2018) also fall within this category, where the opponent is assumed to optimize its policy directly. For these algorithms, the opponent is assumed to use the gradient to update its policies. The computation of policy gradient calls for a global Q function (Zhang et al., 2018) or centralized critic (Lowe et al., 2017), which evaluates the quality of joint actions of all agents. One way to construct this global Q function of the centralized critic is to maintain a copy of other agents' Q functions. The corresponding information structure includes others' actions and rewards. Another approach is to make agents share their Q-tables or the parameters of the neural networks that approximate Q functions (Zhang et al., 2018). In the latter case, the information structure is the same as that in distributed Q-learning (Kar et al., 2013).

4.1.3. Calibrated opponents

Compared with algorithms within the category "conjectured opponents", algorithms from "calibrated opponents" move one step further: the learner can calibrate its conjecture in order to ensure that the conjectured model is consistent with the actual history of plays. In other words, algorithms from this category still follow a conjecture but with additional correction/calibration mechanisms. The learner's goal is to identify which behavioral model the opponent follows and detect the switch from one model to another as quickly as possible.

To show how the calibration mechanism works in a learning process, we discuss the MAL algorithm proposed in Conitzer and Sandholm (2007) for dealing with learning in repeated games: AWESOME (Adapt When Everybody is Stationary, Otherwise Move to Equilibrium). As its name suggests, when the opponent appears to be playing stationary strategies, AWESOME adapts to play the best response regarding these strategies. When the opponent appears to be adapting their strategies, AWESOME resorts to an equilibrium strategy. Hence, in the learning process, the learner needs to form a belief about whether the opponent is stationary or not, and constantly calls the calibration mechanism to adjust its belief.

The information structure of AWESOME is

$$\mathcal{I}_i^0 = \{\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}\}, \quad \mathcal{I}_i^t = \{\{A_j\}_{j\in\mathcal{N}}\},$$

where the learner has complete knowledge about the game, which enables it to compute the equilibrium strategy profile (π_i^*, π_{-i}^*). This complete domain knowledge is necessary for AWESOME, as the learner sticks to the equilibrium strategy π_i^* when the opponent is thought to be non-stationary. In the learning process, the agent needs a few interactions to learn and calibrate the opponent's model, and hence, the learning process consists of a series of epochs. The belief remains constant throughout the epoch, and will be adjusted at the beginning of the next epoch. For ease of the exposition, we denote the belief in the kth epoch by $\gamma_i^{t_k}$. Similarly, other notations introduced in Definition 2 with the t_k superscript denotes the corresponding quantities, mappings in the kth epoch.

In AWESOME, the belief is generated according to

$$\gamma_i^{t_k} = \begin{cases} \bar{\pi}_{-i}^{t_{k-1}}, & \text{if the opponent is thought to be stationary} \\ \pi_{-i}^*, & \text{otherwise} \end{cases},$$

where $\bar{\pi}_{-i}^{t_{k-1}}$ is the frequency of the opponent's play in the t_{k-1}-th epochs. In AWESOME, the calibration mechanism is a hypothesis test on whether the implemented actions of the opponent in the latest epoch (t_{k-1}-th epoch) are samples independently drew from the distribution $\bar{\pi}_{-i}^{t_{k-2}}$. If the opponent is stationary, there should not be much difference between $\bar{\pi}_{-i}^{t_{k-2}}$ and $\bar{\pi}_{-i}^{t_{k-1}}$, and AWESOME maintains the stationarity hypothesis if the ℓ_1 norm of the different is below some threshold.

Information structures. In AWESOME, the calibration mechanism is based on a statistical test, where the samples come from the repeated interactions. In a similar vein, Banerjee, Liu, and How (2017) and Hadoux, Beynier, and Weng (2014) propose RL algorithms under non-stationary environments, where the learner relies on hypothesis testing for detecting the change of the environment. These algorithms require that the learner has complete knowledge about the underlying MDP, and can fully observe the states, actions, and rewards in the learning process. Based on this information, the agent can compute the likelihood ratios for hypothesis testing.

Even though statistical tests provide a mathematically sound approach for detecting the change of the opponent's policy or behavioral model, it requires much of the domain knowledge to compute the likelihood ratio. When the learner has limited knowledge about the environment, it may also detect the change based on its own realized payoffs, making the learning process more self-dependent. The WoLF principle is based on the following heuristic: the learner should adapt fast when it is doing more poorly than expected. When it is doing better than expected, it should be cautious by diminishing the learning stepsize since the other players are likely to change their policy. Compared with statistical tests in algorithms like AWESOME, the calibration mechanism in WoLF-based algorithms only requires the agent to observe the realized payoffs in the repeated plays. The same intuition behind WoLF algorithms has also been explored in Marden et al. (2009) and Young (2009), and it has been shown in these works that under a properly designed calibration, agents' limiting behaviors arrive at equilibrium points.

4.1.4. Sophisticated opponent

Algorithms within the above categories above all assume that the opponent is following a particular behavioral model, which can be a stationary one ("stationary opponents"), a pre-defined one ("conjectured opponents"), or an unknown model that needs to be learned ("calibrated opponents"). The opponent in these models is assumed to be adaptive but not sophisticated, meaning that it adapts its behaviors according to a certain rule, and there is not strategic reasoning on the opponent's side.

In this last class, termed "sophisticated opponents", it is assumed that the opponent is also reasoning about others' decision making. Accordingly, MAL algorithms within this class model the opponent's behavior and model the opponent's strategic reasoning, which leads to nested reasoning: the learner would ponder how the opponent is reasoning about the learner's decision-making.

One illustrative example of algorithms within this category is the sophisticated experience-weighted attraction (s-EWA) (Camerer, Ho, & Chong, 2002), where two players repeatedly play the same normal-form game, with the following information structure

$$\mathcal{I}_i^0 = \{\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}\}, \quad \mathcal{I}_i^t = \{\{A_j^t\}_{j\in\mathcal{N}}\}.$$

With a slight abuse of notations, in s-EWA, the learner's belief of the opponent's policy is given by

$$\gamma_i^t = \Gamma_i^t(\mathcal{I}_i^{0:t}) = \Gamma_i^t(\gamma_{-i}^t), \tag{5}$$

where γ_{-i}^t denotes the opponent's belief of the learner's policy at time t. Similarly, we have

$$\gamma_{-i}^t = \Gamma_{-i}^t(\mathcal{I}_{-i}^{0:t}) = \Gamma_{-i}^t(\gamma_i). \tag{6}$$

Since agents share the same domain knowledge and online observations, $\mathcal{I}_i^{0:t} = \mathcal{I}_{-i}^{0:t}$, combining (5) and (6) leads to the following fixed-point characterization of γ_i^t

$$\gamma_i^t = \Gamma_i^t(\gamma_{-i}^t) = \Gamma_i^t \circ \Gamma_{-i}^t(\gamma_i^t), \tag{7}$$

based on which, the learner takes the best response

$$\pi_i^t = \Pi_i^t(\mathcal{I}_i^{0:t}, \gamma_i^t) = \arg\max_{x\in\Delta(\mathcal{A}_i)} R_i(x, \gamma_i^t).$$

Since s-EWA assumes that sophisticated agents believe others are sophisticated, and those others think others are sophisticated, so on so forth, it creates a whirlpool of recursive thinking as demonstrated in (7), which leads to equilibrium concepts. It has been shown in Camerer et al. (2002) that when agents are all sophisticated and believe others are sophisticated, the learning outcome is a Nash equilibrium.

Information structures. Different from previously discussed information structures, in s-EWA, the domain knowledge includes the reward functions, and the online observations are joint actions of all agents at each time, which are common information. This common information provides agents with the same jump-off point when carrying out the iterative reasoning process, simplifying the theoretical analysis. When agents are allowed to observe private information, the signaling effect of their actions must be taken into account, and the resulting belief hierarchy is quite challenging when developing the algorithm (Ouyang et al., 2016).

Since research works on MAL within this category are relatively scarce, we comment on some models and related information structures

considered in planning problems. The level-k and cognitive hierarchy models (Camerer, Ho, & Chong, 2004; Costa-Gomes, Crawford, & Broseta, 2001) are mainly applied to analyze the iterative reasoning process. The model involves an initial set of zero-level strategies, usually uniform distributions over the action spaces, representing non-strategic behaviors. The next-level strategy is essentially the best response against the current level. Using the language of the belief generation in Definition 2, the learner's belief is produced by (5), where the opponent's belief about the learner γ_{-i}^t depends on which level the opponent believes the learner is in.

Inspired by the cognitive hierarchy, Gmytrasiewicz and Doshi (2005) propose a formal sequential decision-making model called interactive POMDP (I-POMDP), which considers what an agent knows and believes about what other agents know and believe. In this model, the state variable incorporates models of how agents reason, which is unobservable to all. The agent's belief of the true state tells how the agent believes another agent reasons. Parameterized I-POMDP (Wunder, Kaisers, Yaros, & Littman, 2011) is more closely related to level-k theory (Costa-Gomes et al., 2001). The idea is to compute a policy that maximizes the rewards against the distribution of agents over previous levels or selects representative agents from these levels, by solving the POMDP formed by them.

4.2. Policy generation

In our previous discussion, we have mentioned how the learner generates its policy π_i^t based on the belief γ_i^t. In this subsection, a more detailed treatment on policy generation is provided. We also summarize policy generation approaches primarily used in the literature.

Best response. The most direct way of generating policies is to best respond to γ_i^t which is the opponent's belief generated by the learner. Mathematically speaking, the best response policy is given by

$$\pi_i^t = \arg\max_{x\in\Delta(\mathcal{A}_i)} R_i(x, \gamma_i^t), \tag{BR}$$

where R_i denotes the reward function. R_i can also be replaced by its estimate, such as the Q function. MAL algorithms, such as fictitious play (Brown, 1951) and its variants (Eksin & Ribeiro, 2017; Swenson et al., 2017), utilize (BR) for producing π_i^t. When other players are believed to play the equilibrium strategies, the best response to these strategies also leads to an equilibrium strategy for the learner. In the sense, value-based MARL algorithms, such as minmax Q-learning (Littman, 1994), Nash Q-learning (Hu & Wellman, 2003) and Correlated Q-learning (Greenwald & Hall, 2003), also relies on the best response idea.

Smoothed best response. Since the best response mapping in (BR) always seeks the maximum, the resulting policy may be myopic and exploitable in Li et al. (2021). In order to balance the exploitation and exploration, a regularization term can be added in (BR) so that the probability of choosing suboptimal actions is greater than zero. This regularized best response is referred to as the smoothed best response, and mathematically, it is defined as

$$\pi_i^t = \arg\max_{x\in\Delta(\mathcal{A}_i)} R_i(x, \gamma_i^t) + \epsilon h(x), \tag{SBR}$$

where $h(\cdot)$ is the regularizer with strong convexity and ϵ is called the exploration parameter, determining how likely the suboptimal actions will be chosen. When $h(x)$ is the entropy function, the resulting policy is called softmax policy (Neu, Jonsson, & Gómez, 2017) or Boltzmann–Gibbs policy (Zhu et al., 2010). We refer the reader to Li et al. (2021) for more details on the selection of the regularizer and the exploitation–exploration trade-off.

Gradient response. Both the best response (BR) and the smoothed best response work for learning with finite actions. When dealing with infinite spaces, for example, the policy space $\Delta(\mathcal{A}_i)$, the agent may rely on the gradient information to search better policies.

$$\pi_i^t = \pi_i^{t-1} + \eta_i \frac{\partial R_i(\pi_i^{t-1}, \gamma_i^{t-1})}{\partial \pi_i}, \tag{GR}$$

where η_i is the learning rate for the learner. When the gradient $\frac{\partial R_i(\pi_i^{t-1}, \gamma_i^{t-1})}{\partial \pi_i}$ can be estimated from collected samples, the gradient response does not directly rely on the reward function, leading to a policy-based method in MAL. The gradient response (GR) has been widely applied in learning in repeated games, such as WoLF-based algorithms (Bowling, 2004; Bowling & Veloso, 2002), as well as learning in Markov games, such as policy gradients (Silver et al., 2014) and actor–critic methods (Foerster et al., 2018; Lowe et al., 2017).

5. Discussions

Following the four categorizations of MAL algorithms, we discuss their strengths and limitations regarding the theoretical analysis and practical implementations. A short summary of our discussion is presented in Table 2.

5.1. Strengths and limitations

The discussion in this subsection is not meant to be comprehensive since new developments and interpretations are being brought up in this burgeoning research field. Instead, we focus on the following two aspects when discussing the strengths and limitations of algorithms from different categories. We first comment on the information structure of MAL algorithms from the introduced categories, which determines the applicability of these algorithms. Then, we discuss theoretical guarantees that can be obtained for algorithms within these categories.

Stationary Opponent. Because of the simple assumption about the opponent, algorithms within this category generally adopt simple information structures, and most of them do not require extra information in addition to action observations from the opponent. Besides, under this assumption, the opponent can be viewed as a part of the learning environment, essentially stationary, from the learner's perspective. Under this assumption, single-agent reinforcement learning methods can be easily extended to MARL (Hernandez-Leal et al., 2017).

Convergence to equilibrium or the best response policy has been proven in particular scenarios. For example, when agents all use the same fictitious play in repeated games or Markov games with certain payoff structures (Leslie et al., 2020; Sayin, Parise, & Ozdaglar, 2020), the resulting beliefs converge to a Nash equilibrium. However, in general, theoretical guarantees do not hold when the stationary opponent assumption fails, which has been reported in the literature as a motivating example of the non-stationarity issue (Littman, 1994). The use of algorithms within this class can be considered when no extra information can be obtained from the environment.

Conjecture Opponent. As a slightly more advanced model than "Stationary Opponent", algorithms within the category "Conjectured Opponent" may require more than just action observations, depending on the specific model the learner applies. Similar to our argument in "Stationary Opponent", when the opponent's behavior pattern is appropriately modeled, theoretical results, especially convergence analysis, is no longer a daunting task, given fruitful tools such as stochastic approximation (Benaïm, Hofbauer, & Sorin, 2005) and online convex/ linear optimization (Shalev-Shwartz, 2011). The majority of MAL research works focus on this model.

Thanks to the restrictive assumption of the opponent, algorithms within the first two classes under certain regularity conditions converge to stationary policies when dealing with MAL in RG or MG (Zhang et al.,

2019). These obtained stationary policies are best-response policies for the opponent's model, which is the best the learner can hope in the face of other adaptive agents in the multi-agent system. Since adopting different opponent models leads to different interpretations of the best response policy, algorithms within this "Conjectured Opponent" can be leveraged to search policies with different properties. For example, when the opponent is assumed to be the worst-case as assumed in minimax Q-learning (Littman, 1994), the obtained policy can be regarded as a robust solution to the MAL problem. The idea of robustness has been explored in many engineering applications (Marden & Shamma, 2018), especially in security domains (Zhu & Başar, 2013, 2015).

One limitation of this class is the constrained adaptability of these algorithms. The successful implementation of algorithms within this class requires that the learner is aware of the decision-making scheme employed by the opponent. Even though the knowledge of others' decision-making schemes is not mandatory when rolling out the algorithm, when the conjecture is wrong, the failure of achieving high rewards or other related criteria in the learning process is not surprising. Considering its strengths and limitations, algorithms within this class can be utilized when the learner has access to the opponent's learning scheme or some information regarding its behavioral pattern. More importantly, the opponent's decision-making can be designed so that the resulting best response policy enjoys desired properties, and then algorithms can be considered to search the desired policy.

Calibrated Opponent. The model "Calibrated Opponent" is more advanced than the previous models, and algorithms within this class do not require a strong prior assumption about the opponent's learning process. Instead, the learner is directed to construct a model of the opponent's decision-making during the learning, which brings algorithms to a broader audience. It should be noted that due to the calibration mechanism, the constructed model may be time-varying, depending on the learner's observations of the opponent as well as the opponent's inherent decision-making.

On the one hand, the calibration mechanism makes the learner less exploitable (Bowling, 2004; Conitzer & Sandholm, 2007), as it constantly corrects its constructed model based on the online observations. As we have illustrated in the previous section, the idea of quickest change detection enables the learner to adjust the constructed model swiftly once there is a misalignment between what the opponent is projected to do and what it really did. The adjustment of the opponent model further leads to the adaption of learner's policies, which is designed explicitly for the adjusted opponent model. In other words, the calibration mechanism enables the learner to do the right thing at the right time. Another advantage of these algorithms is that the learned model of the opponent can be reused if the opponent returns to the same strategy. The idea of reusing learned models or knowledge is closely related to transfer learning (Zhang & Bareinboim, 2017) or causal reinforcement learning (Bannon et al., 2020; Buesing et al., 2018), where information from previous interactions can be reused in order to reduce the sample complexity. We will include more details on this topic when discussing future directions.

Despite the strength, the limitations of these algorithms are also due to the calibration mechanism. First, the learner requires sufficient rounds of interactions to learn and construct the opponent's model, and the associated sample complexity can be prohibitive. Second, with the calibration mechanism, if the learner constantly calibrates the opponent's model, equilibrium convergence analysis can be quite challenging. In general, for algorithms within this class, the theoretical analysis is more involved, and for the most of existing works, only performance guarantees are available (Marden & Shamma, 2018). In the literature, the most used notion is regret, which is the gap between the average performance under current policies and the best policy in hindsight. When the regret is diminishing or upper bounded, these algorithms achieve no-regret or low regret. The idea of regret can be further extended under different circumstances. Weighing its strengths and limitations, we suggest that algorithms from this category can be applied to learn the opponent's model when the learner possesses limited domain knowledge.

Sophisticated Opponent. Distinct from all other classes, the class "Sophisticated Opponent" is more related to behavioral game theory (Camerer, 2011; Wright & Leyton-Brown, 2010), where players perform complex strategic reasoning. Compared with other classes, research on this type of learning is still in its infancy, and algorithms within this class mostly fall within the realm of behavioral economics and psychological studies. One strength of these algorithms is that they are suitable for predicting agents' transient behavioral or strategic moves in a short period. It is because agents determine their policy by reasoning how others may react instead of modeling and estimating the opponent, and hence, agents requires less rounds of interactions. However, the process of strategic reasoning necessitates high computational costs to solve them (Camerer et al., 2004). Even though in simple examples (Camerer et al., 2002), theoretically analyzing the learning outcomes is viable, producing interesting interpretations of concepts in behavioral game theory (Camerer, 2011; Wright & Leyton-Brown, 2010), analytical results are scarce in this field of study.

5.2. The value of information

According to the definition of the information structure in Definition 2, $\mathcal{I}_i^0$ and $\mathcal{I}_i^t$ are subsets of the complete domain knowledge $\mathcal{G}$ and the set of variables $\{S^t, \{M_j^t\}_{j\in\mathcal{N}}, \{A_j^{t-1}\}_{j\in\mathcal{N}}, \{R_j^{t-1}\}_{j\in\mathcal{N}}\}$, respectively. Therefore, we can use the notion of inclusion relation in set theory to compare information structures of different agents.

Definition 5 (*Information Superiority, Inferiority and Equality*)**.** In a MAL process with a horizon T, one agent i is said to be informationally superior to another agent j, if the information structure of agent i at time t is a proper superset of that of agent j, for $t \in \{0, 1, \dots, T\}$, i.e.,

$$\mathcal{I}_j^t \subsetneq \mathcal{I}_i^t, \quad \text{for all } t \in \{0, 1, \dots, T\}.$$

In this case, agent j is said to be informationally inferior to agent i. Furthermore, agent i is said to informationally equal to agent j, if the information structures of the two agents coincide.

It is natural to conjecture that information superiority leads to more accurate beliefs about the opponent, and hence, results in higher rewards in the learning process. For example, in MARL, having access to everyone's actions and realized payoffs enables the learner to construct Q tables of other agents, and further to learn its equilibrium policy (Greenwald & Hall, 2003; Hu & Wellman, 2003). By contrast, the numerical results in Littman (1994) demonstrate that unobservability of the opponent's realized payoffs renders Q-learning ineffective when facing multiple agents.

However, by the following example, we argue that this conjecture does not hold for every situation. It is likely that acquiring more information may lead to worse outcomes, which we termed as *Information Paradox*.

Example 1 (*Information Paradox*)**.** We consider a repeated zero-sum game between two players, who have the same action space $\mathcal{A}_1 = \mathcal{A}_2 = [-1, 1]$. The reward functions are defined as $R_1(a_1, a_2) = -R_2(a_1, a_2) = -a_1 \cdot a_2$. The player 1 is regarded as the learner who adopts fictitious play, whereas the player 2 is unintelligent, and use the following policy

$$\pi_2^t = \begin{cases} -0.5, & t = 1, \\ 1, & t \text{ is even}, \\ -1, & t \text{ is odd and greater than 1}. \end{cases}$$

In the repeated play, if the learner utilizes fictitious play with $a_1^1 = 0$ as the initialization, then the learner's action a_i^t and the immediate reward $R_1(a_1^t, a_2^t)$, as well as the cumulative reward $\sum_{k=1}^t R_1(a_1^k, a_2^k)$ can be summarized in Table 1. As shown in Table 1, under the information structure $\mathcal{I}_1^0 = \{\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}\}$, $\mathcal{I}_i^t = \{\{A_j^t\}_{j\in\mathcal{N}}\}, t \geq 1$, the cumulative rewards for the learner tends to $-\infty$, implying that the learner has been exploited by the unintelligent opponent.

Table 1
A summary of the repeated play. Even facing an unintelligent opponent, the learner fails to form a correct belief about its opponent and keeps being exploited by the opponent.

Round	Action	Immediate reward	Cumulative reward
$t=1$	$a_1^1=0$	0	0
$t=2$	$a_1^2=1$	-1	-1
$t=3$	$a_1^3=-1$	-1	-2
$t=4$	$a_1^4=1$	-1	-3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t=2k-1$	$a_1^{2k-1}=-1$	-1	$-2k+2$
$t=2k$	$a_1^{2k}=1$	-1	$-2k+1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

On the contrary, let $\mathcal{I}_i^t$ be the empty set for all $t \geq 1$, then the learner is in fact better off under this new information structure, denoted by $\tilde{\mathcal{I}}_i^t$. In this case, the learner is only aware of the game and cannot observe anything during the learning. It may use its security strategy (maxmin strategy or worst-case strategy) (Zamir, 2009), which leads to a bounded cumulative reward.

The information paradox example provides a counterexample to the aforementioned conjecture that information superiority leads to better outcomes in learning. The counterexample necessitates the introduction of a metric, termed the value of information (VoI), which quantitatively evaluates the impact of the information structures on MAL algorithms. To formally define the value of information, we introduce the following notations. Let $\mathcal{OM}$ be a given opponent model selected from the four categories: "stationary opponents", "conjectured opponents", "calibrated opponents" and "sophisticated opponents". Denote $\bar{R}_{\mathcal{OM}}(\mathcal{I}_i^{0:T})$ the time-averaged expected reward of agent i under the information structure $\mathcal{I}_i^{0:T}$ using the model $\mathcal{OM}$. Accordingly, we denote $\bar{R}(\mathcal{G})$ the time-averaged expected reward of the optimal solution using planning methods, e.g., linear programming for solving Markov games (Filar & Vrieze, 2012). $\mathcal{G}$ is the MASDM problem defined in Definition 1.

Definition 6 (*The Value of Information*)**.** For a given MASDM problem $\mathcal{G}$, when agent i applies the model $\mathcal{OM}$, the value of the information structure $\mathcal{I}_i^{0:T}$ is defined as

$$\mathrm{VoI}_{\mathcal{OM}}(\mathcal{I}_i^{0:T}) := \frac{\bar{R}_{\mathcal{OM}}(\mathcal{I}_i^{0:T})}{\bar{R}(\mathcal{G})}.$$

When analyzing the difference brought up by different information structures in the same MAL problem, we can consider computing the ratio of one information structure over another. In the information paradox example, the optimal policy for the learner is

$$\pi_1^t = \begin{cases} -1, & t \text{ is even}, \\ 1, & t \text{ is odd}. \end{cases}$$

The time-average reward under this policy is $\bar{R}(\mathcal{G}) = 1$, as T goes to infinity. Under the original information structure $\mathcal{I}_i^{0:T}$, we have $\bar{R}_{\mathcal{OM}}(\mathcal{I}_i^{0:T}) = -T/2$. While the average reward under $\tilde{\mathcal{I}}_i^{0:T}$ is in fact a random variable, which depends on the fixed policy adopted by the learner. For simplicity, we assume the policy is the uniform distribution over $[-1, 1]$, then we obtain the expected reward is $\bar{R}_{\mathcal{OM}}(\tilde{\mathcal{I}}_i^{0:T}) = 0$. Therefore, $\mathrm{VoI}_{\mathcal{OM}}(\mathcal{I}_i^{0:T}) = -T/2$, $\mathrm{VoI}_{\mathcal{OM}}(\tilde{\mathcal{I}}_i^{0:T}) = 0$, and

$$\frac{\mathrm{VoI}_{\mathcal{OM}}(\mathcal{I}_i^{0:T})}{\mathrm{VoI}_{\mathcal{OM}}(\tilde{\mathcal{I}}_i^{0:T})} = -\infty,$$

implying that the original information structure deteriorates the learning, even though it enjoys information superiority.

Regarding the non-stationarity issue in MAL, we point out that the learning process may not be convergent, nevertheless, $\bar{R}_{\mathcal{OM}}(\mathcal{I}_i^{0:T})$ can be replaced by the upper or lower bounds of the averaged rewards. Our argument above still applies to non-convergent learning processes.

Finally, it should be noted that in Definition 6, the performance of the optimal solution $\bar{\mathcal{G}}$ is chosen as the baseline, which bridge the gap between the studies of learning and planning. Direct comparisons can be made between learning methods and planning methods. Even though many MAL works focus on convergence to certain equilibrium point, it is reasonable to expect that learner may even do better than simply employing the equilibrium strategy. For example, when it is possible for it to exploit other irrational agents, the learner may achieve more than the equilibrium payoffs.

5.3. Future directions

As we have mentioned at the beginning of this paper, as an active research area, MAL is still in its infancy, and there are many open questions. In this section, we present several promising lines of research.

5.3.1. Design of information structures

The *Information Paradox* in Example 1 indicates that the relationship between information superiority and agents' learning performance is not monotonic: more information does not necessarily leads to better outcome. The root cause of this non-monotonicity is the mismatch between the information structure and the belief and policy generation adopted by the learner. By wrongly assuming that the opponent is stationary, the learner makes poor predictions about the opponent's play, and then best responds to wrong beliefs, which resulting in itself being exploited by the unintelligent opponent.

Considering the three key components of MAL, i.e., information structures, belief and policy generation, there are three kinds of remedies to free the learner from being exploited. The first one is to find an information structure that fits the belief and policy generation. As shown in Example 1, the learner gets better off by not observing anything, if its belief generation follows "stationary opponent" model. The second approach is to adjust the learner's belief generation that allows for a consistent conjecture on opponent's non-stationary behavior. For example, if the learner adopts AWESOME learning (Conitzer & Sandholm, 2007) discussed in Section 4.1.3, it can detect the non-stationarity of opponent's strategy within one epoch, and hence, resort to equilibrium strategy, i.e., maxmin strategy, achieving a lower bound for possible loss. Finally, for the policy generation, a simple switch from best response to smoothed best response (see Section 4.2) leads to a policy generation called follow-the-regularized-leader, which is extensively studied in MAB problems (see Section 2), and is shown to achieve no-regret performance asymptotically (Shalev-Shwartz, 2011).

From the above discussion, we can see that the success of a MAL algorithm rests on the proper alignment of information structures, belief generation and policy generation. The existing literature has mostly focuses on designing belief and policy generation under a given information structure, such as learning under partial observations (Hernandez-Leal et al., 2019; Kaelbling et al., 1998) and coordinated learning over networks (Li et al., 2021; Liu et al., 2020; Marden & Shamma, 2018; Zhu, Tembine, & Başar, 2011).

In contrast, the design of information structures has remained largely an uncharted territory. One goal of the design is to ensure the compatibility of the information structure with respect to the agent's belief and policy generation. With a proper design of information structures, each agent is enabled to make the most of its observations, and learns to adapt to others' non-stationary behaviors efficiently. In addition to efficient learning with reasonable sample and computation complexity, the goal of the design also includes equipping the agents with informational adaptability and resiliency, when the MAS is deployed in a dynamic, uncertain and adversarial environment. Specifically, following the same spirit of meta learning (Finn, Abbeel, & Levine, 2017), by exposing agents to a family of properly designed information structures, one can endow these learning agents with informational adaptability. In this case, agents are not subject to any

prescribed information structure, instead, they can quickly adapt to a collection of closely related information structures in an online manner through the interactions with other agents and the environment. This informational adaptability makes the MAL resilient-by-design: when part of the agents are compromised by the adversary, the rest can adapt to the new information structure, and reconfigure their learning based on the type of agents that they interact with. In this sense, a proper design of information structures can increase the composability and modularity of MAL, leading to a mosaic operation of MAS (Chen & Zhu, 2019b).

5.3.2. Heterogeneity in MAL

In our presentation of MAL algorithms, to simplify our argument, we mainly address two-player cases. However, from the definition of MAL in Definition 2, a MAL algorithm is said to be comprised of three components: the information structure $\mathcal{I}_i^t$, the belief generation Γ_i^t and the policy generation Π_i^t, all of which are player-dependent. In other words, it is possible to encounter several agents with different learning characteristics, which include different information structures as well as various belief and policy generation processes. The fact that agents within the same multi-agent system may possess distinct learning capabilities are referred to as the heterogeneity of the system.

The heterogeneity is one of the complicating factors in developing and analyzing MAL algorithms. This heterogeneity is still theoretically manageable for algorithms within "calibrated opponent" and "sophisticate opponent", as these algorithms do not require too much information regarding the opponent's decision-making. However, it should be noted that the size of the sample and computation complexity due to the heterogeneity may render algorithms practically infeasible when it comes to the implementation of these algorithms. This complexity issue will be further discussed in Section 5.3.3.

On the other hand, the heterogeneity issue brings up great challenges when using algorithms within the "Stationary Opponent" and the "Conjecture Opponent". First, these algorithms heavily rely on domain knowledge, for example, the opponent's behavioral patterns. However, in a heterogeneous system, it is not reasonable to assume all agents are stationary opponents or follow the same conjectured model, limiting the application of these algorithms. Second, even if the assumption holds, it is much more complicated to study the learner's limiting behavior and the stabilized system outcome. This is because the resulting dynamical systems of the learning schemes are much more involved, and classical analyzing techniques, such as certain Lyapunov functions in game-theoretic learning are not readily available. For example, the convergence of fictitious play in network systems often requires the introduction of an additional inertia term (Swenson et al., 2017) and special game structures (Eksin & Ribeiro, 2017). For more detailed discussions on heterogeneous learning, we refer to the reader to Zhu, Tembine, and Basar (2013).

One way to cope with such a complex learning environment is to characterize them across different dimensions. For example, agents having the same reward structure can be labeled as one type, or agents with the same information structure can be viewed as a team (Sunehag et al., 2018; Tang et al., 2021). It may also be possible that agents can be classified by the learning algorithms they use (Tardos et al., 2018). Once agents are labeled according to some rules, each group of agents can be viewed of the same type, where types are distributed according to a prior distribution. Then, each type is treated as a new decision-maker, and our proposed MAL framework still applies, which leads to population-based MAL algorithms (Bard, Nicholas, Szepesvári, & Bowling, 2015; Tembine et al., 2014). Besides this population approach, a multi-scale or multi-resolution approach is also worth exploring. The multi-scale idea is intuitive: all agents are organized into different groups, and the MAL problem is first solved at the group level, which serves as an initialization for the search of the individual-level solution (Bouvrie & Maggioni, 2012; Li & Zhu, 2019).

5.3.3. The challenge of scalability

To handle non-stationarity, following our characterization of MAL algorithms, each agent needs to consider the decision-making of others. Depending on the specific model the agent utilizes, it may require information regarding joint actions, rewards, and messages of all agents, whose dimension increases exponentially with the number of agents. This is also referred to as the combinatorial nature of MAL (Zhang et al., 2019), another issue in MAL studies, and can become even more challenging when considering heterogeneity. The real-world applications of multi-agent systems, such as infrastructure networks (Chen & Zhu, 2019a), usually include complex underlying topologies and heterogeneous agents, which require a blend of multiple MAL learning schemes, resulting in prohibitive sample complexity and computation complexity. The scalability issue used to and continues to be one of the primary factors that prevent MAL from being massively deployed in reality.

As a long-standing challenge in MAL and even in broader artificial intelligence research, the scalability issue receives much attention from the community, and there are many possible remedies. The most straightforward one is to resort to function approximation (Geramifard et al., 2013), especially deep neural networks (LeCun et al., 2015), which has achieved many successes in past decades (Brown & Sandholm, 2017; Lanctot et al., 2019; Mnih et al., 2015; Moravík et al., 2017; Silver et al., 2014). However, the theoretical analysis of MAL with deep neural networks is almost uncharted territory due to the limited understanding of deep learning theory. In addition, the interdependence between deep learning and decision-making further complicates the matter: poor approximation leads to poor decisions, which may further affect the representation learning. It remains unclear how the representation learning is connected to agents' decision-making, and the mutual influence between function approximation and decision-making is quite difficult to quantify.

Another approach is to leverage the inherent structure of the environment. For example, the environment can be simplified if there exist factorized structures of the reward functions and/or transition dynamics with respect to the action/state dependence (Guestrin, Koller, & Parr, 2002). The resulting problem is simpler than the original one, which helps reduce the computation complexity. We refer the reader to Guestrin, Koller, and Parr (2001) and Kok and Vlassis (2004) for the original heuristic ideas and (Rashid et al., 2018; Sunehag et al., 2018) for recent progress. However, this factorized structure does not solve the root of the scalability problem.

The third approach is the idea of information reuse and knowledge transfer we have mentioned in the previous subsection. The specific examples include batch reinforcement learning (Bannon et al., 2020) and causal reinforcement learning (Bannon et al., 2020; Buesing et al., 2018; Zhang & Bareinboim, 2017), where past experiences are reused to model the current situations. By leveraging the learned knowledge, agents do not need too many observations/samples to rebuild the model from scratch. We refer the reader to Bannon et al. (2020) for a review on these topics.

5.3.4. Novel learning objectives

Different from single-agent learning, where the goal is to maximize the long-term rewards, the learning objectives of MAL can be vague. This unclarity of what to be learned in MAL is the fundamental question in many early MAL works, as argued in Shoham et al. (2007).

The most common goal in MAL is still related to rewards maximization. The idea is to achieve optimality for multiple agents, which is described by some equilibrium. For example, if the algorithm finally converges to Nash equilibrium, no agent will deviate from the learned policy, leading to a stable point of the learning dynamics. This is undoubtedly a proper solution concept in game theory, assuming that the agents are all perfectly rational. In addition to the equilibrium concept, the notion of regret captures agents' rationality from another angle.

As we have mentioned before, regret measures the algorithm's performance compared with the best hindsight static strategy. Compared with the equilibrium concept, regret is a less restrictive criterion: convergence to the equilibrium indicates that the MAL algorithm of question achieves the optimality with other agents' presence. In contrast, convergence to a zero regret implies that the agent is not exploited by others (Lattimore & Szepesvári, 2020). Naturally, in certain games, non-exploitation leads to the equilibria (Xu & Zhao, 2020). In this case, algorithms asymptotically achieving zero average regret guarantees the convergence to the equilibria.

However, the two notions introduced above are related to the long-term behaviors of the agents within the multi-agent system. From the system viewpoint, MAL algorithms that aim at convergence to the equilibrium or the no-regret policy, are more concerned with the stabilized system-level performance than with the ongoing process of learning. However, as argued in Hernandez-Leal et al. (2019), the transient behaviors of learning agents also matters, especially in some safety-critical MAL applications, such as autonomous driving (Shalev-Shwartz, Shammah, & Shashua, 2016) and security-critical cyber–physical systems (Zhu & Başar, 2015). In these applications, the movements of the learning agents must be constrained due to the safety requirement. Hence, in these applications, the performance guarantee of transient behaviors is equally important as that of limiting behaviors. Another motivation behind this transient performance guarantee is that in some MAL problems in the security domain (Zhu & Başar, 2015), the interactions are often of limited horizons, which renders the convergence analysis infeasible. To sum up, in addition to current long-term performance criteria, there is a call for new learning objectives that account for the desired transient behaviors in MAL. Even though previous works consider combining constrained dynamic programming (Altman, 1999; Borkar, 2005) with existing MAL frameworks, the area still remains open for future investigation.

In addition to the goals concerning optimizing the return, several other goals have also attracted increasing attention from MAL researchers and practitioners. For example, Foerster, Assael, Freitas, and Whiteson (2016b) and Kim et al. (2019) have investigated learning to communicate, where a communication protocol is learned to achieve better coordination. Recent studies on the communication efficiency of MAL are also inspired by the idea that better communication leads to better coordination (Kim, Cho and Sung, 2019; Liu et al., 2020; Ren, Haupt, & Guo, 2021) . Other important objectives revolve around the robustness and resilience of MAL, and there are some attempts on developing MAL algorithms for robustly learning with either malicious/adversarial or failed/dysfunctional learning agents (Li et al., 2019). Some of the existing works concerning the goals mentioned above provide only empirical studies, leaving plenty of room for theoretical investigation.

6. MAL in security applications

In this section, we review some MAL applications in the security domain. We first demonstrate that our proposed MAL framework provides a mathematical toolset to analyze the learning processes considered in these security applications, and then we discuss the importance of information structures in security problems

Most security research and applications focus on intelligent attacks because they can cause the worst damage to the target. In general, many security problems can be formulated as a two-agent learning problem with one attacker and one defender. It can be easily extended to multi-agent cases if there are multiple attackers and defenders. During the attack–defense process, the attacker and defender learn the optimal attack and defense strategies from their observations and eventually reach an equilibrium solution, where the attacker and the defender cannot perform better by deviating the current attack/defense strategy.

When applying learning to security applications, the information structure is of vital importance because it can affect security performance. In most cases, intelligent attackers have information advantages over defenders, which means attackers' information such as the type and relative parameters is not pre-known for security defenders. However, the information structure determines the defender's perception of its surroundings, as well as its awareness of the potential security threats. Therefore, to overcome the information restrictions and perform better defense, the defender can use learning to gather the threat information and to model the attacker.[3] We elaborate on the impact of the information structure in security applications from the following two perspectives.

First, information structures can affect the learning paradigms (Huang & Zhu, 2020b). For example, suppose that the defender is only uncertain of some parameters about the attacker's strategies. In this case, the defender can use *Bayesian learning* to estimate the parameter from the past information for better defense (Huang & Zhu, 2020a). If both the attacker and defender observe perturbed rewards of each other but share no information, then *distributed learning* is more suitable for learning to defend (Zhu & Başar, 2013). If both attacker and defender face an unknown environment, *reinforcement learning* can be applied to learn the optimal attack/defense policy (Huang & Zhu, 2019a, 2019b). We can also observe that the complexity of information structures is related to the complexity of the learning algorithm. For example, Bayesian learning is enough to handle the parameter estimation problems, and RL is unnecessary. As the information structure of interest becomes simpler, we need more sophisticated learning algorithms to learn good attack/defense strategies. This perspective confirms the value and the importance of information structures we discussed before.

Second, a slight change in the information structure may lead to security vulnerability or even system failure. For security applications, the performance of the security strategy can be measured by the value of the objective function and the convergence of the learning algorithm. For example, Huang and Zhu (2020a) investigates the Advanced Persistent Threats (ATP) in cyber–physical systems with one attacker and one defender. Due to the information advantage, the defender does not know the type of the attacker and needs to learn from observations for better defense. Three different ISs are used for defense strategy learning: complete information for both agents, incomplete information for the defender only, and incomplete information for both agents. From their case study, we can observe that the complete information yields a much better utility for the defender than two incomplete information cases, which means the complete information can best protect systems from APT attacks. Accordingly, the complete information yields the lowest utility for the attacker, which means that the system is the least likely to be compromised. In Nguyen, Alpcan, and Basar (2009), the convergence of successful defense strategy relies on the defender's information structure. Additional perturbation or missing information may lead to convergence failure, which means the defender cannot defend the target satisfactorily, causing security vulnerability.

6.1. An illustrative example

We take the Moving Target Defense (MTD) problem in networks (for example, IoT networks) as an example to better illustrate how our framework analyzes the learning for security applications.

MTD is a proactive defense mechanism that allows the defender to dynamically change the security strategies to limit the exposure of network vulnerabilities by increasing the attack cost (Jajodia, Ghosh, Swarup, Wang, & Wang, 2011). We adopt the setting in Zhu and Başar (2013) for MDT, where the defender has to protect a multi-layer system

[3] The same argument applies to the attacker if he also needs to learn the defender's behavior.

Table 2
A summary of M L algorithms from the four categories discussed in this paper. Related characteristics of these algorithms are provided in this summary, such as contexts where they are applied, information structures, as well as the corresponding theoretical guarantees.

Category	Algorithms	Contexts	Information structures		Theoretical guarantees
			Domain knowledge	Online observability	
Stationary Opponents	Fictitious Play (FP) (Brown, 1951)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, R_i$	$\{A_j\}_{j\in\mathcal{N}}$	NE[a]
	Joint action learner (Claus & Boutilier, 1998)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$\{A_j\}_{j\in\mathcal{N}}$	NE
	Two-timescale Q-learning (Leslie & Collins, 2003)	RG	$\mathcal{A}_i$	A_i, R_i	NE
	Individual Q-learning (Leslie & Collins, 2005)	RG	$\mathcal{A}_i$	A_i, R_i	NE
	Distributed FP (Eksin & Ribeiro, 2017)	RG	$\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, R_i$	$\{A_j\}_{j\in\mathcal{N}(i)}, \{M_j\}_{j\in\mathcal{N}(i)}$	NE
	Distributed Best Response (BR)(Swenson et al., 2017)	RG	$\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}$	$\{A_j\}_{j\in\mathcal{N}(i)}, \{M_j\}_{j\in\mathcal{N}(i)}, R_i$	NE
	FP in MG (Sayin et al., 2020)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \mathcal{T}, R_i$	$S, \{A_j\}_{j\in\mathcal{N}}, R_i$	NE
	BR in MG (Leslie et al., 2020)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \mathcal{T}, R_i$	$S, \{A_j\}_{j\in\mathcal{N}}, R_i$	NE
Conjectured Opponents	Minmax Q-learning (Littman, 1994)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$S, \{A_j\}_{j\in\mathcal{N}}, R_i$	NE
	Nash Q-learning (Hu & Wellman, 2003)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$S, \{A_j\}_{j\in\mathcal{N}}, R_i$	NE
	Correlated Q-learning (Greenwald & Hall, 2003)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$S, \{A_j\}_{j\in\mathcal{N}}, R_i$	CE[b]
	Distributed Q-learning (Kar et al., 2013)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$S, \{A_j\}_{j\in\mathcal{N}(i)}, \{M_j\}_{j\in\mathcal{N}(i)}, R_i$	Optimality[c]
	Multi-Agent Actor Critic (Lowe et al., 2017)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$S, \{A_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	Empirical study[d]
	Counterfactual multi-agent policy gradients (Foerster et al., 2018)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$S, \{A_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	Optimality
	Multi-agent actor–critic with networked agents (Zhang et al., 2018)	MG	$\mathcal{S}, \mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}$	$S, \{A_j\}_{j\in\mathcal{N}}, \{M_j\}_{j\in\mathcal{N}(i)}, R_i$	Optimality
	Gradient-based learning (Mazumdar et al., 2020)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, R_i$	$\{A_j\}_{j\in\mathcal{N}}$	NE
Calibrated Opponents	Win-or-learn-fast gradient ascent (Bowling & Veloso, 2002)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	$\{A_j\}_{j\in\mathcal{N}}$	NE
	Win-or-learn-fast policy hill climbing (Bowling, 2004)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	$\{A_j\}_{j\in\mathcal{N}}$	No regret
	Change or learn fast (Cote, Lazaric, & Restelli, 2006)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	$\{A_j\}_{j\in\mathcal{N}}$	Empirical study
	AWESOME (Conitzer & Sandholm, 2007)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	$\{A_j\}_{j\in\mathcal{N}}$	Best response or NE
	Learning by trial and error (Young, 2009)	RG	$\mathcal{A}_i$	A_i, R_i	NE
	Payoff-based learning (Marden et al., 2009)	RG	$\mathcal{A}_i$	A_i, R_i	NE
	RL with change-point detection (Hadoux et al., 2014)	MDP	$\mathcal{S}, \mathcal{A}, \mathcal{T}, R$	S, A, R	Optimality
	RL with quickest change detection (Banerjee et al., 2017)	MDP	$\mathcal{S}, \mathcal{A}, \mathcal{T}, R$	S, A, R	Optimality
Sophisticated Opponents	Level-K (Costa-Gomes et al., 2001)	One-shot game	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	–	Empirical study
	Sophisticated experience-weighted attraction learning (Camerer et al., 2002)	RG	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	A_i, R_i	NE
	Cognitive hierarchy (Camerer et al., 2004)	One-shot game	$\mathcal{N}, \{\mathcal{A}_j\}_{j\in\mathcal{N}}, \{R_j\}_{j\in\mathcal{N}}$	–	Empirical study
	Interactive POMDP (I-POMDP) (Gmytrasiewicz & Doshi, 2005)	MASDM	$\mathcal{G}$	–	Empirical study
	Parameterized I-POMDP (Wunder et al., 2011)	MASDM	$\mathcal{G}$	–	Optimality

[a]NE means convergence to Nash equilibrium.
[b]CE means convergence to correlated equilibrium.
[c]Optimality means that the algorithm achieves a pre-defined optimal condition, depending on the specific learning task.
[d]Empirical study means that there is only empirical results available for the performance of the algorithm.
– For offline planning problems, online observability is not a matter to be discussed.

with N layers. At layer $l \in \{1, \dots, N\}$, there are n_l system vulnerabilities that the attacker can exploit to compromise the system. We denote the vulnerability set as $\mathcal{V}_l := \{v_{l,1}, \dots, v_{l,n_l}\}$. A configuration $c_{l,i}$ constitutes a subset of vulnerabilities in $\mathcal{V}_l$, and the subset is denoted as the vulnerability map (also called the attack surface): $\pi_l(c_{l,i})$. There are m_l possible configuration for layer l and the feasible configuration set $C_l := \{c_{l,1}, \dots, c_{l,m_l}\}$. The attacker can launch an attack $a_{l,i} \in \mathcal{A}_l := \{a_{l,1}, \dots, a_{l,n_l}\}$ to exploit the vulnerability $v_{l,i} \in \mathcal{V}_l$. The attack can successfully cause damage r_l if $a_{l,i} \in \pi_l(c_{l,i})$. Otherwise, the damage is zero. Therefore, the defender tries to select the configuration $c_{l,i} \in C_l$ to avoid the attack and protect the system. The attacker aims to maximize the total attack damage by launching attacks for each layer, while the defender seeks to minimize the overall damage by choosing the proper layerwise configurations. The objective is to find the equilibrium attack/defense strategies. The information is incomplete in this case due to practical considerations. More specifically, both attacker and defender can only observe disturbed reward function $\hat{r}_l$, and the opponent's actions are also unknown. Therefore, two players have to use learning to gain more information about each other for better attack and defense. The MTD problem can be viewed as a two-agent learning problem.

Due to the noncooperative environment, there is no communication between two agents at any time. So the information structure at layer l and time t in MTD for the defender and the attacker are ${}^{D}\mathcal{I}_l^t = \{\hat{r}_l^t, \mathbb{C}_l^t \in$

$C_l\}$ and ${}^{A}\mathcal{I}_l^t = \{\hat{r}_l^t, \mathbb{a}_l^t \in \mathcal{A}_l\}$, respectively.[4] Each agent uses the observed reward sample to estimate the new average reward for future belief and policy generation:

$$
\begin{aligned}
{}^{D}\bar{r}_l^{t+1}(c_{l,h}) &= {}^{D}\bar{r}_l^{t}(c_{l,h}) + {}^{D}\mu^t \mathbf{1}_{\{\mathbb{c}_l^t = c_{l,h}\}}(\hat{r}_l^t - {}^{D}\bar{r}_l^{t}(c_{l,h})), \quad \forall c_{l,h} \in C_l,\\
{}^{A}\bar{r}_l^{t+1}(a_{l,h}) &= {}^{A}\bar{r}_l^{t}(a_{l,h}) + {}^{A}\mu^t \mathbf{1}_{\{\mathbb{a}_l^t = a_{l,h}\}}(\hat{r}_l^t - {}^{A}\bar{r}_l^{t}(a_{l,h})), \quad \forall a_{l,h} \in \mathcal{A}_l,
\end{aligned}
$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function and ${}^{D}\mu^t, {}^{A}\mu^t$ represent the learning rate for the defender and the attacker respectively. For belief generation, each agent deals with the conjectured opponent. It is worth mentioning that there is no explicit belief variable or belief mapping in this case. The belief of the opponents is included in the average reward $\bar{r}_l^t$ and policy generation. In other words, the average reward ${}^{D}\bar{r}_l^t$ and ${}^{A}\bar{r}_l^t$ are indicators of the belief on the attacker and the defender respectively. For policy generation, we first denote ΔC_l and $\Delta\mathcal{A}_l$ as sets of all possible mixed strategies over C_l and $\mathcal{A}_l$ respectively, and denote $\mathbf{f}_l$ and $\mathbf{g}_l$ as probability distribution vectors in ΔC_l and $\Delta\mathcal{A}_l$ respectively. The defender and the attacker solve an regularized optimization problem (DP) and (AP) respectively to generate the next-step policy:

$$
\begin{aligned}
(DP): \sup_{\mathbf{f}_l^{t+1}\in\Delta C_l} & -\sum_{h=1}^{m_l} f_{l,h}^{t+1}\cdot {}^{D}\hat{r}_l^t(c_{l,h}) - {}^{D}\epsilon_l^t \sum_{h=1}^{m_l} f_{l,h}^{t+1}\log\left(\frac{f_{l,h}^{t+1}}{f_{l,h}^{t}}\right),\\
(AP): \sup_{\mathbf{g}_l^{t+1}\in\Delta\mathcal{A}_l} & -\sum_{h=1}^{n_l} g_{l,h}^{t+1}\cdot {}^{A}\hat{r}_l^t(a_{l,h}) - {}^{A}\epsilon_l^t \sum_{h=1}^{n_l} g_{l,h}^{t+1}\log\left(\frac{g_{l,h}^{t+1}}{g_{l,h}^{t}}\right),
\end{aligned}
$$

where ${}^{D}\epsilon_l^t$ and ${}^{A}\epsilon_l^t$ are regularization parameters. Two optimization problems aim to find the best mixed strategies under the current average cost. Since there is no explicit communication, the best response dynamics can be decoupled as (DP) and (AP). It is worth mentioning that by adding a regularization term to the policy generation problems (DP) and (AP), a closed-loop solution for next-step policy can be obtained, which leads to the following analytic learning dynamics:

$$
\begin{aligned}
f_{l,h}^{t+1} &= \left(1 - {}^{S}\lambda_l^t\right) f_{l,h}^t + {}^{S}\lambda_l^t \frac{f_{l,h}^t\cdot\exp\left(-\frac{{}^{S}\hat{r}_l^t(c_{l,h})}{{}^{S}\epsilon_l^t}\right)}{\sum_{h'=1}^{m_l} f_{l,h'}^t\cdot\exp\left(-\frac{{}^{S}\hat{r}_l^t(c_{l,h'})}{{}^{S}\epsilon_l^t}\right)},\\
g_{l,h}^{t+1} &= \left(1 - {}^{A}\lambda_l^t\right) g_{l,h}^t + {}^{A}\lambda_l^t \frac{g_{l,h}^t\cdot\exp\left(-\frac{{}^{A}\hat{r}_l^t(a_{l,h})}{{}^{A}\epsilon_l^t}\right)}{\sum_{h'=1}^{n_l} g_{l,h'}^t\cdot\exp\left(-\frac{{}^{A}\hat{r}_l^t(a_{l,h'})}{{}^{A}\epsilon_l^t}\right)}.
\end{aligned}
$$

It is proved that under mild conditions, the learning algorithm will converge to the equilibrium solution. The optimal attack/defense strategies can be successfully learned.

We mention that if the defender has perfect information (which contains accurate action sets and the reward), the MDT problem exists a unique mixed-strategy equilibrium solution, which can be explored by learning algorithms such as fictitious play (Brown, 1951). From the belief generation perspective, each agent only considers a stationary opponent instead of a conjectured opponent. This observation also demonstrates the value of the information structure and corroborates our previous claim: the information structure can affect the complexity of the learning algorithm.

7. Conclusion

With a mathematical characterization of MAL (see Section 2), this review provides a systematic overview of the state-of-the-art MAL algorithms, with a focus on the information structure. We identified several principled approaches on how the learning agent generates a belief of its opponent, based on the information structure, arriving at a new taxonomy of MAL with four categories: *stationary opponents, conjectured opponents, calibrated opponents* and *sophisticated opponent* (see Section 4). For each category, we elaborate on the role of information structures using concrete algorithms and provide an extensive list of state-of-the-art algorithms classified into these categories (see Table 2). Furthermore, the strengths and limitations of these algorithms are discussed in detail in Section 5. To quantitatively discuss the impact of information structures, we introduce a metric *the value of information* (see Section 5.2), which mathematically displays the information paradox (see Example 1): more information does not necessarily lead to better outcomes. Finally, we point out some promising lines of research in MAL, and especially, we highlight the application of MAL in security studies (see Section 6).

The readers are encouraged to position their research works using our framework for ease of reference and navigation of related (future) works. Meanwhile, by introducing the MAL definition and related notions, such as information structures, the value of information, the review seeks to provide the jump-off point for future research works on MAL studies from the information perspective.

[4] The subscript D and A denote the defender and the attacker respectively.

Acknowledgments

This work is partially supported by grants SES-1541164, ECCS-1847056, CNS-2027884, and BCS-2122060 from National Science Foundation (NSF), grant 20-19829 from DOE-NE, and grant W911NF-19-1-0041 from Army Research Office (ARO).

References

Altman, E. (1999). *Constrained Markov decision processes, Vol. 7*. CRC Press.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2–3), 235–256. http://dx.doi.org/10.1023/a:1013689704352.

Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, *107*(3), 797–817. http://dx.doi.org/10.2307/2118364.

Banerjee, T., Liu, M., & How, J. P. (2017). Quickest change detection approach to optimal control in Markov decision processes with model changes. In *2017 American control conference (ACC)* (pp. 399–405). IEEE, http://dx.doi.org/10.23919/acc.2017.7962986.

Bannon, J., Windsor, B., Song, W., & Li, T. (2020). Causality and batch reinforcement learning: complementary approaches to planning in unknown domains. arXiv preprint arXiv:2006.02579.

Bard, N., Nicholas, D., Szepesvári, C., & Bowling, M. (2015). Decision-theoretic clustering of strategies. In *Workshops at the twenty-ninth aaai conference on artificial intelligence* (pp. 17–25).

Beer, R. D., & Gallagher, J. C. (1992). Evolving dynamical neural networks for adaptive behavior. *Adaptive Behavior*, *1*(1), 91–122. http://dx.doi.org/10.1177/105971239200100105.

Benaïm, M., Hofbauer, J., & Sorin, S. (2005). Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, *44*(1), 328–348. http://dx.doi.org/10.1137/s0363012904439301.

Bond, A. H., & Gasser, L. (2014). *Readings in distributed artificial intelligence*. Morgan Kaufmann, http://dx.doi.org/10.1016/C2013-0-07700-6.

Borkar, V. S. (2005). An actor-critic algorithm for constrained markov decision processes. *Systems & Control Letters*, *54*(3), 207–213. http://dx.doi.org/10.1016/j.sysconle.2004.08.007.

Bouvrie, J., & Maggioni, M. (2012). Multiscale markov decision problems: compression, solution, and transfer learning. arXiv preprint arXiv:1212.1143.

Bowling, M. (2004). Convergence and no-regret in multiagent learning. In *NIPS'04, Proceedings of the 17th international conference on neural information processing systems* (pp. 209–216). Cambridge, MA, USA: MIT Press.

Bowling, M., & Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence*, *136*(2), 215–250. http://dx.doi.org/10.1016/S0004-3702(02)00121-2.

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, *13*(1), 374–376.

Brown, N., Lerer, A., Gross, S., & Sandholm, T. (2019). Deep counterfactual regret minimization. In *Proceedings of machine learning research*: *Vol. 97*, *International conference on machine learning* (pp. 793—802). Long Beach, California, USA: PMLR, URL: http://proceedings.mlr.press/v97/brown19b.html.

Brown, N., & Sandholm, T. (2017). Superhuman AI for heads-up no-limit poker: libratus beats top professionals. *Science*, *359*(6374), 418–424. http://dx.doi.org/10.1126/science.aao1733.

Brown, N., & Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science*, *365*(6456), 885–890. http://dx.doi.org/10.1126/science.aay2400.

Bu, J., Ratliff, L. J., & Mesbahi, M. (2019). Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *ArXiv*, arXiv:1911.04672.

Buesing, L., Weber, T., Zwols, Y., Racaniere, S., Guez, A., Lespiau, J.-B., et al. (2018). Woulda, coulda, shoulda: counterfactually-guided policy search. arXiv preprint arXiv:1811.06272.

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 156–172. http://dx.doi.org/10.1109/TSMCC.2007.913919.

Camerer, C. F. (2011). *Behavioral game theory: experiments in strategic interaction*. Princeton University Press, http://dx.doi.org/10.1016/j.socec.2003.10.009.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, *104*(1), 137–188. http://dx.doi.org/10.1006/jeth.2002.2927.

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*(3), 861–898. http://dx.doi.org/10.1162/0033553041502225.

Chen, J., & Zhu, Q. (2019a). A game- and decision-theoretic approach to resilient interdependent network analysis and design. *Springer Briefs in Electrical and Computer Engineering*, 75–102. http://dx.doi.org/10.1007/978-3-030-23444-7_5.

Chen, J., & Zhu, Q. (2019b). A games-in-games approach to mosaic command and control design of dynamic network-of-networks for secure and resilient multi-domain operations. In G. Chen, & K. D. Pham (Eds.), *Sensors and systems for space applications XII, Vol. 11017* (pp. 189–195). SPIE, International Society for Optics and Photonics, http://dx.doi.org/10.1117/12.2526677.

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI '98/IAAI '98, Proceedings of the fifteenth national/tenth conference on artificial intelligence/innovative applications of artificial intelligence* (pp. 746–752). USA: American Association for Artificial Intelligence.

Colorni, A., Dorigo, M., Maniezzo, V., et al. (1991). Distributed optimization by ant colonies. In *Proceedings of the first european conference on artificial life, Vol. 142*, Paris, France (pp. 134–142).

Conitzer, V., & Sandholm, T. (2007). AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, *67*(1–2), 23–43. http://dx.doi.org/10.1007/s10994-006-0143-1.

Costa-Gomes, M., Crawford, V. P., & Broseta, B. (2001). Cognition and behavior in normal-form games: an experimental study. *Econometrica*, *69*(5), 1193–1235. http://dx.doi.org/10.1111/1468-0262.00239.

Cote, E. M. d., Lazaric, A., & Restelli, M. (2006). Learning to cooperate in multi-agent social dilemmas. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems - AAMAS '06* (pp. 783–785). http://dx.doi.org/10.1145/1160633.1160770.

Coussi-Korbel, S., & Fragaszy, D. M. (1995). On the relation between social dynamics and social learning. *Animal Behaviour*, *50*(6), 1441–1453. http://dx.doi.org/10.1016/0003-3472(95)80001-8.

Da Silva, F. L., & Costa, A. H. R. (2019). A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research*, *64*, 645–703. http://dx.doi.org/10.1613/jair.1.11396.

Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-agent systems: a survey. *IEEE Access*, *6*, 28573–28593. http://dx.doi.org/10.1109/ACCESS.2018.2831228.

Eksin, C., & Ribeiro, A. (2017). Distributed fictitious play for multiagent systems in uncertain environments. *IEEE Transactions on Automatic Control*, *63*(4), 1177–1184. http://dx.doi.org/10.1109/tac.2017.2747767.

Filar, J., & Vrieze, K. (2012). *Competitive markov decision processes*. Springer Science & Business Media, http://dx.doi.org/10.1007/978-1-4612-4054-9, Springer Science & Business Media.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126–1135). PMLR.

Foerster, J. N., Assael, Y. M., de Freitas, N., & Whiteson, S. (2016a). Learning to communicate with deep multi-agent reinforcement learning. In *NIPS'16, Proceedings of the 30th international conference on neural information processing systems* (pp. 2145–2153). Red Hook, NY, USA: Curran Associates Inc..

Foerster, J. N., Assael, Y. M., Freitas, N. d., & Whiteson, S. (2016b). Learning to communicate with deep multi-agent reinforcement learning. *ArXiv*, arXiv:1605.06676.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. In *Thirty-second AAAI conference on artificial intelligence, Vol. 32* (pp. 2974–2982).

Fogel, D. B. (1995). *Evolutionary computation: toward a new philosophy of machine intelligence*. Wiley-IEEE Press.

Fudenberg, D., Drew, F., Levine, D. K., & Levine, D. K. (1998). The theory of learning in games. *vol. 2*, MIT Press.

Fudenberg, D., & Levine, D. K. (1989). Reputation and equilibrium selection in games with a patient player. *JSTOR*, *57*(4), 759–778. http://dx.doi.org/10.2307/1913771, URL: http://www.jstor.org/stable/1913771.

Geramifard, A., Walsh, T. J., Tellex, S., Chowdhary, G., Roy, N., & How, J. P. (2013). A tutorial on linear function approximators for dynamic programming and reinforcement learning. *Foundations and Trends® in Machine Learning*, *6*(4), 375–451. http://dx.doi.org/10.1561/2200000042.

Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, *24*(1), 49–79. http://dx.doi.org/10.1613/jair.1579.

Greenwald, A., & Hall, K. (2003). Correlated-q learning. In *ICML'03: Vol. 20, Proceedings of the twentieth international conference on machine learning* (pp. 242–249).

Guestrin, C., Koller, D., & Parr, R. (2001). Multiagent planning with factored MDPs. In *Proceedings of the 14th international conference on neural information processing systems: natural and synthetic* (pp. 1523–1530). MIT Press.

Guestrin, C., Koller, D., & Parr, R. (2002). Multiagent planning with factored MDPs. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, Vol. 14*. MIT Press, URL: https://proceedings.neurips.cc/paper/2001/file/7af6266cc52234b5aa339b16695f7fc4-Paper.pdf.

Hadoux, E., Beynier, A., & Weng, P. (2014). Sequential decision-making under non-stationary environments via sequential change-point detection. In *Learning over multiple contexts (LMCE)*. Nancy, France: URL: https://hal.archives-ouvertes.fr/hal-01200817.

Heinrich, J., Lanctot, M., & Silver, D. (2015). Fictitious self-play in extensive-form games. In F. Bach, & D. Blei (Eds.), *Proceedings of machine learning research: Vol. 37, Proceedings of the 32nd international conference on machine learning* (pp. 805–813). Lille, France: PMLR, URL: http://proceedings.mlr.press/v37/heinrich15.html.

Hernandez-Leal, P., Kaisers, M., Baarslag, T., & Cote, E. M. d. (2017). A survey of learning in multiagent environments: dealing with non-stationarity. *ArXiv*, arXiv:1707.09183.

Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 750–797. http://dx.doi.org/10.1007/s10458-019-09421-1.

Hertz, J. A., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation*. Addison-Wesley, http://dx.doi.org/10.1201/9780429499661.

Hofbauer, J., & Sigmund, K. (2003). Evolutionary game dynamics. *American Mathematical Society. Bulletin*, *40*(4), 479–519. http://dx.doi.org/10.1090/s0273-0979-03-00988-1.

Hu, J., & Wellman, M. P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, *4*(Nov), 1039–1069.

Huang, L., & Zhu, Q. (2019a). Adaptive honeypot engagement through reinforcement learning of semi-markov decision processes. In *International conference on decision and game theory for security* (pp. 196–216). Springer, http://dx.doi.org/10.1007/978-3-030-32430-8_13.

Huang, Y., & Zhu, Q. (2019b). Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International conference on decision and game theory for security* (pp. 217–237). Springer, http://dx.doi.org/10.1007/978-3-030-32430-8_14.

Huang, L., & Zhu, Q. (2020a). A dynamic games approach to proactive defense strategies against advanced persistent threats in cyber-physical systems. *Computers & Security*, *89*, Article 101660. http://dx.doi.org/10.1016/j.cose.2019.101660.

Huang, L., & Zhu, Q. (2020b). Strategic learning for active, adaptive, and autonomous cyber defense. In *Adaptive autonomous secure cyber systems* (pp. 205–230). Springer, http://dx.doi.org/10.1007/978-3-030-33432-1_10.

Jackson, M. O., & Zenou, Y. (2015). Chapter 3 games on networks. *Handbook of Game Theory with Economic Applications*, *4*, 95–163. http://dx.doi.org/10.1016/b978-0-444-53766-9.00003-3.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., et al. (2019). Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *Science*, *364*(6443), 859–865. http://dx.doi.org/10.1126/science.aau6249, arXiv:1807.01281.

Jajodia, S., Ghosh, A. K., Swarup, V., Wang, C., & Wang, X. S. (2011). *Moving target defense: creating asymmetric uncertainty for cyber threats, Vol. 54*. Springer Science & Business Media, http://dx.doi.org/10.1007/978-1-4614-0977-9.

Jin, C., Allen-Zhu, Z., Bubeck, S., & Jordan, M. I. (2018). Is Q-learning provably efficient? arXiv preprint arXiv:1807.03765.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1–2), 99–134. http://dx.doi.org/10.1016/s0004-3702(98)00023-x.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, *4*, 237–285. http://dx.doi.org/10.1613/jair.301.

Kar, S., Moura, J. M. F., & Poor, H. V. (2013). QD-Learning: a collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Transactions on Signal Processing*, *61*(7), 1848–1862. http://dx.doi.org/10.1109/tsp.2013.2241057, arXiv:1205.0047.

Kaufmann, E., Korda, N., & Munos, R. (2012). Thompson sampling: an asymptotically optimal finite-time analysis. In *Algorithmic learning theory* (pp. 199–213). Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-34106-9_18.

Kim, W., Cho, M., & Sung, Y. (2019). Message-dropout: an efficient training method for multi-agent deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 6079–6086). http://dx.doi.org/10.1609/aaai.v33i01.33016079.

Kim, D., Moon, S., Hostallero, D., Kang, W. J., Lee, T., Son, K., et al. (2019). Learning to schedule communication in multi-agent reinforcement learning. *ArXiv*, arXiv:1902.01554.

Kok, J. R., & Vlassis, N. (2004). Sparse cooperative Q-learning. In *Twenty-first international conference on machine learning - ICML '04* (p. 61). http://dx.doi.org/10.1145/1015330.1015410.

Kuhn, H. W. (2016). 11. Extensive Games and the problem of information. In H. W. Kuhn, & A. W. Tucker (Eds.), *Contributions to the theory of games (AM-28), Volume II* (pp. 193–216). Princeton University Press, http://dx.doi.org/10.1515/9781400881970-012.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., et al. (2019). OpenSpiel: A framework for reinforcement learning in games. *ArXiv*, arXiv:1908.09453.

Lattimore, T., & Szepesvári, C. (2020). *Cambridge core*, *Bandit algorithms*. Cambridge University Press: http://dx.doi.org/10.1017/9781108571401, URL: https://www.cambridge.org/core/books/bandit-algorithms/8E39FD004E6CE036680F90DD0C6F09FC.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Leslie, D. S., & Collins, E. J. (2003). Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Annals of Applied Probability*, *13*(4), 1231–1251. http://dx.doi.org/10.1214/aoap/1069786497.

Leslie, D. S., & Collins, E. J. (2005). Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, *44*(2), 495–514. http://dx.doi.org/10.1137/s0363012903437976.

Leslie, D. S., Perkins, S., & Xu, Z. (2020). Best-response dynamics in zero-sum stochastic games. *Journal of Economic Theory*, *189*, Article 105095. http://dx.doi.org/10.1016/j.jet.2020.105095.

Li, T., Peng, G., Zhu, Q., & Basar, T. (2021). The confluence of networks, games and learning. arXiv:2105.08158,

Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., & Russell, S. (2019). Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 4213–4220). http://dx.doi.org/10.1609/aaai.v33i01.33014213.

Li, T., & Zhu, Q. (2019). On convergence rate of adaptive multiscale value function approximation for reinforcement learning. In *2019 IEEE 29th international workshop on machine learning for signal processing (MLSP)* (pp. 1–6). http://dx.doi.org/10.1109/mlsp.2019.8918816.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *ICML'94*, *Proceedings of the 11th international conference on international conference on machine learning* (pp. 157–163). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Littman, M. L. (2001). Value-function reinforcement learning in markov games. *Cognitive Systems Research*, *2*(1), 55–66. http://dx.doi.org/10.1016/S1389-0417(01)00015-8.

Liu, S., Li, T., & Zhu, Q. (2020). Communication-efficient distributed machine learning over strategic networks: a two-layer game approach. arXiv preprint arXiv:2011.01455.

Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., & Dauphin, Y. (2019). On the pitfalls of measuring emergent communication. *ArXiv*, arXiv:1903.05168.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, *Advances in neural information processing systems 30* (pp. 6379–6390). Curran Associates, Inc., URL: http://papers.nips.cc/paper/7217-multi-agent-actor-critic-for-mixed-cooperative-competitive-environments.pdf.

Luo, R., & Kay, M. (1992). Data fusion and sensor integration: state-of-the-art 1990s. *Data Fusion in Robotics and Machine Intelligence*, 7–135.

Marden, J. R., & Shamma, J. S. (2018). Game-theoretic learning in distributed control. In *Handbook of dynamic game theory* (pp. 511–546). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-44374-4_9.

Marden, J. R., Young, H. P., Arslan, G., & Shamma, J. S. (2009). Payoff-based dynamics for multiplayer weakly acyclic games. *SIAM Journal on Control and Optimization*, *48*(1), 373–396. http://dx.doi.org/10.1137/070680199.

Mataric, M. J. (1998). Using communication to reduce locality in distributed multiagent learning. *Journal of Experimental & Theoretical Artificial Intelligence*, *10*(3), 357–369. http://dx.doi.org/10.1080/095281398146806.

Mazumdar, E., Ratliff, L. J., & Sastry, S. S. (2020). On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, *2*(1), 103–131. http://dx.doi.org/10.1137/18m1231298, arXiv:1804.05464.

Mertens, J.-F., Sorin, S., & Zamir, S. (2015). *Cambridge core*, *Repeated games*. Cambridge: Cambridge University Press, http://dx.doi.org/10.1017/cbo9781139343275, URL: https://www.cambridge.org/core/books/repeated-games/6E933D5A1B0C1116E90EC7333FF72841.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. http://dx.doi.org/10.1038/nature14236.

Moravík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., et al. (2017). DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, *356*(6337), 508–513. http://dx.doi.org/10.1126/science.aam6960, arXiv:1701.01724.

Myerson, R. B. (1979). Incentive compatibility and the bargaining problem. *Econometrica*, *47*(1), 61–73. http://dx.doi.org/10.2307/1912346.

Myerson, R. B. (1991). *Game theory: analysis of conflict*. Cambridge, MA: Harvard University Press.

Naghizadeh, P., Gorlatova, M., Lan, A. S., & Chiang, M. (2019). Hurts to be too early: Benefits and drawbacks of communication in multi-agent learning. In *IEEE INFOCOM 2019 - IEEE conference on computer communications* (pp. 622–630). http://dx.doi.org/10.1109/INFOCOM.2019.8737652.

Neu, G., Jonsson, A., & Gómez, V. (2017). A unified view of entropy-regularized markov decision processes. arXiv preprint arXiv:1705.07798.

Nguyen, K. C., Alpcan, T., & Basar, T. (2009). Security games with incomplete information. In *2009 IEEE international conference on communications* (pp. 1–6). IEEE.

O'Hare, G. M., & Jennings, N. R. (1996). *Foundations of distributed artificial intelligence, Vol. 9*. John Wiley & Sons.

Ouyang, Y., Tavafoghi, H., & Teneketzis, D. (2016). Dynamic games with asymmetric information: common information based perfect bayesian equilibria and sequential decomposition. *IEEE Transactions on Automatic Control*, *62*(1), 222–237. http://dx.doi.org/10.1109/tac.2016.2544936.

Park, K. F., & Shapira, Z. (2017). Risk and uncertainty. In M. Augier, & D. J. Teece (Eds.), *The palgrave encyclopedia of strategic management* (pp. 1–7). London: Palgrave Macmillan UK, http://dx.doi.org/10.1057/978-1-349-94848-2_250-1.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

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

Ren, J., Haupt, J., & Guo, Z. (2021). Communication-efficient hierarchical distributed optimization for multi-agent policy evaluation. *Journal of Computer Science*, *49*, Article 101280. http://dx.doi.org/10.1016/j.jocs.2020.101280.

Sandholm, T. (2007). Perspectives on multiagent learning. *Artificial Intelligence*, *171*(7), 382–391. http://dx.doi.org/10.1016/j.artint.2007.02.004.

Sayin, M. O., Parise, F., & Ozdaglar, A. (2020). Fictitious play in zero-sum stochastic games. *ArXiv*, arXiv:2010.04223.

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, *4*(2), 107–194. http://dx.doi.org/10.1561/2200000018.

Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv:1610.03295.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, *39*(10), 1095–1100. http://dx.doi.org/10.1073/pnas.39.10.1095.

Shoham, Y., Powers, R., & Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, *171*(7), 365–377. http://dx.doi.org/10.1016/j.artint.2006.02.006.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *International conference on machine learning, Vol. 32* (pp. 387–395). URL: http://proceedings.mlr.press/v32/silver14.html.

Solan, E., & Vieille, N. (2015). Stochastic games. *Proceedings of the National Academy of Sciences*, *112*(45), 13743–13746. http://dx.doi.org/10.1073/pnas.1513508112.

Stone, P., & Veloso, M. (2000). Multiagent systems: a survey from a machine learning perspective. *Autonomous Robots*, *8*(3), 345–383. http://dx.doi.org/10.1023/a:1008942012299.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., et al. (2018). Value-decomposition networks for cooperative multi-agent learning based on team reward. In *AAMAS '18*, *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 2085–2087). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction*. MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1–2), 181–211. http://dx.doi.org/10.1016/s0004-3702(99)00052-1.

Swenson, B., Eksin, C., Kar, S., & Ribeiro, A. (2017). Distributed inertial best-response dynamics. *IEEE Transactions on Automatic Control*, *63*(12), 4294–4300. http://dx.doi.org/10.1109/tac.2018.2817161, arXiv:1605.00601.

Swenson, B., Murray, R., & Kar, S. (2018). On best-response dynamics in potential games. *SIAM Journal on Control and Optimization*, *56*(4), 2734–2767. http://dx.doi.org/10.1137/17m1139461.

Tang, D., Tavafoghi, H., Subramanian, V., Nayyar, A., & Teneketzis, D. (2021). Dynamic games among teams with delayed intra-team information sharing. *ArXiv*, arXiv:2102.11920.

Tardos, E., Elkind, E., Vohra, R., Braverman, M., Mao, J., Schneider, J., et al. (2018). Selling to a no-regret buyer. In *Proceedings of the 2018 ACM conference on economics and computation* (pp. 523–538). http://dx.doi.org/10.1145/3219166.3219233.

Tembine, H., Zhu, Q., & Baar, T. (2014). Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, *59*(4), 835–850. http://dx.doi.org/10.1109/tac.2013.2289711.

Tuyls, K., & Weiss, G. (2012). Multiagent learning: basics, challenges, and prospects. *AI Magazine*, *33*(3), 41. http://dx.doi.org/10.1609/aimag.v33i3.2426, URL: https://ojs.aaai.org/index.php/aimagazine/article/view/2426.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292. http://dx.doi.org/10.1007/BF00992698.

Weibull, J. W. (1997). *Evolutionary game theory*. MIT Press.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John wiley & sons.

Wright, J. R., & Leyton-Brown, K. (2010). Beyond equilibrium: predicting human behavior in normal-form games. In *24th AAAI conference on artificial intelligence*. http://dx.doi.org/10.1145/1807406.1807449.

Wunder, M., Kaisers, M., Yaros, J. R., & Littman, M. (2011). Using iterated reasoning to predict opponent strategies. In *AAMAS '11, The 10th international conference on autonomous agents and multiagent systems - Volume 2* (pp. 593–600). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Xu, X., & Zhao, Q. (2020). Distributed no-regret learning in multiagent systems: challenges and recent developments. *IEEE Signal Processing Magazine*, *37*(3), 84–91. http://dx.doi.org/10.1109/msp.2020.2973963.

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE*, *87*(9), 1423–1447. http://dx.doi.org/10.1109/5.784219.

Young, H. P. (2009). Learning by trial and error. *Games and Economic Behavior*, *65*(2), 626–643. http://dx.doi.org/10.1016/j.geb.2008.02.011.

Zamir, S. (2009). BayesIan games: games with incomplete information. In R. A. Meyers (Ed.), *Encyclopedia of complexity and systems science* (pp. 426–441). New York, NY: Springer New York, http://dx.doi.org/10.1007/978-0-387-30440-3_29.

Zhang, J., & Bareinboim, E. (2017). Transfer learning in multi-armed bandits: a causal approach. In *Proceedings of the 16th conference on autonomous agents and multiagent systems* (pp. 1778–1780). http://dx.doi.org/10.24963/ijcai.2017/186.

Zhang, K., Yang, Z., & Baar, T. (2019). Multi-agent reinforcement learning: a selective overview of theories and algorithms. arXiv:1911.10635,

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

Zhu, Q., & Başar, T. (2013). Game-theoretic approach to feedback-driven multi-stage moving target defense. In *International conference on decision and game theory for security* (pp. 246–263). Springer, http://dx.doi.org/10.1007/978-3-319-02786-9_15.

Zhu, Q., & Başar, T. (2015). Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems. *IEEE Control Systems Magazine*, *35*(1), 46–65. http://dx.doi.org/10.1109/MCS.2014.2364710.

Zhu, Q., Tembine, H., & Baar, T. (2010). Heterogeneous learning in zero-sum stochastic games with incomplete information. In *49th IEEE conference on decision and control (CDC)* (pp. 219–224). http://dx.doi.org/10.1109/cdc.2010.5718053.

Zhu, Q., Tembine, H., & Başar, T. (2011). Distributed strategic learning with application to network security. In *Proceedings of the 2011 American control conference* (pp. 4057–4062). IEEE, http://dx.doi.org/10.1109/ACC.2011.5991373.

Zhu, Q., Tembine, H., & Basar, T. (2013). Hybrid learning in stochastic games and its applications in network security. *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, *17*(14), 305–329. http://dx.doi.org/10.1002/9781118453988.ch14.

Zinkevich, M., Johanson, M., Bowling, M., & Piccione, C. (2007). Regret minimization in games with incomplete information. In *NIPS'07*, *Proceedings of the 20th international conference on neural information processing systems* (pp. 1729–1736). Red Hook, NY, USA: Curran Associates Inc..