Cooperative and Competitive Multi-Agent Systems: From Optimization to Games

Jianrui Wang, Yitian Hong, Jiali Wang, Jiapeng Xu, Yang Tang, *Senior Member, IEEE*, Qing-Long Han, *Fellow, IEEE*, and Jürgen Kurths

***Abstract*—Multi-agent systems can solve scientific issues related to complex systems that are difficult or impossible for a single agent to solve through mutual collaboration and cooperation optimization. In a multi-agent system, agents with a certain degree of autonomy generate complex interactions due to the correlation and coordination, which is manifested as cooperative/competitive behavior. This survey focuses on multi-agent cooperative optimization and cooperative/non-cooperative games. Starting from cooperative optimization, the studies on distributed optimization and federated optimization are summarized. The survey mainly focuses on distributed online optimization and its application in privacy protection, and overviews federated optimization from the perspective of privacy protection mechanisms. Then, cooperative games and non-cooperative games are introduced to expand the cooperative optimization problems from two aspects of minimizing global costs and minimizing individual costs, respectively. Multi-agent cooperative and non-cooperative behaviors are modeled by games from both static and dynamic aspects, according to whether each player can make decisions based on the information of other players. Finally, future directions for cooperative optimization, cooperative/non-cooperative games, and their applications are discussed.**

***Index Terms*—Cooperative games, counterfactual regret minimization, distributed optimization, federated optimization, fictitious self-play, mean field games, multi-agent reinforcement learning, non-cooperative games.**

Manuscript received January 5, 2022; revised February 8, 2022; accepted February 26, 2022. This work was supported in part by the National Natural Science Foundation of China (Basic Science Center Program: 61988101), the Sino-German Center for Research Promotion (M-0066), the International (Regional) Cooperation and Exchange Project (61720106008), the Programme of Introducing Talents of Discipline to Universities (the 111 Project) (B17017), and the Program of Shanghai Academic Research Leader (20XD1401300). Recommended by Associate Editor Tao Yang. *(Jianrui Wang and Yitian Hong contributed equally to this work. Corresponding author: Yang Tang and Qing-Long Han.)*

Citation: J. R. Wang, Y. T. Hong, J. L. Wang, J. P. Xu, Y. Tang, Q.-L. Han, and J. Kurths, "Cooperative and competitive multi-agent systems: From optimization to games," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 5, pp. 763–783, May 2022.

J. R. Wang, Y. T. Hong, J. L. Wang and Y. Tang are with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mail: jianruiwang@mail.ecust.edu.cn; ythong1314@mail.ecust.edu.cn; jialiwang@mail.ecust.edu.cn; yangtang@ecust.edu.cn).

J. P. Xu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: jxu@uwindsor.ca).

Q.-L. Han is with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, VIC 3122, Australia (e-mail: qhan@swin.edu.au).

J. Kurths is with the Potsdam Institute for Climate Impact Research, 14473 Potsdam, and also with the Institute of Physics, Humboldt University of Berlin, 12489 Berlin, Germany (e-mail: juergen.kurths@pikpotsdam.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105506

I. Introduction

In the past few decades, multi-agent systems have received considerable attention [1]–[3] due to the development of automatic control, computer technology and artificial intelligence. Multi-agent systems are able to solve complex problems through interactions between agents, and successfully improve the efficiency and robustness [4]–[6], which is a current research hotspot in various fields, such as intelligent transportation [7] and smart grids [8]. In multi-agent systems, the interactions between agents may be reflected in cooperation and/or competition. There are several factors that influence the cooperative/competitive behaviors of the agents, such as cost functions, neighboring information and environmental factors, which affect the performance of multi-agent systems in the optimization and decision-making process [9]. This survey analyzes cooperative/non-cooperative behaviors of multi-agent systems from optimization to games.

From the perspective of cooperative optimization, agents cooperate and diffuse information so as to obtain a globally optimal solution [10]. Cooperative optimization can be divided into centralized optimization and distributed optimization from the perspective of network structures [11]. Due to the advantages of processing large-scale data and improving system robustness [12], this survey focuses on distributed optimization. Since there is no prior knowledge of the objective functions when the information is highly uncertain and unpredictable in many applications, distributed online optimization is put forward and able to address this problem. The survey summarizes recent work on distributed online optimization and its applications concerned with privacy protection. Afterwards, specifically for the privacy protection issues, federated optimization [13] is introduced to provide additional privacy protection by leveraging the computing power of users to save significant network bandwidth.

In multi-agent systems, due to the heterogeneity of agents, the complexity of working environments, and the diversity of system objectives, game theory is introduced to model the cooperative/competitive behaviors for individual/global optimization goals. Games can be divided into cooperative games and non-cooperative games, judging from the cooperative/non-cooperative behaviors of agents. Similarly, games can also be classified into static games and dynamic

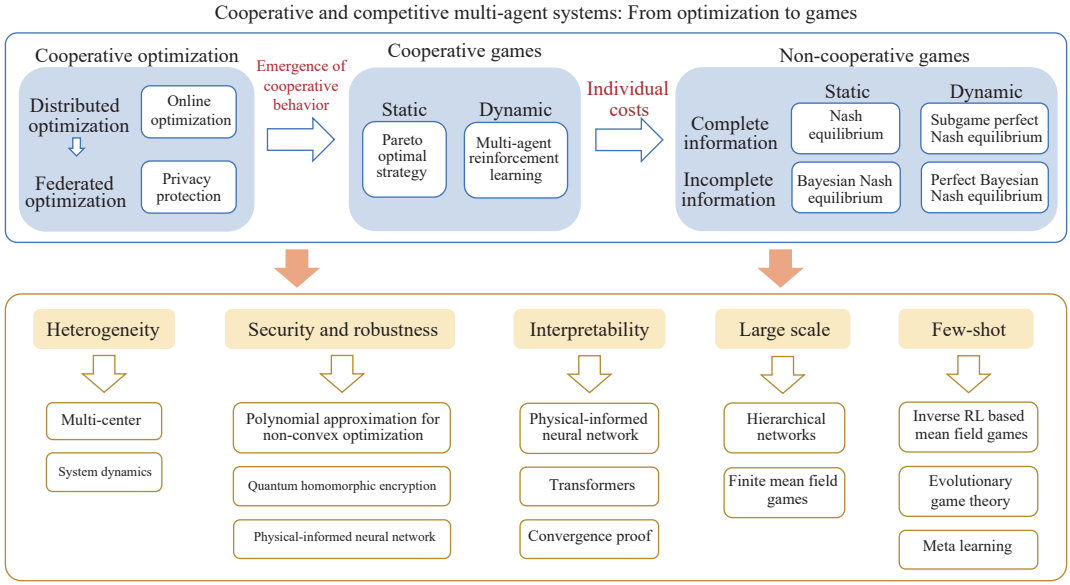

Fig. 1. The overall structure of the survey.

games according to the action sequence of agents [14]. Cooperative games are used to ensure cooperation behavior by considering minimizing the global cost function [15]. In static cooperative games, agents seek for Pareto optimality through multi-objective optimization [16]. While in dynamic cooperative games, Markov decision processes are generalized to multiple cooperating decision-makers, and agents obtain a series of optimal strategies [17].

However, agents may not cooperate in all situations, which means that there may exist competitions in multi-agent systems, such as pursuit-evasion games [18] and capture-the-flag games [19], [20]. Unlike cooperative games, non-cooperative games maximize individual payoffs or minimize individual costs to achieve the Nash equilibrium (NE) [14]. In static non-cooperative games, complete information is usually considered to analytically solve the NE solution. In practice, however, agents do not necessarily know what their exact costs are or what types of opponents they are. Therefore, considering the situation of incomplete information, Bayesian games are useful to obtain the NE solution [21]. Similarly, in dynamic non-cooperative games, under the condition of complete information, researchers often minimize the cost function and maximize the reward function to find the optimal strategy for normal-form games [14], and construct a subgame and solve the NE through Nash backward induction [22] for extensive-form games. Under the condition of incomplete information, game-theoretic solutions, such as regret matching [23] and fictitious play [24], are widely used and combined with artificial intelligence algorithms, such as reinforcement learning (RL) [25] and deep learning [26]. Moreover, mean field games (MFG) are considered to analytically solve the NE solution for large-scale dynamic non-cooperative games [27]. In recent years, researches on dynamic non-cooperative games have made great breakthroughs, such as AlphaZero [28], Pluribus [29] and AlphaStar [30].

In this survey, we analyze the cooperative and competitive of multi-agent systems from three aspects: Cooperative optimization, cooperative games, and non-cooperative games, and summarize future works and applications. At present, many researchers have published surveys on cooperative or/and non-cooperative multi-agent systems. Unlike the survey on cooperative optimization [31], we focus on the recent work of distributed online optimization and the research of distributed online optimization and federated optimization in privacy protection; unlike the survey on games [32], we focus on both cooperative games and non-cooperative games from the perspective of static games and dynamic games, respectively; unlike the survey on both cooperative optimization and games [33], we bridge the transition from cooperative optimization to games, that is, cooperative games, which aims at analyzing the emergence of cooperative behavior.

This survey is organized as follows. Section II introduces cooperative optimization composed of distributed optimization and federated optimization. Section III focuses on cooperative games and non-cooperative games, from static and dynamic aspects respectively. Finally, Section IV summarizes potential future work and applications. The overall structure of this survey is reflected in Fig. 1, and the abbreviations in this survey are summarized in Table I.

II. Cooperative Optimization

Cooperative optimization is inspired by the principle of teamwork, which considers the interactions among agents to optimize variables, and provided with the objective function. Cooperative optimization has a solid theoretical foundation and outstanding practical performance, which has aroused widespread attention and payoffs. In this section, we mainly focus on two kinds of cooperative optimizations with different structures but similar fields of application, namely distributed optimization [31] and federated optimization [13]. The differences between the two are shown in Table II. In distributed optimization, balanced and independent and identically distributed (IID) data are processed on fewer devices with unlimited communication, while in federated optimization, unbalanced and non-IID data are processed on

TABLE I
SUMMARY OF ABBREVIATIONS IN THIS SURVEY

Abbreviation	Full name
AC	Actor-critic algorithms
BNE	Bayesian Nash equilibrium
CC	Centralized critic
COM	Communication learning
CFR	Counterfactual regret minimization
CTDE	Centralized training and decentralized execution
DDPG	Deep deterministic policy gradient
DOO	Distributed online optimization
DOCO	Distributed online convex optimization
DQN	Deep Q network
FEDAVG	Federated averaging
FEDSGD	Federated stochastic gradient descent
FP	Fictitious play
IGM	Individual-global-max
IQL	Independent Q-learning
MADDPG	Multi-agent deep deterministic policy gradient
MARL	Multi-agent reinforcement learning
MFG	Mean field games
NE	Nash equilibrium
NFSP	Neural fictitious self-play
PBNE	Perfect Bayesian Nash quilibrium
QL	Q-learning
RL	Reinforcement learning
RNN	Recurrent neural network
VD	Value decomposition

TABLE II
THE DIFFERENCES BETWEEN DISTRIBUTED OPTIMIZATION AND FEDERATED OPTIMIZATION

	Distributed optimization	Federated optimization
Central node	No	Yes
Number of nodes	Normal	Massively
Data	Balanced, and IID	Unbalanced and non-IID
Communication	Unlimited	Limited
Scenario	Geographic distribution	Data distribution

massive devices with limited communication, which is available to a central server.

A. Distributed Optimization

In distributed optimization of multi-agent systems, agents cooperate to minimize the global cost function, which is a sum of local cost functions, based on its own information and neighboring information. Distributed optimization is widely studied in various fields, such as wireless sensor networks [34] and integrated energy networks [35]. Unlike the existing surveys, such as distributed optimization algorithms for undirected graphs [31] and distributed optimization for electric power systems [36], this survey mainly focuses on the latest research hotspots in recent years, i.e., distributed online optimization and its applications in privacy protection.

Distributed online optimization: In traditional distributed optimization issues, it is usually assumed that every agent knows its local private cost function in advance [31]. However, in certain practical scenarios, there is no prior knowledge and the cost function is time-varying because of the uncertain information. Therefore, the distributed online optimization (DOO) is proposed to minimize the accumulation of time-varying global cost functions [37]. In order to make the problem mathematically tractable, the existing works in [38], [39] are often settled in convex optimization. Distributed online convex optimization (DOCO) aims at minimizing the objective function and reducing the regret, thereby quantifying the performance of DOO algorithms [40]. Regret can be divided into two types, static regret represents the comparison between the computed candidate optima and the best-fixed decision in hindsight, while dynamic regret considers the nature of the minimum value of the global cost function [41].

The DOCO problem can be constructed with a global cost function f_t [39], which is composed of a set of local cost functions over a communication network with n agents, i.e.,

$$f_t(x) := \frac{1}{n}\sum_{i=1}^{n} f_{i,t}(x) \tag{1}$$

where each agent $i \in \{1,\ldots,n\}$ is required to generate a decision point $x \in \mathbb{R}^d$. The global cost function is represented by $f_{i,t} : \mathbb{R}^d \to \mathbb{R}$ at each time step $t \in \{1,\ldots,T\}$ for all i with possible constraints, where T is the total time of the optimization process. In that way, static regret is given in the distributed setting by

$$R_T^s := \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} f_t(x_{i,t}) - \min_{x}\sum_{t=1}^{T} f_t(x_t). \tag{2}$$

Similarly, in the dynamic scenario, dynamic regret can be defined as

$$R_T^d := \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} f_t(x_{i,t}) - \sum_{t=1}^{T}\min_{x_t} f_t(x_t). \tag{3}$$

Then, we focus on the solution to the DOCO problems. We first consider an ideal situation without constraints in DOCO, in that way, the online subgradient is an idea to solve the DOCO problems [42]. In recent years, with exact gradient information, gradient tracking based methods are suitable for handling time-varying cost functions. The dependency assumption is removed in [43], where the local objective functions are strongly convex in the gradient tracking algorithm. Then, when taking both exact gradients and stochastic/noisy gradients into consideration, distributed online gradient tracking algorithm is proposed in [39] to handle the DOCO problems with an aggregative variable. In addition, it is not necessary to assume any prior information on system evolution when considering sparse time-varying optimization, and the model proposed in [44] is suitable for adversarial frameworks as well. Considering the heterogeneity of network nodes, the distributed any-batch mirror descent

TABLE III
SUMMARY OF DOO

Year	Reference	Approach	Communication topology	Constraints type	Convex/Non-convex	Convergence	Privacy protection
2016	Hosseini *et al.* [45]	Dual sub-gradient averaging	Directed	Global set	Convex	$O(\sqrt{T})$	√
2017	Shahrampour and Jadbabaie [37]	Mirror descent algorithm	Undirected	Global set	Convex	$O(\sqrt{T})$	
2018	Mazzi *et al.* [46]	Automatically alleviating contingencies	Directed	Ramping constraints	Convex	Sublinear	
2019	Lu *et al.* [47]	Auxiliary optimization strategies	Directed	Global set	Non-convex	$O\left(T^{3/4}\cdot\sqrt{T}\right)$	
2019	Zhang *et al.* [43]	Gradient tracking algorithm	Undirected	Unconstrained	Convex	Linear	
2020	Fosson [44]	Sparse time-varying algorithm	Undirected	Unconstrained	Convex	Sublinear	
2020	Yi *et al.* [48]	Primal-dual dynamic mirror descent algorithm	Directed	Coupled inequality and local set	Convex	Sublinear	
2020	Eshraghi and Liang [38]	Any-batch mirror descent algorithm	Undirected	Unconstrained	Convex	$O(\sqrt{T})$	
2020	Yuan *et al.* [49]	Approximate mirror descent algorithm	Undirected	Global set	Convex	$O(1/\sqrt{T})$	
2020	Lu *et al.* [50]	Differentially stochastic subgradient-push algorithm	Directed	Unconstrained	Convex	$O(\sqrt{T})$	√
2021	Li *et al.* [39]	Primal-dual push-sum algorithm	Directed	Global set	Convex	$O(1/\sqrt{T})$	
2021	Lu and Wang [51]	Primal-dual dynamic mirror descent algorithm	Directed	Local set	Non-convex	$O(\sqrt{T})$	

algorithm is proposed in [38], which is based on the distributed mirror descent with a fixed computing time every round, ensuring the speed of the overall convergence process of heterogeneous nodes. Node heterogeneity is also reflected in the objective function consists of two parts, a time-varying loss function and a regularization function. In this case, DOO algorithms based on approximate mirror descent are raised in [49].

In addition, considering that constraints are common in actual distributed optimization problems [45], [48], auxiliary constraints are added to the distributed optimization algorithm to improve the practicality of the optimization results. For the case where there are constraints in DOCO problems, when DOCO problems are studied under the global/local set constraints, a dual sub-gradient averaging algorithm is proposed in [45] in the DOCO case. Besides, the celebrated mirror descent algorithm is improved in [37] and a regret bound is established to satisfy the constraints. It is worth noting that when the locally cost functions are considered strongly pseudoconvex, auxiliary optimization strategies are utilized to handle DOO in [47]. After that, they assume the objective function to be nonconvex, and propose a DOO algorithm based on the consensus and the mirror descent algorithm in [51]. When DOCO is constrained by coupled inequality constraints, the objective functions and constraint functions will be revealed over time. In [48], a distributed online primal-dual dynamic mirror descent algorithm is raised to investigate the time-varying coupled inequality constraints. After that, the primal-dual algorithm is modified in [39], promoting the result to unbalanced graphs without making any assumption about bounded parameters.

Privacy protection: In some applications, local information is likely to be private and needs to be safely protected, such as smart grids, financial systems, medical treatments, etc. Although there exist research works on privacy protection in distributed optimization [35], most of them focus on static optimization. Note that when the agents share information, the privacy of the agents may be leaked at any time, so it is necessary to introduce DOO algorithms to solve privacy protection issues. Security is the top priority and is often treated as a constraint in DOCO issues [46]. Take smart grid as an example, contingencies, such as line overloads and voltage violations, should be immediately and automatically mitigated by the network itself. A distributed correction optimization approach is put forward in [52], in order to mitigate line overloads in an online closed-loop way. The system measurements and security constraints in the proposed DOCO algorithm ensure security.

Under the premise of ensuring security, there exist two options for implementing privacy protection in distributed optimization, that is, differential privacy and information encryption [53]. From the aspect of privacy masking, differential privacy is one of the insightful privacy strategies. In the DOCO problem of [50], differential privacy is introduced, and a distributed stochastic subgradient-push algorithm considering differential privacy is also proposed. Although the local cost functions are masked, they can also achieve sublinear regrets for diverse cost functions. However, the increased noise that makes differential privacy more accessible will inevitably affect the data availability, leading to a trade-off between privacy level and calculation accuracy [54]. Therefore, information encryption is utilized to protect privacy indirectly, because encryption techniques cannot be directly applied to distributed optimization without third-party assistance. As seen from Table III, the privacy protection in DOO requires in-depth research, and there are several expansion directions, such as complex constraints and non-convex. It is worth noting that there are already distributed

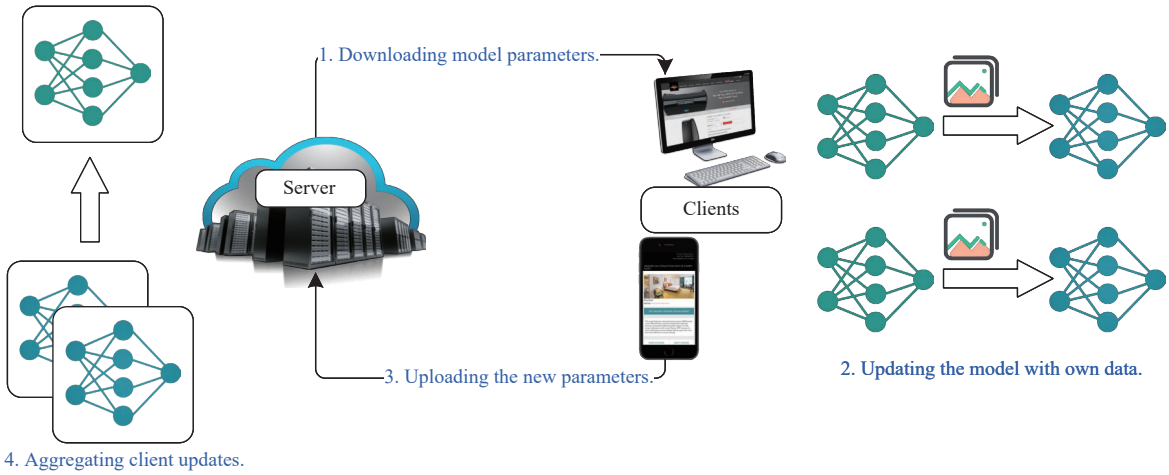

Fig. 2. The federated learning framework.

offline optimization algorithms to solve constrained/unconstrained non-convex optimization problems over directed/undirected communication graphs [55]–[59]. Therefore, non-convex DOO will be a trend in future research to address physical applications where optimization problems are non-convex.

B. Federated Optimization

In distributed optimization, the data centers require to process a large amount of data on fewer devices. In federated optimization, the data volume of a single device is small, while the number of devices is huge [60]. This data distribution is very common in modern mobile devices, but the individual data are basically privacy sensitive. Federated optimization is a structural concept first proposed by Konečnỳ *et al.* [13], which conducts structural improvements to distributed optimization especially for privacy protection issues. It is worth noting that most machine learning algorithms require a large amount of data for training.

How to integrate multi-party data for machine learning training under the premise of privacy is a worth considering problem. Without considering individual data privacy, data sharing is a simple way to realize the training of machine learning algorithms. As individuals pay more attention to individual privacy, general data sharing is not allowed. Federated learning is a solution for federated optimization, which focuses on adapting traditional machine learning to the realistic background of privacy protection and data island. According to different data distribution in practical problems, federated learning can be divided into three categories [61]: Horizontal federated learning, vertical federated learning and federated transfer learning.

Fig. 2 briefly shows a very common structure of horizontal federated learning, specifically as follows:

1) Participants download the latest global model from the server;

2) Participants update local model parameters according to their existing data;

3) Participants pass the trained local model back to the server in an encrypted form;

4) The server consolidates the various models received and updates the global model parameters.

Repeat these four steps until convergence. In this way, distributed training is realized without disclosing privacy. Unlike privacy protection in distributed optimization, thanks to the structure of federated optimization, model aggregation is a unique privacy protection mechanism for federated learning. Similarly to distributed optimization, federated learning also uses differential privacy and information encryption to achieve privacy protection. Therefore, we focus on three main categories of privacy mechanisms in federated learning: model aggregation [60], homomorphic encryption [62] and differential privacy [63].

Model aggregation: Model aggregation is one of the most famous privacy mechanisms in federated learning, which trains a global model with distributed data, but cannot directly use data from each client. A natural way is to use the model parameters of the distributed clients to do an aggregation. Under the privacy mechanism of model aggregation, federated stochastic gradient descent (FEDSGD) and federated averaging (FEDAVG) are proposed in [60], which lay the foundation for model aggregation algorithms. FEDSGD directly migrates the traditional optimization algorithm – stochastic gradient descent (SGD) into the federated framework. This method has high computational efficiency, but it needs a large number of expensive communication rounds with the clients to update the global model's parameters. Unlike FEDSGD, FEDAVG expects each client to conduct multiple rounds of training for local data, and then transmit the models' parameters to the server, the server later conducts a weighted average of the parameters of local models and finally integrates them into the global model. By distributing the computation to each client, FEDAVG greatly reduces the communication cost of the training process.

Most of the research results focus on developing algorithms to further reduce communication based on FEDAVG [64]–[68]. Considering the permutation invariance of neural networks, that is, changing the arrangement of parameters, the performance of neural networks remains unchanged. It is inefficient to directly average the network parameters of each client according to the coordinates. Federated neural matching proposes the idea of matching the client neural network before average [64]. Federated matched averaging further extends

federated neural matching from fully connected network module to recurrent neural network (RNN) module and convolution module [65]. The convergence rate reduction caused by client-drift issue in FEDAVG is considered in [66] and the authors use control variable to correct the issue in client's local updates. To achieve less updates in the global server network, the contribution of client capabilities is estimated at [68]. In addition, a attention mechanism is introduced in [67] to improve the model aggregation capability.

The model aggregation method uses plaintext parameters instead of the raw data to achieve privacy protection and proves that FEDAVG converges on some data problems that are non-independent and identically distributed. However, the plaintext parameters still has the risk of privacy disclosure [69].

Homomorphic encryption: Data protection has always been a problem considered by traditional cryptography. Traditional cryptography realizes data protection by encrypting data [70]. Only by using the correct key, the raw data can be extracted from ciphertext. In this case, data can be stored anywhere, and privacy disclosure will not occur. In cryptography, if the cost of cracking the key exceeds the gain, we believe it is computationally, or conditionally secure. Homomorphic encryption is a form of encryption which allows users to perform the calculation of encrypted data without decryption. These calculation results are retained in the form of encryption, and the decryption of the calculation result will produce the same output as the operation on unencrypted data [71]. Assume that E is an encryption algorithm, M represents the information that needs to be encrypted, $\odot$ represents one of the calculation operations. If the following equation is satisfied:

$$E(m_1)\odot E(m_2)=E(m_1\odot m_2),\ \ \forall m_1,\ m_2\in M \tag{4}$$

the encryption algorithm E has a homomorphic property for the operation $\odot$. The current mainstream homomorphic Encryption algorithms include partially homomorphic encryption and fully homomorphic encryption [72]. Fully homomorphic encryption is powerful, but partially homomorphic encryption requires less performance overhead. Most of homomorphic encryption works focus on how to combine with federated learning in different scenarios [62], [73], [74].

Under vertical data distribution, combining additively homomorphic encryption with federated logistic regression to achieve encrypted training is proposed in [62]. Moreover, on the cross-silo federated learning, the minimum homomorphic encryption requirement is discussed in [73]. In a more practical application background such as different industrial devices, traditional kmeans and AdaBoost combined with homomorphic encryption are discussed in [74].

Differential privacy: Recall that model aggregation has the risk of privacy disclosure, and the use of homomorphic encryption technology can ensure data security, but even partially homomorphic encryption algorithms have the problem of large computational costs. How to make a compromise between computational cost and privacy protection, some researchers turn their perspective to differential privacy. Differential privacy is mainly used to prevent privacy disclosure caused by statistical information. Adding perturbation is the most commonly used method in differential privacy [75]. Each participant disrupts his/her own data, and then sends the disturbed data to the server. In this case, no attackers can get individual privacy data by statistical calculation of the information on the communication channels. The primary perturbation mechanism can be roughly divided into three categories: Laplace mechanism and exponential mechanism and randomized response mechanism. Similarly to homomorphic encryption, federated learning combined with differential privacy also mainly considers application in various scenarios: Aircomp federated learning [76], blockchain-enabled federated learning [74] and mobile edge computing [63].

Among the three privacy mechanisms mentioned above, the model aggregation algorithm is the simplest one to implement but has the high risk of privacy disclosure. Homomorphic encryption has the best confidentiality but is computationally complex. Differential privacy ranks the second among the three methods in terms of complexity and confidentiality, but the accuracy may decline due to perturbation. It is necessary to select appropriate privacy mechanisms according to specific application scenarios.

III. Game Theory

In game theory, a game concerns the optimal strategic interactions between agents. Games provide complex systems science with a systematic approach for deciding a series of Pareto-efficient strategies in cooperative situations or the best strategy in non-cooperative situations, which have attracted a great deal of attention in recent years. From the perspective of the timing of behavior, games can be divided into static games and dynamic games [14]. Static games mean that the participants choose at the same time or although not at the same time, but the later participant does not know what specific actions the former participant took; on the other hand, dynamic games mean that the participants' actions have a sequential order, and the later participant can observe the actions chosen by the former participant.

A. Cooperative Games

Cooperative games are also called positive sum games and concerned with the participants forming alliances and working together to seek to achieve their common goals [77]. Similarly to cooperative optimization, cooperative games also consider coordinating multiple individuals to achieve unified goals. The difference is that cooperative optimization focuses on the analysis of optimization goals, while cooperative games focus on the emergence of cooperative behaviors.

Static cooperative games: If two or more players agree to cooperate while playing games, they will help each other minimize their costs, as long as they do not decrease their payoffs. This leads to the concept of absolutely cooperative solutions or Pareto optimal solutions in games. Such a solution ensures that each player cannot change their strategy for a better solution. From this perspective, we explore the Pareto optimal solutions under the background of static, continuous cooperative games. The Pareto optimal solutions can be described as follows:

$$U_i(\hat{\pi}_i) \geq U_i(\pi_i), \quad \forall i \in \{1,\ldots,n\}, \quad \forall \pi_i \in \Pi_i, \ \forall s \in \mathcal{S} \tag{5}$$

where π_i is a joint policy, $U_i(s)$ is the expected long-term return of agent i in state s, Π_i is the set of all possible policies for agent i, and $\{\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n\}$ are the Pareto optimal solutions of the cooperative games. Usually there is not only one Pareto optimal solution, and the hyperplane composed of multiple Pareto optimal solutions is called Pareto front. The Pareto front solutions are also a set of undominated solutions [16].

Nowadays, most of researches on Pareto front are based on multi-objective optimization [78]. Correspondingly, our problem can also be described as

$$\{\min(-U_1(\pi_1)), \min(-U_1(\pi_2)), \ldots, \min(-U_n(\pi_n))\}. \tag{6}$$

To solve Pareto optimality, a classical method is to convert it into a single-objective optimization problem, that is, to obtain a point on the Pareto front in each simulation. There are some representative methods at present, such as Tchybeshev, weighted sum, perturbation, geometric mean, min–max, goal programming, and physical programming [79]. In contrast, evolutionary computation, which can obtain a large number of Pareto solutions after one simulation, is the current mainstream method [80].

Dynamic cooperative games: The dynamic cooperative games refer to the games that the participants take alternate actions to achieve a common goal. In general, we consider dynamic cooperative games from the perspective of Markov/stochastic games [17], which generalize Markov decision processes to multiple interacting decision-makers.

Multi-agent reinforcement learning (MARL) has made a great success in cooperative Markov games, especially when modeling and analysis for games are difficult. A lot of works exist in the fields of smart grid, traffic light control, home energy management and smart factories [81]–[83]. In cooperative MARL, each agent gets a shared reward, the cooperative Markov games can be modeled as maximizing the accumulation of rewards, and the Pareto optimal solutions are reduced to a unique optimal solution. The generation of the optimal strategy mainly relies on the trial and error in MARL progress. This section provides a brief survey on deep cooperative MARL, where the function relations are obtained by training neural networks.

One naive approach to solving an MARL problem is to train all agents using a central controller. In this structure, each agent needs a real-time data transmission with the central controller, and each action of the agents is strictly implemented under the requirements of the central controller. However, in most cases, communication is expensive, and it is difficult to achieve a real-time communication. In addition, as the number of agents increases, the neural network used by deep RL will also increase greatly, which has high requirements for the capacity of the central controller. Another naive approach to solving an MARL problem is to train each agent with individual RL [84]. This method has the following good characteristics compared with the centralized method:

1) Do not need a real-time data transmission and the decisions are generated by each agent itself;

2) No large-scale problems and neural networks are scattered across agents.

The strategy learning of each agent affects the strategy learning of other agents. For a single agent, if other agents are regarded as part of the environment, this environment is equivalent to a dynamic environment. The environment will change due to the actions of other agents, not just over time. This decentralized decision making can be modeled as decentralized partially observable Markov decision processes (Dec-POMDPs) [85], where action selection is under uncertainty and incomplete knowledge. In Dec-POMDPs, the training stability of single agent RL is poor, which directly hinders the extension to MARL. Fig. 3 shows the structure transformation of three kinds of multi-agent interactions in centralized training and decentralized execution (CTDE) framework, where Fig. 3(a) represents the use of a centralized controller to control all agents; Fig. 3(b) represents the realization of multi-agent control with limited communication; Fig. 3(c) represents controlling each agent through individual controller [81].

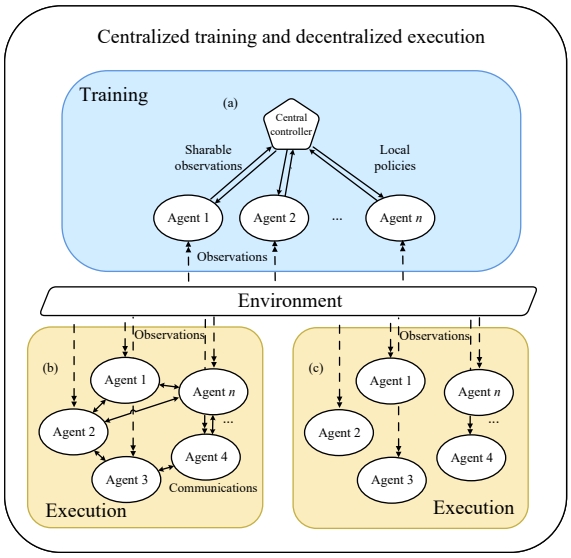

Fig. 3. Transformation of three representative information structures in centralized training and decentralized execution framework. (a) Training process: Centralized setting with significant communication costs; (b) Execution process: Decentralized setting with networked agents based on limited communication bandwidth; (c) Execution process: Fully decentralized setting without communication.

We will discuss cooperative MARL briefly through three main research areas:

1) Learning a communication protocol [86];

2) Learning a reward decomposition protocol [87];

3) Learning a critic network guaranteed stability training with global observation [88].

They all provide a reasonable breakthrough point for solving Dec-POMDPs problems.

Communication protocol learning: The cooperation in limited communication bandwidth depends on a communication protocol to coordinate each agent. One appropriate method to solve this question is to regard the communication

as an action of each agent, that is, communication-action. In this case, each agent needs to decide what message is transmitted, which agent to transmit, and what the message received means. As shown in Fig. 3(b), each robot has the ability to communicate with other robots. Using deep RL to realize the learning of communication protocols is first considered in [86]. At each time instant, agents need to simultaneously determine an environmental-action and a communication-action. The environmental-action will affect environmental transfer and the communicate-action will affect the environmental-action selection in other agents. As discussed before, Dec-POMDPs significantly affects the training stability of MARL algorithm. Agents can use the information of other agents to stabilize individual RL training. Moreover, they focus on the setting with CTDE framework. There is no limit to communication during training, but communication is only via the limited-bandwidth channel during execution (testing).

After that, more efficient communication strategies have become an important part of communication protocol learning. It is found in [89] that agents communicating by continues symbols has outperformed communicating with discrete symbols, because the information corresponding to continuous communication is differentiable and it can be trained efficiently with back-propagation. The lousy interaction effect between agents with continuous communication is considered in [9], where the received message may be harmful to training efficiency in some cases, and a gating mechanism allowing agents to block their communication is proposed. Besides, they think that it is difficult for each agent to know its individual contribution with a global reward, especially on large-scale tasks and it is proposed to train with its individualized reward. Similarly to [9], the lousy interaction effect question is also considered in [90] and the authors propose a targeted communication architecture for MARL. This targeted communication architecture is realized through an attention mechanism. The sender broadcasts a key that encodes the information to be transmitted, and then the receiver measures the correlation of information through this key, so as to determine whether a communication connection is established. They also suggest a multi-round communication structure, in which agents perform multi-round communication rather than just one round before taking action. Relatively special, it is believed that learning communication-actions will increase exploration space in [91], thereby reducing learning efficiency. In this way, a centralized RL neural network without communication is trained, then the decentralized executable neural network with communication is derived from the trained centralized RL neural network via policy distillation.

In cooperative MARL, communication can make up for the lack of individual information caused by limited observation, and stabilize the training. Unlike using a centralized controller, communication protocol learning is mainly considered in a limited communication capacity scenario. Agents learn communication protocols and cooperation strategies at the same time. However, in some scenarios, agents have no communication ability, the trained model obtained by the following two methods can be used in non-communication scenarios.

Reward decomposition protocol learning: In cooperative MARL, personalized reward mechanisms can achieve remarkable improvement in performance. However, it is very complex to design individual rewards for each agent, and it is difficult to ensure that individual rewards can achieve optimal cooperation. A more general solution is considered in [87] to let agents learn such reward decomposition protocol and introduce an additive value-decomposition network (VDN) over individual agents. VDN assumes that the joint action value function Q_{tot} can be additively decomposed into each action value function Q_i of n agents, $\forall i \in 1,\ldots,n$, i.e.,

$$Q_{tot} = \sum_{i=1}^{n} Q_i. \tag{7}$$

VDN provides the prototype of the value decomposition method. It is worth mentioning that Q_{tot} is the accumulation of global reward, and the decomposition of Q_{tot} is also equivalent to the decomposition of reward. Since then, a more theoretical exposition of the value decomposition method is made in [92] and the authors further propose the individual-global-max (IGM) condition. Suppose that τ^i, u^i denote all historical actions and observations respectively of agent i while $\boldsymbol{\tau}$, $\boldsymbol{u}$ represent the historical actions and observations respectively of all agents $1 \sim n$, IGM condition can be formulated as

$$\arg\max_{\boldsymbol{u}} Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) = \begin{pmatrix} \arg\max_{u^1} Q_1\left(\tau^1, u^1\right) \\ \vdots \\ \arg\max_{u^n} Q_n\left(\tau^n, u^n\right) \end{pmatrix}. \tag{8}$$

IGM describes such a relationship that the combination formed by the optimal actions of each agent is also the optimal action combination on the centralized controller. In this case, when each agent finds the local optimal action, and the global optimal action combination is found. The additional construction in VDN just satisfies such an IGM condition. Specifically, each agent's Q is constructed through a network, and these networks are added to form the global network of Q_{tot}. Under the CTDE framework, VDN trains the global network to realize the allocation of the total value of Q_{tot} automatically. However, VDN severely limits the complexity and learning ability of the centralized action value network. Based on VDN, a based on VDN, a new structure named QMIX [92] is developed, which is structure is developed, which is extended to a monotonicity constraint on the relationship between Q_{tot} and Q_i, i.e.,

$$\frac{\partial Q_{tot}}{\partial Q_i} \geq 0, \quad \forall i \in \{1,\ldots,n\}. \tag{9}$$

QMIX combines each Q_i into Q_{tot} through a mix network not as a simple sum as in VDN, and the mix network is restricted to have positive weights so as to achieve a monotonous constraint. Moreover, it can easily introduce the global information into the mix network to further enhance the global network.

However, the monotonic constraint is also too strict (the monotonic constraint is a sufficient condition for the IGM condition). Much work has since focused on how to relax constraints between Q_{tot} and Q_i (make the constraints between Q_{tot} and Q_i be a necessary and sufficient condition for the IGM condition). QTRAN structure [93] guarantees a more general decomposition than QMIX and VDN with a complex loss design. Weighted-QMIX (WQMIX) structure [94] provides greater learning rate for better action combinations to achieve preference for optimal action combinations. Unlike QTRAN and WQMIX realizing IGM conditions through approximation, QPLEX structure [95] realizes the strict construction of IGM conditions through a duplex dueling architecture. Qatten structure [96] introduces a multi-head attention mechanism based on QMIX to improve the approximation ability of Q_{tot}. The previous work of value decomposition is mainly based on the deep Q-learning (QL) structure, in [97], the value-decomposition methods are extended to the actor-critic framework.

How to decompose the reward generated by the grand alliance to each participant has always been a hot topic in the cooperative games. Traditional methods are inspired by marginal contribution like the Shpaley value [98]. Here, the reward decomposition protocol is obtained by neural network based on IGM conditions. However, most of the works are constructed on Q-learning (QL), and can not solve the problems of cooperation with continuous action space.

Global observation critic network learning: Unlike the reward decomposition protocol learning method, the global observation critic network learning method based on actor-critic algorithms (AC) is suitable for solving the cooperation problem of continuous action space. The MARL instability problem is formulated as a probabilistic form

$$P(s' \mid s,a,\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_n) \neq P\left(s' \mid s,a,\boldsymbol{\pi}'_1,\ldots,\boldsymbol{\pi}'_n\right),\ \ \forall \boldsymbol{\pi}_i \neq \boldsymbol{\pi}'_i \tag{10}$$

where s, s' represent the current state and the next state respectively in MDPs (Markov decision processes), $\{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\ldots,\boldsymbol{\pi}_n\}$, $\{\boldsymbol{\pi}'_1,\boldsymbol{\pi}'_2,\ldots,\boldsymbol{\pi}'_n\}$ represents two different strategy sets of n agents, a represent one of the action to be done in the current state. This formula depicts the fact that state transition probability is uncertain when other agents' strategies are not known. A sensational method [88] to stabilize the state transition probability is as following:

$$\begin{aligned} P(s' \mid s,a_1,\ldots,a_n,\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_n) &= P(s' \mid s,a_1,\ldots,a_n) \\ &= P\left(s' \mid s,a_1,\ldots,a_n,\boldsymbol{\pi}'_1,\ldots,\ \boldsymbol{\pi}'_n\right),\ \ \forall \boldsymbol{\pi}_i \neq \boldsymbol{\pi}'_i. \end{aligned} \tag{11}$$

Specifically, as long as we know the action of each agent, the state transition probability is determined. Similarly to most MARL methods, they take advantage of the CTDE framework that allows additional information to stabilize training as long as it is not used during the test. Considering that traditional Q-learning can not use different information in testing and training, they propose multi-agent deep deterministic policy gradient (MADDPG) algorithm based on actor-critic policy gradient methods instead. MADDPG simply achieves training stability by adding additional information to the critic network. Most of the research in global observation critic network learning is focused on MADDPG and its extensions. In [99], it is shown that MADDPG fails in some simple observable cooperative tasks. A recurrent multi-agent actor-critic architectures for message passing and movement then proposed to integrate RNN into MADDPG. Moreover, considering the similarity between MARL with natural language processing problem, the meaning of each agent's action (a word in a paragraph) depends on the entire system (the paragraph context). Then, the self-attention mechanism is integrated into MADDPG in [100]. Other researches consider improving algorithm efficiency from the training process, such as: modeling multi-agent policy transfer learning with option learning [101] and using a distributed prioritized experience replay [102].

The methods above extend the single agent RL algorithms to multi-agent RL in order to solve the dynamic cooperative games. As seen from Table IV, the current work of cooperative MARL tends to use a CTDE structure and RNN module to stabilize the learning process. The information structure transformations of the three methods are shown in Fig. 3. Most of the recent works on cooperative MARL use additional global information to stabilize the training process, which is based on the structure shown in Fig. 3(a), but the trained models have different landing forms in execution process. The first research area achieves cooperation by learning communication protocols, so the trained models need a limited communication bandwidth, which is based on the structure shown in Fig. 3(b). The other two research areas are not built in the context of communication, so the trained model can be applied in non-communication scenarios, which is based on the structure shown in Fig. 3(c).

B. Non-Cooperative Games

Non-cooperative games are suitable for analyzing the behaviors of players who believe that their own payoffs conflict with the payoffs of others in games. In non-cooperative games, individual payoffs are maximized and individual costs are minimized to achieve the Nash equilibrium (NE) [14]. For example, the optimal strategies of capture-the-flag differential games are obtained by geometric method in [19]. Unlike cooperative games, competitive behaviors between agents exist in non-cooperative games and agents cannot reach a binding agreement.

In multi-agent systems, we first define the NE, that is

$$\forall \pi_i \in \Pi_i,\ \ \forall s \in \mathcal{S},\ U_i\left(\pi_i^*,\pi_{-i}^*\right) \geq U_i\left(\pi_i,\pi_{-i}^*\right);\ \ \forall i \in \{1,\ldots,n\} \tag{12}$$

where π^* is a joint policy, $U_i(s)$ is the expected long-term payoff of agents i in state s and Π_i is the set of all possible strategies for agents i. Basically, it means that each agent would rather not change his/her strategies if he/she wants to obtain long-term payoffs.

Static non-cooperative games: In static games, players choose their strategies at the same time [103]. In this situation, each player does not know the choices of the others and he (or she) needs information to make decisions. The static non-

TABLE IV
SUMMARY OF COOPERATIVE MULTI-AGENT REINFORCEMENT LEARNING METHOD

Year	Reference	Research area	QL/AC	Single agent RL benchmark	Communication	Centralized training and decentralized execution	Neural network module
1993	Tan [104]	IQL	QL				
2016	Foerster *et al.* [86]	COM	QL	DQN	√		RNN
2016	Sukhbaatar *et al.* [89]	COM			√		RNN
2017	Lowe *et al.* [88]	CC	AC	DDPG	√	√	
2018	Sunehag *et al.* [87]	VD	QL	DQN	√	√	RNN
2018	Singh *et al.* [9]	COM	AC	REINFORCE	√	√	RNN
2018	Rashid *et al.* [92]	VD	QL	DQN		√	RNN
2019	Das *et al.* [90]	COM	AC	REINFORCE	√	√	RNN
2019	Son *et al.* [93]	VD	QL	DQN		√	RNN
2020	Chen [91]	COM	AC	DDPG	√	√	RNN
2020	Rashid *et al.* [94]	VD	QL	DQN		√	RNN
2020	Yang [96]	VD	QL	DQN		√	RNN/Attention
2020	Wang *et al.* [99]	CC	AC	DDPG	√	√	RNN
2021	Liu *et al.* [100]	CC	AC	DDPG	√	√	RNN/Attention
2021	Su *et al.* [97]	VD	AC			√	RNN
2021	Wang *et al.* [95]	VD	QL	DQN		√	RNN/Attention

cooperative games can be divided into complete information games and incomplete information games according to the players' understanding of other players. In complete information games, each player has accurate information about other players' characteristics, strategy spaces and cost functions. Otherwise, it refers to incomplete information games. In addition, static non-cooperative games can be divided into complete information static games and incomplete information static games. The equilibrium concepts corresponding to the two kinds of games are NE and Bayesian Nash equilibrium (BNE).

Complete information static games are the most basic type of non-cooperative games, such as classic Prisoner's dilemma, Cournot games, Chicken games and so on. Static games with complete information can be seen in many different situations, such as auction bids, network security, transport efficiency and so on. With the rapid development of the network, network security [105] has attracted more and more attention. It is different from the general research on information security based on the relationship between attackers and defenders. Aiming at the strategies selection of open testing service for medium-sized software vendors, the authors propose a research method based on the static games model of software vendor, white and black with complete information. In addition, static games with complete information also play an important role in transportation. For example, collaborative control of connected autonomous vehicles has the potential to significantly reduce negative impacts on the environment while improving driving safety and traffic efficiency [106], [107].

Incomplete information static games are also called Bayesian games. In incomplete information games [108], at least one player cannot determine the cost function of another player. The standard expressions of n-player static Bayesian games are as follows: The action spaces of the participants $A_1,\ldots,A_n$ and their type spaces $T_1,\ldots,T_n$, their inferences $p_1,\ldots,p_n$, and their cost functions $u_1,\ldots,u_n$. In addition, the type of participant i determines the benefit function $u_i(a_1,a_2,\ldots,a_n;t)$ of participant i. The inference $p_i(t_{-i}|t_i)$ of player i describes the uncertainty of possible type t_{-i} for other $n-1$ participants given i's own type t_i, where $t_{-i}=\{t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n\}$. The games can be denoted as $G=(A_1,\ldots,A_n;T_1,\ldots,T_n;p_1,\ldots,p_n;u_1,\ldots,u_n)$. Incomplete information makes game analysis complicated, so many researchers use Harsanyi transformation to transform incomplete information games into complete but imperfect information games.

In a static Bayesian games $G=(A_1,\ldots,A_n;T_1,\ldots,T_n;p_1,\ldots,p_n;u_1,\ldots,u_n)$, one strategy of player i is a function $s_i(t_i)$. Strategic combination $s^*=\{s_1^*,s_2^*,\ldots,s_n^*\}$ is a pure strategic BNE, if for each player i and for each t_i in the type set T_i of i, s^* satisfies

$$\max_{a_i\in A_i}\sum_{t_{-i}}\{u_i[s_1^*(t_1),\ldots,s_{i-1}^*(t_{i-1}),a_i,s_{i+1}^*(t_{i+1}),\ldots,s_n^*(t_n);t_i]p_i(t_{-i}|t_i)\}. \tag{13}$$

When a player's strategic combination in a static Bayesian games is a BNE, no player is willing to change his strategies, even if the change involves only one action of one type.

In recent years, researchers have carried out many studies on static games with incomplete information. With the increasing popularity of electric vehicles and the lack of poor battery life, an electricity transaction scheme based on Bayesian games pricing in the blockchain-supported internet of vehicles is proposed, and users' approximate satisfactions are verified in [109]. When the Markov games with incomplete information generated by the defense strategies of

moving target are in a certain network state, they can be regarded as static games with incomplete information. In view of the problem that network defense based on incomplete information games ignores the types of defenders, which leads to an improper selection of defense strategies, an active defense strategy selection based on static Bayesian game theory is proposed in [110] and the authors construct a static Bayesian game model. 5G networks communication with incomplete information games mechanism based on static games repeated model is studied in [111], and the proposed mechanism provided a more realistic and universal model for resource allocation of communication between multi-unit devices. A memory-based game theory defense method under incomplete information is mentioned in [112].

Dynamic non-cooperative games: From the perspective of cooperative games, the MARL methods mentioned above are also applicable to dynamic non-cooperative games [9], [88]. Specifically, the cooperative agents in the non-cooperative game form alliances. The agents within each alliance cooperate in the game, and use the same cost function and MARL method, while the agents in different alliances perform non-cooperative behaviors, and use different cost functions or even different MARL methods. Finally, the best strategy for non-cooperative games can be obtained.

From the perspective of non-cooperative games, it is appropriate for analyzing the behaviors of agents whose payoffs completely conflict with the payoffs of others in the system. Therefore, the dynamics of multi-agent systems are modeled and the cost functions are characterized, and then the optimal strategy is obtained by minimizing the cost function and maximizing the reward function. Similarly to static non-cooperative games, this survey also introduces two aspects of dynamic non-cooperative games: Complete information and incomplete information. The equilibrium concepts corresponding to these two refer to subgame perfect Nash equilibrium and perfect Bayesian Nash equilibrium (PBNE), respectively.

In complete dynamic non-cooperative games, take pursuit-evasion games [18] and reach-avoid games [113] as examples. In the reach-avoid games, the attackers need to reach the target without being intercepted by the defenders, whereas the defenders need to capture the attackers before they reach the target. Moreover, capture-the-flag games are more complex than reach-avoid games in game settings. In capture-the-flag games [19], [20], two opposing teams control their respective areas and protect the flags in that area. Capture-the-flag games consist of two phases, pre-capture and post-capture, and each phase contains a number of potential competitive objectives. The victory condition is that agents enter the opponent's area to capture the flag, and then return to their own area without being captured by the opponent.

In capture-the-flag games, players must consider not only the capture region but also the return region when choosing their actions. The basic settings of capture-the-flag games can be seen in Fig. 4. Numerical solutions to the Hamilton-Jacobi-Isaacs (HJI) equations are used to describe the initial conditions and strategies for each agent to win [114]. More

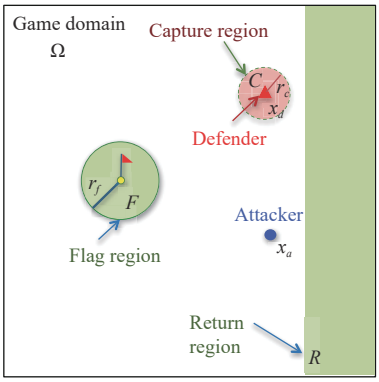

Fig. 4. The basic setting of capture-the-flag games [114].

generally, the winning strategies of the opponents in capture-the-flag games can be viewed as a solution to a zero-sum differential game. The goal of each team can be expressed as a value function that can be maximized or minimized by changing their strategies. With some processing, the value function can be calculated from the solution of the HJI equation. The payoff function of each team can be expressed as

$$J(u(t)) = \phi(x(t)) + \int_0^T L(x,u,t)dt \tag{14}$$

where $x(t)$ is the state of system. The value of the game is

$$V(\boldsymbol{x}_0) = \min\max J(u(t);\boldsymbol{x}_0). \tag{15}$$

The solution to the HJI equation can be obtained through the method of [115]. In the equation, the Hamiltonian can be expressed as

$$H(x,\lambda,u,t) = \lambda^T(t)f(x,u,t) + L(x,u,t) \tag{16}$$

where λ is the costate vector and u is the optimal control.

Moreover, there have been a considerable amount of studies on capture-the-flag games up to now. In addition to solving the optimal strategies by the HJI equation, there are many researchers using geometric methods to consider the optimal problems of capture-the-flag games. For example, the two-stage synergistic optimal strategy and winning region for capture-the-flag games are proposed in [19]. The optimal capture point can be obtained by Apollonius circle. In [116], for the two-speed ratios, an attack zone method based on the Voronoi diagram and Apollonius circle is proposed to construct barriers analytically.

In addition, dynamic games are often formulated in an extensive form [14], described as game tree, which defines the order, information, alternatives and outcomes of dynamic games. In other words, extensive-form games generalize normal-form games by modeling sequential and simultaneous moves, as well as private information. The stable state of dynamic non-cooperative games can be solved by the NE and backward induction, named subgame perfect NE [117]. At each stage of the game, backward induction determines the player's optimal strategy for the last step, i.e., NE. Then, taking the last player's move as a given, the best move for the penultimate player is determined. This process continues backwarding until the best action at each point of the games is determined. In other words, Nash backward induction [22] is

used to determine the NE of each subgame of the original games. However, in practical games, such as Texas Hold'em and StarCraft II, the information of the games is often imperfect/incomplete. Therefore, the results of Nash backward induction often fail to conduct the NE [118].

In incomplete dynamic non-cooperative games, as mentioned in previous content, with the help of the Harsanyi transformation, the incomplete information games are able to be converted to an equivalent complete but imperfect information games. In that way, we only need to consider the case of imperfect dynamic non-cooperative games.

Finding an approximate NE solution for imperfect information games has received a great deal of attention recently [119]. At the basis of regret matching [23], no-regret learning [120] has merged as a useful tool to solve the issue. By means of minimizing counterfactual regret, the overall regret is decreased, so that the average strategy is able to approach an NE. Counterfactual regret minimization (CFR) [121] is the most popular and effective family of algorithms for finding the PBNE in imperfect information games. Specifically speaking, in CFR games, counterfactual regret value and cost function are computed and decomposed into every information set. Let σ_i^t be the strategies of the player i on round t. The counterfactual value and the instantaneous regret for an action a in information set I on iteration t are as follows respectively

$$u_i(I,\sigma)=\sum_{z\in Z_I}\pi_{-i}^{\sigma}(z[I])\pi^{\sigma}(z[I],z)u_i(z)$$

$$r^t(I,a)=v_i\left(I,\sigma_{I\to a}^t\right)-u_i\left(I,\sigma^t\right). \tag{17}$$

Then, players use regret matching to compute the strategies of next iteration at the basis of the sum of counterfactual regret values. On the iteration $T+1$, player i selects action $a\in A(I)$ according to probability

$$\sigma_i^{T+1}(I,a)=\begin{cases}\dfrac{R_i^{T,+}(I,a)}{\sum_{a\in A(I)}R_i^{T,+}(I,a)}, & \text{if } \displaystyle\sum_{a\in A(I)}R_i^{T,+}(I,a)>0\\ \dfrac{1}{|A(I)|}, & \text{otherwise}\end{cases} \tag{18}$$

where $R_+^T(I,a)=\max\left\{R^T(I,a),0\right\}$ to make the accumulated regret positive. Therefore, the average strategy over all iterations converses to the NE of two-player zero sum games.

However, original CFR only works for simplified state space, action space and the converge efficiency demands improvement. Sampling-based CFR [122], [123] and deep learning-based CFR [123]–[125] are proposed to solve the large-scale games effectively. It is worth noting that sampling-based CFR uses two large tabular-based memories M_S and M_R to record the average strategy and cumulative regret for all information sets, while in deep learning-based CFR, two deep neural networks are utilized to learn M_S and M_R, respectively. The difference of sampling-based CFR only traverses a subset of the game tree. Discounted CFR [122] is proposed to utilize regret matching to reduce calculation time and storage memory; Monte Carlo is combined with standard or vanilla CFR to compute the PBNE strategies in [123]. In another way, deep learning-based CFR relies on two independent deep neural networks to represent the regret value and strategies while approximating the weighted average NE strategy. By introducing deep neural networks to CFR, double neural CFR [123] and deep CFR [124] are proposed to match the performance of CFR algorithms. On the basis of deep CFR, single deep CFR [125] simplifies the neural approximation by the obtained value networks, which directly adjust the NE strategy. CFR-MIX [126], built on double neural CFR, utilizes strategy representation to represent a joint action strategy, so as to represent cumulative regret clearly.

In a word, regret matching mainly focuses on every information set in a game tree, modeling each relevant player, and iteratively solving the NE. That is to say, each player can only observe the received payoffs and choose a better reply with respect to the average realized payoff. There is no need to know the number of players and the cost functions in the games. On the contrary, with the perspective of modeling the games from a global, fictitious play (FP) [127] is a good choice. Similarly to regret matching, FP also finds or approaches the NE in an iterative manner. Specifically, in FP, each player can observe other players' actions and choose the best response to his belief. It is necessary for players to acquire the knowledge of cost functions.

FP is a method of learning the NE in normal-form games by playing the game repeatedly and choosing the best response according to each other's average behavior. Fictitious self-play (FSP) [24] extends FP to extensive-form games with sampling-based machine learning approach. Therefore, the curse of dimensionality is broken. Combining neural network function approximation with FSP, neural fictitious self-play (NFSP) [128] is proposed. NFSP consists of two neural networks, one is the best response policy network, learning an approximate best response to the history behaviors of other players by RL; the other is the average policy network, trained from memorized behaviors of the best response policy by supervised learning. It is worth noting that NFSP is the first end-to-end RL method to approach to the NE in imperfect information games without any prior knowledge. The strategies are iteratively updated to approximate the PBNE

$$\Pi_{t+1}^i\in(1-\alpha_{t+1})\Pi_t^i+\alpha_{t+1}B_{\epsilon_t}^i\left(\Pi_t^{-i}\right) \tag{19}$$

where B_ϵ^i is the ϵ-best response of player i, Π_{t+1}^i is the average strategy of player i in iteration $t+1$ and α_{t+1} is a mix probability of the best response and the average strategy.

In order to improve the stability and speed of RL training, asynchronous NFSP [129] is proposed, which shares deep Q network (DQN) and supervises learning network between players in parallel, so as to accumulate and calculate gradients. In that way, the data storage memory is smaller than NFSP. With the help of policy gradient method, self-play pretrains the model to improve the scale of NFSP. For the improvement of RL progress, the authors propose a situation evaluation reward in each round and a future reward. In order to improve efficiency in large-scale extensive-form games, network security games are combined with NFSP in [130], enabling NFSP with high-level actions. The best response policy

network is based on deep recurrent neural network, designed for natural languages domain, to define the noticeable legal actions of network security games. For the average policy network, the proposed metric-based few-shot learning [131] utilizes information contained in graphs to transfer actions and states between metrics.

Large-scale dynamic non-cooperative games: In large-scale dynamic non-cooperative multi-agent systems, it is unrealistic for agents to be able to access the real-time state information of all other agents, which leads to challenges brought by imperfect information games. Mean field games (MFG) theory provides a promising solution to large-scale dynamic non-cooperative games [27]. In MFGs, a set of decentralized control laws are obtained for a single agents, where each agent only needs its own state information, giving decentralized ϵ-NE for finite N agents with ϵ approaching zero as $N \to \infty$ [132], [133]. In MFGs, agents may be coupled by their cost functions, dynamics and even observation processes. A general model for each agent can be described by

$$dx_i(t) = \frac{1}{N}\sum_{j=1}^{N} f_i(x_i(t), u_i(t), x_j(t))dt + \sigma dw_i(t) \tag{20}$$

where for agent i, $x_i(t) \in \mathbb{R}^n$ is the state, $u_i(t) \in \mathbb{R}^m$ the control input, $w_i(t) \in \mathbb{R}^p$ the standard Wiener motion, and f_i a general nonlinear function. The objective of each agent is to minimize its own cost function described by

$$J_i(u_i, u_{-i}) = \mathbb{E}\int_0^T \frac{1}{N}\sum_{j=1}^{N} g_i(t, x_i(t), u_i(t), x_j(t))dt \tag{21}$$

where u_{-i} denotes $(u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_N)$ and g_i is a general nonlinear function. When f_i in (20) is linear and g_i in (21) is quadratic, analytical ϵ-NE can be obtained for MFGs [133].

In MFGs, an important concept is the Nash certainty equivalence (NCE) principle introduced in [132], which provides a way to obtain a specific estimate of the mean-field term, thus providing an approximated PBNE.

Since MFG theory provides a powerful methodology for obtaining decentralized control strategies for large-scale dynamic non-cooperative multi-agent systems, it has received considerable attention since the pioneering work in [132], [133]. By introducing a major agent, MFGs for major-minor agent systems are studied in [134]. Risk-sensitive and robust performance indexes are considered in [135]. There are also a few works focusing on discrete-time settings, such as [136].

Applications of dynamic non-cooperative games: The difficulty level of dynamic non-cooperative games is ranked from easy to difficult as follows:

I) Games with complete information, such as AlphaGo and AlphaZero;

II) Games with incomplete information and a single goal, such as Texas Hold'em;

III) Games with incomplete information and complex goals, such as StarCraft II.

The examples of these three categories lie in Fig. 5.

In Category I, we take Go as an example. Due to the challenges of huge search space and sparse reward, AlphaGo [137] utilizes supervised learning from human experts and the self-play RL framework to train the deep neural networks, and it is the first time that AI defeats a professional in Go games. The self-play RL framework is designed to generate new datasets and directly approximate the optimal value function. Then, AlphaGo Zero [138] improves the Go games by using self-play RL from scratch, without human data and any domain knowledge other than the basic rules. After that, AlphaZero [28] is put forward, which has succeeded in generalizing the structure from Go games to other zero-sum games of complete and perfect information, such as chess and shogi by means of a general self-play RL algorithm. MuZero [139] extends AlphaZero to a wider situation, including a single agent and non-zero rewards for intermediate time steps.

In Category II, we take Texas Hold'em as an example. Libratus, put forward by Brown and Sandhoim in [140], focuses on three application-independent algorithms and defeats four top human specialists in no-limit two-player Texas Hold'em. The first algorithm proposes the blueprint strategy, with the help of Monte Carlo CFR [123], that simplifies the action space by skipping suboptimal actions for repeated iterations in the game. The second algorithm performs the real-time calculation of the sub-game strategies based on the current game state without conflicting with the blueprint strategy. The third algorithm aims at improving the blueprint strategy by absorbing the actual actions of the opponents. Then, Pluribus [141] has conducted a landmark extension from the scale of two-player Texas Hold'em to six-player mode. In order to improve the efficiency of Monte Carlo CFR in Pluribus, during the training stage of [29], linear weighted discounts are introduced in the early stages, and negative regret behaviors are strategically pruned in the later stages.

In Category III, we take StarCraft II as an example. Alphastar [30] is designed for StarCraft II, which is composed of four leagues, named main agents, main exploiters, league exploiters and past players, respectively. Main agents are trained to defeat all players, main explorers are designed to find the weakness of main agents on the basis of human information, and league exploiters aim at finding the weakness of the system, training with the historical data stored by past players. Prioritized fictitious self-play (PFSP) is proposed to calculate the best response to the mixed previous strategies, so that main agents will give priority to players with low winning rates. The performance of Alphastar can reach the Grandmaster level in all three StarCraft II races. After that, StarCraft Commander [142] optimizes the Alphastar in computational resources and robustness, which is able to acquire most of the Alphastar performance with fewer replays and datasets. In addition, during the RL training stage, branching agents are proposed to improve training efficiency and balance the strength and diversity of strategies.

As seen from Table V, the current work of dynamic non-cooperative games mainly focuses on discrete action space with complete information in the situation of perfect competition, more difficult game types and complex situations remain to be considered.

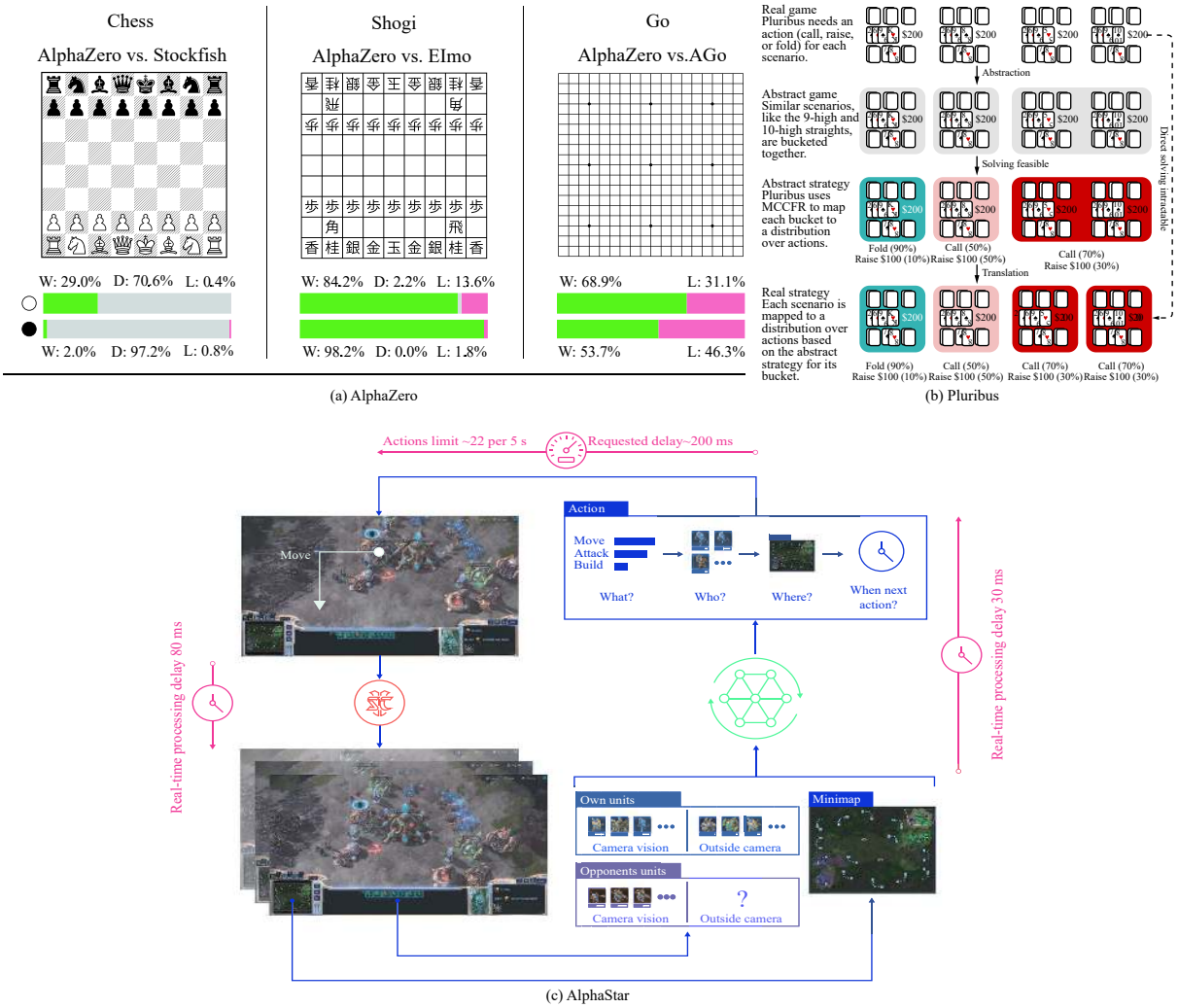

Fig. 5. Difficulty category examples of non-cooperative games. (a) AlphaZero belongs to Category I [28]; (b) Pluribus belongs to Category II [29]; (c) AlphaStar belongs to Category III [30].

TABLE V
SUMMARY OF DYNAMIC NON-COOPERATIVE GAMES

Year	Reference	Perfect competition	Complete/Incomplete games	Algorithm of solving the NE	Application
2018	Silver *et al.* [28]	√	Complete	Self-play	chess and shogi
2019	Brown and Sandholm [122]	√	Complete	Discounted CFR	HUNL poker
2019	Steinberger [125]	√	Complete	Single Deep CFR	5-Flop Hold'em poker
2019	Brown and Sandholm [141]		Incomplete	Monte Carlo CFR	Texas Hold'em poker
2019	Vinyals *et al.* [30]		Incomplete	Prioritized FSP	StarCraft II
2020	Li *et al.* [123]	√	Complete	Monte Carlo CFR	Leduc Hold'em poker
2021	Wang *et al.* [142]		Incomplete	PFSP	StarCraft II
2021	Li *et al.* [126]		Incomplete	CFR-MIX	Goofspiel
2021	Chen *et al.* [143]	√	Complete	Optimization NFSP	OpenSpiel
2021	Xue *et al.* [130]	√	Complete	NSG-NFSP	Goofspiel

IV. FUTURE WORK

A. Cooperative Optimization

Non-convex: Distributed optimization for solving non-convex optimization problems has attracted considerable attention in many fields. In the current research, distributed algorithms have been developed to solve unconstrained/constrained non-convex DOO issues with undirected/directed graphs [51], [144]. It is worth applying the existing distributed optimization algorithms to non-convex optimization problems with theoretical guarantees, switching the goal from finding a globally optimal solution to finding a stationary point or a

local extremum. In that way, the polynomial dependence of the complexity of the first-order methods on the dimension of the problem and the desired accuracy can be obtained by polynomial approximation, thereby achieving convergence to the global optimal from the algorithm level.

Heterogeneity: Heterogeneity greatly hinders the integration in federated optimization, which contains only one central server integrating data from different clients. However, inspired by the existing research of data heterogeneity [145], multi-center is expected to be a solution to client heterogeneity. Specifically, different local central servers are used to integrate different types of clients. Then, according to the actual application scenarios, it is determined whether multiple local central servers need to be integrated into a global central server.

Security: From the perspective of privacy protection, the security of the cooperative optimization system is improved by differential privacy, information encryption and model aggregation. With the development of quantum computers, traditional encryption technologies will no longer be secure. Therefore, quantum encryption is expected to become the mainstream encryption technology for privacy protection in the process of cooperative optimization. Recently, there has been some applied research on quantum homomorphic encryption [146]. In the future, extending the quantum homomorphic encryption to cooperative optimization is a promising direction.

Lightweight network: A lightweight network scale is one way to ensure security and robustness. In addition to end-to-end training using learning methods, such as neural architecture search and network pruning, combining learning methods with analytical methods have been recently considered to reduce network complexity while maintaining model robustness and security. For example, the physics-informed neural network (PINN) [147] is able to approximate and solve complex partial differential equations using deep neural networks, RL, etc., while adding interpretability to the learning methods. In the future, it is an interesting idea to utilize PINN for optimization problems that are easy to model but difficult to solve.

B. Cooperative/Non-Cooperative Games

Since non-cooperative games may include some cooperative behaviors, cooperative and non-cooperative games have overlaps in challenges. Therefore, for future work, the ideas below are not limited to cooperative games or non-cooperative games alone.

Heterogeneity: For many-to-many heterogeneous cooperative games, such as matching games [148], the current literature [149] mainly focuses on multi-connectivity, multi-channel and multi-radio conditions. An interesting future direction is to increase the consideration of heterogeneous information from system dynamics, such as acceleration and jerk constraints, in order to achieve more adequate allocation of limited resources among heterogeneous agents.

Interpretability: For ante-hoc interpretability, it makes the model itself interpretable. In Bayesian games, due to the assumption of conditional independence, existing work [21] transforms the decision-making process of the game models into probability calculations. In addition, the attention mechanism makes the model well interpretable by focusing on the objects of interest in games. In future, introducing transformers to allocate limited resources and assist the decision-making process of the games is a promising direction. For post-hoc interpretability, it uses interpretability techniques to interpret the model. In cooperative games, few researchers have explored why the CTDE structure can achieve improved results in various domains. A meaningful task is to prove that the iterative process of CTDE is a contraction mapping, and then prove that the CTDE algorithm converges to the optimal strategy through Banach's fixed point theorem. In addition, PINN endows deep neural networks with interpretability that is not available on the basis of mathematical knowledge and physical laws. Encouragingly, games can provide interpretable support for research in other fields. Inspired by Eigengame [150], it is an attractive idea to abstract the goals and construct game models for non-game problems and choose the corresponding optimal strategy for some of the agents involved. Therefore, the analysis method is changed and the interpretability is improved.

Large scale: Considering the scale of agents in multi-agent cooperative/competitive systems, the action combination space increases exponentially. It increases the difficulty of finding the NE or globally optimal solutions. By means of hierarchical networks, such as hierarchical RL [151] and hierarchical graph attention network [152], the action combination space is dispersed to each layer. Then, the cooperative/competitive relationship between agents can be described. Therefore, the multi-dimensional game equilibrium can be accurately obtained through end-to-end training. In addition, since each agent needs to know global information in the training process of CTDE algorithms, with the increase of the number of agents, the communication cost will become a challenge. Inspired by federated learning, which is widely used to optimize of a large number of client models, MARL combined with federated learning is expected to reduce the communication costs in the future. Solutions to non-cooperative games will be upgraded to MFGs as agents scale from massive to infinite. The existing literature focuses on, for example, mean-field MARL [153], major-minor agent systems [134], and risk-sensitive and robust performance [135]. A feasible direction for MFG in the future is to explore MFG with a finite number of agents.

Few-shot: Games in multiple fields, such as economy, public opinion and energy, lack consideration of information factors in an open environment. Due to the low-texture and high-dynamic characteristics of the open environment, the environmental data samples are few and difficult to collect. Therefore, there are often incomplete/imperfect information in games, such as deceptive information, uncertain interaction information, asymmetry network, etc. Incomplete/imperfect information makes the agents unable to identify cooperative/competitive intentions, and it is difficult to predict the actions of the agent, thereby affecting the formulation of game strategies. The existing literature [154] utilizes few-shot learning to improve the accuracy of perception and decision-

making of game model in real scenes. On this basis, inverse RL based MFGs [155] can be constructed for handling intent identification and the NE solution, and evolutionary game theory [156] is used to predict the changes in agent actions in real time. In the future, it will be a popular idea to solve the few-shot problems in games from the perspective of transferability, i.e., game setting transfer, framework transfer and strategy transfer. From an algorithmic perspective, meta learning and its variations are good choices for transferability, based on the latest meta-learning ideas, i.e., graph neural network meta learning [157] and domain-augmented meta learning [158], covering migration from communication topology transfer to game environment transfer. From an environmental perspective, universal decision intelligence platform, such as OpenDILab [159], can integrate existing decision-making AI algorithms to achieve transferability and scalability of various computing scales.

C. Applications

In the field of integrated energy networks, consider using distributed RL to design multi-dimensional cost functions for different types of energy sources and energy-consuming devices with different characteristics. In addition, for lack of prior knowledge and multiple uncertain optimization problems, federated optimization is a feasible way to achieve real-time scheduling and enhance the privacy protection of information from various data sources. Furthermore, redesigning the control inputs is an idea to ensure linear convergence rates when system dynamics are incorporated into the optimization objective.

In the field of information security, from the perspective of data integrity and availability, the problems of network attack in non-cooperative games are considered. It is suggested to break through the limitations of incomplete/imperfect information conditions by means of intention recognition and aggressive behavior prediction, and utilize the evolutionary game theory to solve non-cooperative games with cyber-attacks; From the perspective of data security, the privacy protection problems are solved by constructing a federated optimization framework that does not directly use client data, but trains machine learning algorithms.

In the field of biomedicine, in terms of molecular synthesis, considering the optimization of the predictive molecular model, under the conditions of multi-objective and multi-constraint, a cooperative game model is constructed to obtain the target macromolecular structure with maximum utility. In order to solve the bottleneck of large-scale data transmission in CTDE, it is an feasible idea to consider federated MARL, which can ensure the security and transmission efficiency of the macromolecule database.

V. Conclusions

In this survey, we analyze the cooperative and competitive multi-agent systems, focusing on the progress of optimization and decision-making. Starting with multi-agent cooperative optimization, we survey the distributed optimization and federated optimization focusing on online optimization and privacy protection. Then we consider the possible non-cooperative behaviors between agents during the optimization process. Therefore, the focuses of our survey range from cooperative optimization to cooperative/non-cooperative games. We survey the relevant articles on multi-agent cooperative/non-cooperative games from static and dynamic perspectives. It is worth noting that for static/dynamic non-cooperative games, we also summarize separately from two perspectives of complete/incomplete information. At last, we put forward future work on cooperative optimization, cooperative/non-cooperative games, and their applications.

References

[1] M. Mazouchi, M. B. Naghibi-Sistani, and S. K. H. Sani, "A novel distributed optimal adaptive control algorithm for nonlinear multi-agent differential graphical games," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 331–341, Jan. 2018.

[2] M. Q. Xue, Y. Tang, W. Ren, and F. Qian, "Practical output synchronization for asynchronously switched multi-agent systems with adaption to fast-switching perturbations," *Automatica*, vol. 116, p. 108917, Jun. 2020. DOI: 10.1016/j.automatica.2020.108917.

[3] X. Jin, Y. Shi, Y. Tang, and X. T. Wu, "Event-triggered attitude consensus with absolute and relative attitude measurements," *Automatica*, vol. 122, p. 109245, Dec. 2020. DOI: 10.1016/j.automatica.2020.109245.

[4] Y. Tang, J. Kurths, W. Lin, E. Ott, and L. Kocarev, "Introduction to Focus Issue: When machine learning meets complex systems: Networks, chaos, and nonlinear dynamics," *Chaos*, vol. 30, no. 6, p. 063151, Jun. 2020. DOI: 10.1063/5.0016505.

[5] X. Jin, Y. Shi, Y. Tang, H. Werner, and J. Kurths, "Event-triggered fixed-time attitude consensus with fixed and switching topologies," *IEEE Trans. Autom. Control*, 2021. DOI: 10.1109/TAC.2021.3108514.

[6] L. Ding, Q. L. Han, X. H. Ge, and X. M. Zhang, "An overview of recent advances in event-triggered consensus of multi-agent systems," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1110–1123, Apr. 2018.

[7] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 8, pp. 3152–3168, Aug. 2020.

[8] S. Mao, Z. W. Dong, P. Schultz, Y. Tang, K. Meng, Z. Y. Dong, and F. Qian, "A finite-time distributed optimization algorithm for economic dispatch in smart grids," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 51, no. 4, pp. 2068–2079, Apr. 2021.

[9] A. Singh, T. Jain, and S. Sukhbaatar, "Learning when to communicate at scale in multi-agent cooperative and competitive tasks," in *Proc. 7th Int. Conf. Learning Representations*, New Orleans, USA, 2019, pp. 1–16.

[10] J. S. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[11] Y. M. Wang, S. X. Wang, and L. Wu, "Distributed optimization approaches for emerging power systems operation: A review," *Electr. Power Syst. Res.*, vol. 144, pp. 127–135, Mar. 2017.

[12] L. Ding, L. Y. Wang, G. Y. Yin, W. X. Zheng, and Q. L. Han, "Distributed energy management for smart grids with an event-triggered communication scheme," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 5, pp. 1950–1961, Sep. 2019.

[13] J. Konečnỳ, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," arXiv preprint arXiv: 1511.03575, 2015.

[14] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. 2nd ed. Philadelphia, USA: SIAM, 1998.

[15] E. Semsar-Kazerooni and K. Khorasani, "A game theory approach to multi-agent team cooperation," in *Proc. American Control Conf.*, St. Louis, USA, 2009, pp. 4512–4518.

[16] T. L. Vincent and G. Leitmann, "Control-space properties of cooperative games," *J. Optim. Theory Appl.*, vol. 6, no. 2, pp. 91–113, Aug. 1970.

[17] L. S. Shapley, "Stochastic games," *Proc. Natl. Acad. Sci.*, vol. 39, no. 10, pp. 1095–1100, Oct. 1953.

[18] I. E. Weintraub, M. Pachter, and E. Garcia, "An introduction to pursuit-evasion differential games," in *Proc. American Control Conf.*, Denver, USA, 2020, pp. 1049–1066.

[19] Z. Zhou, J. H. Huang, J. P. Xu, and Y. Tang, "Two-phase jointly optimal strategies and winning regions of the capture-the-flag game," in *Proc. 47th Annu. Conf. IEEE Industrial Electronics Society*, Toronto, Canada, 2021, pp. 1–6.

[20] J. Wang, J. Huang, and Y. Tang, "Swarm intelligence capture-the-flag game with imperfect information based on deep reinforcement learning," *Sci. Sin. Technol.*, 2021. DOI: 10.1360/SST-2021-0382.

[21] S. Zamir, "Bayesian games: Games with incomplete information," in *Complex Social and Behavioral Systems*, M. Sotomayor, D. Pérez-Castrillo, F. Castiglione, Eds. New York, USA: Springer, 2020.

[22] D. Y. Sun, X. Huang, Y. H. Liu, and H. Zhong, "Predictable energy aware routing based on dynamic game theory in wireless sensor networks," *Comput. Electric. Eng.*, vol. 39, no. 6, pp. 1601–1608, Aug. 2013.

[23] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, Sep. 2000.

[24] J. Heinrich, M. Lanctot, and D. Silver, "Fictitious self-play in extensive-form games," in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 805–813.

[25] S. W. Wang, X. Jin, S. Mao, A. V. Vasilakos, and Y. Tang, "Model-free event-triggered optimal consensus control of multiple Euler-Lagrange systems via reinforcement learning," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 1, pp. 246–258, Jan.–Mar. 2021.

[26] C. Z. Zhang, J. R. Wang, G. G. Yen, C. Q. Zhao, Q. Y. Sun, Y. Tang, F. Qian, and J. Kurths, "When autonomous systems meet accuracy and transferability through AI: A survey," *Patterns*, vol. 1, no. 4, p. 100050, Jul. 2020. DOI: 10.1016/j.patter.2020.100050.

[27] H. Tembine, Q. Y. Zhu, and T. Başar, "Risk-sensitive mean-field games," *IEEE Trans. Autom. Control*, vol. 59, no. 4, pp. 835–850, Apr. 2014.

[28] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, Dec. 2018.

[29] A. Blair and A. Saffidine, "AI surpasses humans at six-player poker," *Science*, vol. 365, no. 6456, pp. 864–865, Aug. 2019.

[30] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Y. Wang, T. Pfaff, Y. H. Wu, R. Ring, D. Yogatama, D. Wünsch, K. Mckinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, Oct. 2019.

[31] T. Yang, X. L. Yi, J. F. Wu, Y. Yuan, D. Wu, Z. Y. Meng, Y. G. Hong, H. Wang, Z. L. Lin, and K. H. Johansson, "A survey of distributed optimization," *Ann. Rev. Control*, vol. 47, pp. 278–305, May 2019.

[32] M. Zhu, A. H. Anwar, Z. L. Wan, J. H. Cho, C. A. Kamhoua, and M. P. Singh, "A survey of defensive deception: Approaches using game theory and machine learning," *IEEE Commun. Surv. Tut.*, vol. 23, no. 4, pp. 2460–2493, Oct.–Dec. 2021.

[33] K. Sohrabi and H. Azgomi, "A survey on the combined use of optimization methods and game theory," *Arch. Comput. Methods Eng.*, vol. 27, no. 1, pp. 59–80, Jan. 2020.

[34] Q. J. Shi, C. He, H. Y. Chen, and L. G. Jiang, "Distributed wireless sensor network localization via sequential greedy optimization algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3328–3340, Jun. 2010.

[35] S. Mao, Y. Tang, Z. W. Dong, K. Meng, Z. Y. Dong, and F. Qian, "A privacy preserving distributed optimization algorithm for economic dispatch over time-varying directed networks," *IEEE Trans. Indust. Inf.*, vol. 17, no. 3, pp. 1689–1701, Mar. 2021.

[36] D. K. Molzahn, F. Dörfler, H. Sandberg, S. H. Low, S. Chakrabarti, R. Baldick, and J. Lavaei, "A survey of distributed optimization and control algorithms for electric power systems," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2941–2962, Nov. 2017.

[37] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 714–725, Mar. 2017.

[38] N. Eshraghi and B. Liang, "Distributed online optimization over a heterogeneous network with any-batch mirror descent," in *Proc. 37th Int. Conf. Machine Learning*, 2020, pp. 2933–2942.

[39] X. X. Li, X. L. Yi, and L. H. Xie, "Distributed online convex optimization with an aggregative variable," *IEEE Trans. Control Netw. Syst.*, 2021. DOI: 10.1109/TCNS.2021.3107480.

[40] J. Y. Li, C. Y. Gu, Z. Y. Wu, and T. W. Huang, "Online learning algorithm for distributed convex optimization with time-varying coupled constraints and bandit feedback," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1009–1020, Feb. 2022.

[41] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed estimation of dynamic parameters: Regret analysis," in *Proc. American Control Conf.*, Boston, USA, 2016, pp. 1066–1071.

[42] M. Akbari, B. Gharesifard, and T. Linder, "Distributed online convex optimization on time-varying directed graphs," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 3, pp. 417–428, Sep. 2017.

[43] Y. Zhang, R. J. Ravier, M. M. Zavlanos, and V. Tarokh, "A distributed online convex optimization algorithm with improved dynamic regret," in *Proc. IEEE 58th Conf. Decision and Control*, Nice, France, 2019, pp. 2449–2454.

[44] S. M. Fosson, "Centralized and distributed online learning for sparse time-varying optimization," *IEEE Trans. Autom. Control*, vol. 66, no. 6, pp. 2542–2557, Jun. 2020.

[45] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed convex optimization on dynamic networks," *IEEE Trans. Autom. Control*, vol. 61, no. 11, pp. 3545–3550, Nov. 2016.

[46] N. Mazzi, B. S. Zhang, and D. S. Kirschen, "An online optimization algorithm for alleviating contingencies in transmission networks," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5572–5582, Sep. 2018.

[47] K. H. Lu, G. S. Jing, and L. Wang, "Online distributed optimization with strongly pseudoconvex-sum cost functions," *IEEE Trans. Autom. Control*, vol. 65, no. 1, pp. 426–433, Jan. 2019.

[48] X. L. Yi, X. X. Li, L. H. Xie, and K. H. Johansson, "Distributed online convex optimization with time-varying coupled inequality constraints," *IEEE Trans. Signal Process.*, vol. 68, pp. 731–746, Jan. 2020.

[49] D. M. Yuan, Y. G. Hong, D. W. C. Ho, and S. Y. Xu, "Distributed mirror descent for online composite optimization," *IEEE Trans. on Autom. Control*, vol. 66, no. 2, pp. 714–729, Feb. 2020.

[50] Q. G. Lu, X. F. Liao, T. Xiang, H. Q. Li, and T. W. Huang, "Privacy masking stochastic subgradient-push algorithm for distributed online optimization," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3224–3237, Jun. 2020.

[51] K. H. Lu and L. Wang, "Online distributed optimization with nonconvex objective functions: Sublinearity of first-order optimality condition-based regret," *IEEE Trans. Autom. Control*, 2021. DOI: 10.1109/TAC.2021.3091096..

[52] L. Ding, P. Hu, Z. W. Liu, and G. H. Wen, "Transmission lines overload alleviation: Distributed online optimization approach," *IEEE Trans. Ind. Inf.*, vol. 17, no. 5, pp. 3197–3208, May 2021.

[53] J. L. Raisaro, G. Choi, S. Pradervand, R. Colsenet, N. Jacquemont, N. Rosat, V. Mooser, and J. P. Hubaux, "Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential

privacy," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 5, pp. 1413–1426, Sep.–Oct. 2018.

[54] D. L. Oberski and F. Kreuter, "Differential privacy and social science: An urgent puzzle," *Harv. Data Sci. Rev.*, vol. 2, no. 1, pp. 1–21, Feb. 2020.

[55] M. Y. Hong, D. Hajinezhad, and M. M. Zhao, "Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proc. 34th Int. Conf. Machine Learning*, Sydney, Australia, 2017, pp. 1529–1538.

[56] D. Hajinezhad, M. Y. Hong, and A. Garcia, "ZONE: Zeroth-order nonconvex multi-agent optimization over networks," *IEEE Trans. Autom. Control*, vol. 64, no. 10, pp. 3995–4010, Oct. 2019.

[57] Y. J. Tang, J. S. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 1, pp. 269–281, Mar. 2021.

[58] Z. Y. He, J. P. He, C. L. Chen, and X. P. Guan, "Distributed nonconvex optimization: Gradient-free iterations and globally optimal solution," arXiv preprint arXiv: 2008.00252, 2020.

[59] Z. Y. He, J. P. He, C. L. Chen, and X. P. Guan, "Dependable distributed nonconvex optimization via polynomial approximation," arXiv preprint arXiv: 2101.06127, 2021.

[60] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics*, Fort Lauderdale, USA, 2017, pp. 1273–1282.

[61] Q. Yang, Y. Liu, T. J. Chen, and Y. X. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, p. 12, Mar. 2019.

[62] S. Hardy, W. Henecka, H. Ivey-Law, R. Nock, G. Patrini, G. Smith, and B. Thorne, "Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption," arXiv preprint arXiv: 1711.10677, 2017.

[63] Y. L. Lu, X. H. Huang, Y. Y. Dai, S. Maharjan, and Y. Zhang, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Trans. Ind. Inf.*, vol. 16, no. 3, pp. 2134–2143, Mar. 2020.

[64] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, "Bayesian nonparametric federated learning of neural networks," in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, USA, 2019, pp. 7252–7261.

[65] H. Y. Wang, M. Yurochkin, Y. K. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *Proc. 8th Int. Conf. Learning Representations*, Addis Ababa, Ethiopia, 2020.

[66] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. 37th Int. Conf. Machine Learning*, 2020, pp. 5132–5143.

[67] J. Jiang, S. X. Ji, and G. D. Long, "Decentralized knowledge acquisition for mobile internet applications," *World Wide Web*, vol. 23, no. 5, pp. 2653–2669, Mar. 2020.

[68] H. T. Nguyen, V. Sehwag, S. Hosseinalipour, C. G. Brinton, M. Chiang, and H. V. Poor, "Fast-convergent federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 201–218, Jan. 2021.

[69] R. Shokri, M. Stronati, C. Z. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security and Privacy*, San Jose, USA, 2017, pp. 3–18.

[70] J. L. Hou, R. Xi, P. Liu, and T. L. Liu, "The switching fractional order chaotic system and its application to image encryption," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 2, pp. 381–388, Apr. 2017.

[71] H. Y. Zhao, J. Yan, X. Y. Luo, and X. Gua, "Privacy preserving solution for the asynchronous localization of underwater sensor networks," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 6, pp. 1511–1527, Nov. 2020.

[72] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. 41st Ann. ACM Symp. Theory of Computing*, Bethesda, USA, 2009, pp. 169–178.

[73] C. L. Zhang, S. Y. Li, J. Z. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning," in *Proc. USENIX Ann. Tech. Conf.*, 2020, pp. 493–506.

[74] B. Jia, X. S. Zhang, J. W. Liu, Y. Zhang, K. Huang, and Y. Q. Liang, "Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT," *IEEE Trans. Ind. Inf.*, vol. 18, no. 6, pp. 4049–4058, Jun. 2022.

[75] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Theory of Cryptography Conf.*, New York, USA, 2006, pp. 265–284.

[76] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private aircomp federated learning with power adaptation harnessing receiver noise," in *Proc. IEEE Global Communications Conf.*, Taipei, China, 2020, pp. 1–6.

[77] I. Curiel, *Cooperative Game Theory and Applications: Cooperative Games Arising from Combinatorial Optimization Problems*. Berlin, Germany: Springer Science & Business Media, 2013.

[78] Y. Tang, H. J. Gao, and J. Kurths, "Multiobjective identification of controlling areas in neuronal networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 708–720, May 2013.

[79] Y. C. Jin, *Multi-Objective Machine Learning*. Berlin, Germany: Springer, 2006.

[80] W. Du, Y. Tang, S. Y. S. Leung, L. Tong, A. V. Vasilakos, and F. Qian, "Robust order scheduling in the discrete manufacturing industry: A multiobjective optimization approach," *IEEE Trans. Ind. Inf.*, vol. 14, no. 1, pp. 253–264, Jan. 2018.

[81] K. Q. Zhang, Z. R. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control*, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Eds. Cham, Germany: Springer, 2021, pp. 321–384.

[82] Z. Y. Zuo, Q. L. Han, B. D. Ning, X. H. Ge, and X. M. Zhang, "An overview of recent advances in fixed-time cooperative control of multi-agent systems," *IEEE Trans. Ind. Inf.*, vol. 14, no. 6, pp. 2322–2334, Jun. 2018.

[83] X. H. Ge, Q. L. Han, D. R. Ding, X. M. Zhang, and B. D. Ning, "A survey on recent advances in distributed sampled-data cooperative control of multi-agent systems," *Neurocomputing*, vol. 275, pp. 1684–1701, Jan. 2018.

[84] R. H. Crites and A. G. Barto, "Improving elevator performance using reinforcement learning," in *Proc. Advances in Neural Information Proc. Systems*, Denver, USA, 1995, pp. 1017–1023.

[85] G. E. Monahan, "State of the art-a survey of partially observable Markov decision processes: Theory, models, and algorithms," *Manag. Sci.*, vol. 28, no. 1, pp. 1–16, Jan. 1982.

[86] J. N. Foerster, Y. M. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. 30th Advances in Neural Information Proc. Systems*, Barcelona, Spain, 2016, pp. 2137–2145.

[87] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proc. 17th Int. Conf. Autonomous Agents and Multi-Agent Systems*, Stockholm, Sweden, 2018, pp. 2085–2087.

[88] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. 31st Int. Conf. Neural Information Proc. Systems*, Long Beach, USA, 2017, pp. 6382–6393.

[89] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multi-agent communication with backpropagation," in *Proc. 30th Int. Conf. Neural Information Proc. Systems*, Barcelona, Spain, 2016, pp. 2252–2260.

[90] A. Das, T. Gervet, J. Romoff, D. Batra, D. Parikh, M. Rabbat, and J. Pineau, "TarMAC: Targeted multi-agent communication," in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, USA, 2019, pp. 1538–1546.

[91] G. Chen, "A new framework for multi-agent reinforcement learning–centralized training and exploration with decentralized

execution via policy distillation," in *Proc. 19th Int. Conf. Autonomous Agents and Multi-Agent Systems*, Auckland, New Zealand, 2020, pp. 1801–1803.

[92] T. Rashid, M. Samvelyan, C. S. De Wit, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. 35th Int. Conf. Machine Learning*, Stockholm, Sweden, 2018, pp. 4295–4304.

[93] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, "QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, USA, 2019, pp. 5887–5896.

[94] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, "Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. 34th Ann. Conf. Neural Information Proc. Systems*, Vancouver, Canada, 2020, pp. 10199–10210.

[95] J. H. Wang, Z. Z. Ren, T. Liu, Y. Yu, and C. J. Zhang, "QPLEX: Duplex dueling multi-agent Q-learning," in *Proc. 9th Int. Conf. Learning Representations*, 2021, pp. 1–27.

[96] Y. D. Yang, J. Y. Hao, B. Liao, K. Shao, G. Y. Chen, W. L. Liu, and H. Y. Tang, "Qatten: A general framework for cooperative multi-agent reinforcement learning," arXiv preprint arXiv: 2002.03939, 2020.

[97] J. Y. Su, S. C. Adams, and P. A. Beling, "Value-decomposition multi-agent actor-critics," in *Proc. 35th AAAI Conf. Artificial Intelligence*, 2021, pp. 11352–11360.

[98] E. Winter, "The Shapley value," *Handbook Game Theory Econom. Appl.*, vol. 3, pp. 2025–2054, Aug. 2002.

[99] R. E. Wang, M. Everett, and J. P. How, "R-MADDPG for partially observable environments and limited communication," in *Proc. Workshop in the 36th Int. Conf. Machine Learning*, Long Beach, USA, 2020, pp. 1–9.

[100] K. Liu, Y. Y. Zhao, G. Wang, and B. Peng, "Self-attention-based multi-agent continuous control method in cooperative environments," *Inf. Sci.*, vol. 585, pp. 454–470, Mar. 2021.

[101] T. P. Yang, W. X. Wang, H. Y. Tang, J. Y. Hao, Z. P. Meng, H. Y. Mao, D. Li, W. L. Liu, Y. F. Chen, Y. J. Hu, C. J. Fan, and C. W. Zhang, "An efficient transfer learning framework for multi-agent reinforcement learning," in *Proc. 35th Advances in Neural Information Proc. Systems*, 2021.

[102] C. H. Liu, Z. P. Dai, Y. N. Zhao, J. Crowcroft, D. P. Wu, and K. K. Leung, "Distributed and energy-efficient mobile crowdsensing with charging stations by deep reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 20, no. 1, pp. 130–146, Jan. 2021.

[103] D. Bauso, *Game Theory with Engineering Applications*. Philadelphia, USA: SIAM, 2016.

[104] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proc. 10th Int. Conf. Machine Learning*, Amherst, USA, 1993, pp. 330–337.

[105] P. Liu, W. Y. Zang, and M. Yu, "Incentive-based modeling and inference of attacker intent, objectives, and strategies," *ACM Trans. Inf. Syst. Secur.*, vol. 8, no. 1, pp. 78–118, Feb. 2005.

[106] O. D. Altan, G. Y. Wu, M. J. Barth, K. Boriboonsomsin, and J. A. Stark, "Glidepath: Eco-friendly automated approach and departure at signalized intersections," *IEEE Trans. Intell. Veh.*, vol. 2, no. 4, pp. 266–277, Dec. 2017.

[107] K. Kang and H. A. Rakha, "Game theoretical approach to model decision making for merging maneuvers at freeway on-ramps," *Transp. Res. Rec.*, vol. 2623, no. 1, pp. 19–28, Jan. 2017.

[108] H. White, H. Q. Xu, and K. Chalak, "Causal discourse in a game of incomplete information," *J. Econometrics*, vol. 182, no. 1, pp. 45–58, Sep. 2014.

[109] S. N. Xia, F. L. Lin, Z. Y. Chen, C. B. Tang, Y. J. Ma, and X. H. Yu, "A Bayesian game based vehicle-to-vehicle electricity trading scheme for blockchain-enabled internet of vehicles," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 6856–6868, Jul. 2020.

[110] H. W. Zhang, J. D. Wang, D. K. Yu, J. H. Han, and T. Li, "Active defense strategy selection based on static Bayesian game," in *Proc. 3rd Int. Conf. Cyberspace Technology*, Beijing, China, 2015, pp. 1–7.

[111] J. Huang, C. C. Xing, Y. Qian, and Z. J. Haas, "Resource allocation for multicell device-to-device communications underlaying 5G networks: A game-theoretic mechanism with incomplete information," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2557–2570, Mar. 2018.

[112] S. S. Hasanabadi, A. H. Lashkari, and A. A. Ghorbani, "A memorybased game-theoretic defensive approach for digital forensic investigators," *Forensic Sci. Int. Digital Invest.*, vol. 38, p. 301214, Sep. 2021. DOI: 10.1016/j.fsidi.2021.301214.

[113] R. Yan, Z. Y. Shi, and Y. S. Zhong, "Task assignment for multiplayer reach–avoid games in convex domains via analytical barriers," *IEEE Trans. Rob.*, vol. 36, no. 1, pp. 107–124, Feb. 2020.

[114] H. M. Huang, J. Ding, W. Zhang, and C. J. Tomlin, "Automation-assisted capture-the-flag: A differential game approach," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 1014–1028, Mar. 2015.

[115] M. Mitchell, A. M. Bayen, and C. J. Tomlin, "A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games," *IEEE Trans. Autom. Control*, vol. 50, no. 7, pp. 947–957, Jul. 2005.

[116] R. Yan, Z. Y. Shi, and Y. S. Zhong, "Reach-avoid games with two defenders and one attacker: An analytical approach," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1035–1046, Mar. 2019.

[117] E. Ben-Porath, "Rationality, Nash equilibrium and backwards induction in perfect-information games," *Rev. Econom. Stud.*, vol. 64, no. 1, pp. 23–46, Jan. 1997.

[118] S. D. Levitt, J. A. List, and S. E. Sadoff, "Checkmate: Exploring backward induction among chess players," *Am. Econom. Rev.*, vol. 101, no. 2, pp. 975–990, Apr. 2011.

[119] M. Bowling, N. Burch, M. Johanson, and O. Tammelin, "Heads-up limit Hold'em poker is solved," *Science*, vol. 347, no. 6218, pp. 145–149, Jan. 2015.

[120] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. 14th Int. Conf. Artificial Intelligence and Statistics*, Fort Lauderdale, USA, 2011, pp. 627–635.

[121] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione, "Regret minimization in games with incomplete information," in *Proc. Advances in Neural Information Proc. Systems*, Vancouver, Canada, 2007, pp. 1729–1736.

[122] N. Brown and T. Sandholm, "Solving imperfect-information games via discounted regret minimization," in *Proc. 33rd AAAI Conf. Artificial Intelligence*, Honolulu, USA, 2019, pp. 1829–1836.

[123] H. Li, K. L. Hu, S. H. Zhang, Y. Qi, and L. Song, "Double neural counterfactual regret minimization," in *Proc. 7th Int. Conf. Learning Representations*, Addis Ababa, Ethiopia, 2020, pp. 1–20.

[124] N. Brown, A. Lerer, S. Gross, and T. Sandholm, "Deep counterfactual regret minimization," in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, USA, 2019, pp. 793–802.

[125] E. Steinberger, "Single deep counterfactual regret minimization," arXiv preprint arXiv: 1901.07621, 2019.

[126] S. X. Li, Y. Z. Zhang, X. R. Wang, W. Q. Xue, and B. An, "CFR-MIX: Solving imperfect information extensive-form games with combinatorial action space," in *Proc. 30th Int. Joint Conf. Artificial Intelligence*, Montreal, Canada, 2021, pp. 3663–3669.

[127] D. Monderer and L. S. Shapley, "Fictitious play property for games with identical interests," *J. Econ. Theory*, vol. 68, no. 1, pp. 258–265, Jan. 1996.

[128] J. Heinrich and D. Silver, "Deep reinforcement learning from self-play in imperfect-information games," in *Proc. 3rd Workshops at Advances Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 1–10.

[129] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multi-agent deep reinforcement learning," *Auton. Agents Multi-Agent Syst*, vol. 33, no. 6, pp. 750–797, Oct. 2019.

[130] W. Q. Xue, Y. Z. Zhang, S. X. Li, X. R. Wang, B. An, and C. K. Yeo, "Solving large-scale extensive-form network security games via neural fictitious self-play," in *Proc. 30th AAAI Conf. Artificial Intelligence*, Montreal, Canada, 2021, pp. 3713–3720.

[131] W. B. Li, J. L. Xu, J. Huo, L. Wang, Y. Gao, and J. B. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proc. 33rd AAAI Conf. Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conf. and Ninth AAAI Symp. Educational Advances in Artificial Intelligence*, Honolulu, USA, 2019, pp. 8642–8649.

[132] M. Y. Huang, R. P. Malhamé, and P. E. Caines, "Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle," *Commun. Inf. Syst.*, vol. 6, no. 3, pp. 221–252, Jan. 2006.

[133] M. Y. Huang, P. E. Caines, and R. P. Malhame, "Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ε-Nash equilibria," *IEEE Trans. Autom. Control*, vol. 52, no. 9, pp. 1560–1571, Sep. 2007.

[134] J. Moon and T. Başar, "Linear quadratic mean field Stackelberg differential games," *Automatica*, vol. 97, pp. 200–213, Nov. 2018.

[135] J. Moon and T. Başar, "Linear quadratic risk-sensitive and robust mean field games," *IEEE Trans. Autom. Control*, vol. 62, no. 3, pp. 1062–1077, Mar. 2017.

[136] T. Li and J. F. Zhang, "Decentralized tracking-type games for multi-agent systems with coupled ARX models: Asymptotic Nash equilibria," *Automatica*, vol. 44, no. 3, pp. 713–725, Mar. 2008.

[137] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[138] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. T. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[139] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, "Mastering Atari, Go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, Dec. 2020.

[140] N. Brown and T. Sandholm, "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals," *Science*, vol. 359, no. 6374, pp. 418–424, Dec. 2017.

[141] N. Brown and T. Sandholm, "Superhuman AI for multiplayer poker," *Science*, vol. 365, no. 6456, pp. 885–890, Jul. 2019.

[142] X. J. Wang, J. X. Song, P. H. Qi, P. Peng, Z. K. Tang, W. Zhang, W. M. Li, X. J. Pi, J. J. He, C. Gao, H. T. Long, and Q. Yuan, "SCC: An efficient deep reinforcement learning agent mastering the game of StarCraft II," in *Proc. 38th Int. Conf. Machine Learning*, 2021, pp. 10905–10915.

[143] Y. X. Chen, L. Zhang, S. J. Li, and G. Pan, "Optimize neural fictitious self-play in regret minimization thinking," arXiv preprint arXiv: 2104.10845, 2021.

[144] S. Mao, Z. W. Dong, W. Du, Y. C. Tian, C. Liang, and Y. Tang, "Distributed non-convex event-triggered optimization over time-varying directed networks," *IEEE Trans. Ind. Inf.*, 2021. DOI: 10.1109/TII.2021.3103747.

[145] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," arXiv preprint arXiv: 1906.06629, 2019.

[146] U. Mahadev, "Classical homomorphic encryption for quantum circuits," in *Proc. IEEE 59th Ann. Symp. Foundations of Computer Science*, Paris, France, 2018, pp. 332–338.

[147] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, vol. 378, pp. 686–707, Feb. 2019.

[148] A. E. Roth and M. A. O. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Cambridge, UK: Cambridge University Press, 1992, pp. 485–541.

[149] S. Kumabe and T. Maehara, "Convexity of b-matching games." in *Proc. 29th Int. Joint Conf. Artificial Intelligence*, Yokohama, Japan, 2020, pp. 261–267.

[150] I. Gemp, B. McWilliams, C. Vernade, and T. Graepel, "Eigengame unloaded: When playing games is better than optimizing," arXiv preprint arXiv: 2102.04152, 2021.

[151] S. Pateria, B. Subagdja, A. H. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," *ACM Comput. Surv.*, vol. 54, no. 5, p. 109, Jun. 2022.

[152] H. Ryu, H. Shin, and J. Park, "Multi-agent actor-critic with hierarchical graph attention network," in *Proc. 34th AAAI Conf. Artificial Intelligence*, New York, USA, 2020, pp. 7236–7243.

[153] L. X. Wang, Z. R. Yang, and Z. R. Wang, "Breaking the curse of many agents: Provable mean embedding Q-iteration for mean-field reinforcement learning," in *Proc. 37th Int. Conf. Machine Learning*, 2020, pp. 10092–10103.

[154] F. Schubert, M. Awiszus, and B. Rosenhahn, "TOAD-GAN: A flexible framework for few-shot level generation in token-based games," *IEEE Trans. Games*, 2021. DOI: 10.1109/TG.2021.3069833.

[155] Y. Chen, J. M. Liu, and B. Khoussainov, "Maximum entropy inverse reinforcement learning for mean field games," arXiv preprint arXiv: 2104.14654, 2021.

[156] A. R. Tilman, J. B. Plotkin, and E. Akçay, "Evolutionary games with environmental feedbacks," *Nat. Commun.*, vol. 11, no. 1, p. 915, Feb. 2020. DOI: 10.1038/s41467-020-14531-6.

[157] D. Mandal, S. Medya, B. Uzzi, and C. Aggarwal, "Meta-learning with graph neural networks: Methods and applications," arXiv preprint arXiv: 2103.00137, 2021.

[158] Y. Shu, Z. J. Cao, C. Y. Wang, J. M. Wang, and M. S. Long, "Open domain generalization with domain-augmented meta-learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, USA, 2021, pp. 9619–9628.

[159] DI-engine Contributors, "DI-engine: OpenDILab decision intelligence engine," 2021. [Online]. Available: https://github.com/opendilab/DI-engine.

Jianrui Wang received the B.S. degree in automation from Dalian Maritime University in 2019. He is currently a Ph.D. candidate in control science and engineering at the School of Information Science and Engineering, East China University of Science and Technology. His current research interests include deep reinforcement learning, multi-agent reinforcement learning, game theory, and their applications.

Yitian Hong received the B.S. degree in intelligence science and technology from Xidian University in 2020. He is currently a Ph.D. candidate in computer science and technology at the School of Information Science and Engineering, East China University of Science and Technology. His current research interests include multi-agent deep reinforcement learning, federated learning, and their applications.

Jiali Wang received the M.S. degree in mathematics from the College of Mathematics and Computer Science, Zhejiang Normal University in 2020. She is currently a Ph.D. candidate in control science and engineering at the School of Information Science and Engineering, East China University of Science and Technology. Her current research interests include game theory, optimal control and their applications.

Jiapeng Xu received the Ph.D. degree in control science and engineering from the School of Information Science and Engineering, East China University of Science and Technology in 2021. From October 2019 to October 2020, he was a Visiting Scholar in the Department of Electrical Engineering, University of Notre Dame, USA. He is currently a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, University of Windsor, Canada. His research interests include networked feedback control, event-triggered state estimation, distributed filtering and mean-field games.

Yang Tang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Donghua University in 2006 and 2010, respectively. From 2008 to 2010, he was a Research Associate with The Hong Kong Polytechnic University, China. From 2011 to 2015, he was a Post-Doctoral Researcher with the Humboldt University of Berlin, Germany, and with the Potsdam Institute for Climate Impact Research, Germany. He is now a Professor with the East China University of Science and Technology. His current research interests include distributed estimation/control/optimization, cyber-physical systems, hybrid dynamical systems, computer vision, reinforcement learning and their applications.

He was a recipient of the Alexander von Humboldt Fellowship and has been the ISI Highly Cited Researchers Award by Clarivate Analytics from 2017. He is a Senior Board Member of Scientific Reports, an Associate Editor of *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Circuits and Systems−I*: *Regular Papers*, *IEEE Transactions on Emerging Topics in Computational Intelligence*, *IEEE Systems Journal and Engineering Applications of Artificial Intelligence* (IFAC Journal), etc. He is a Leading Guest Editor for special issues in *IEEE Transactions on Emerging Topics in Computational Intelligence* and *IEEE Transactions on Cognitive and Developmental Systems*.

Qing-Long Han (Fellow, IEEE) received the B.Sc. degree in mathematics from Shandong Normal University in 1983, and the M.Sc. and Ph.D. degrees in control engineering from East China University of Science and Technology in 1992 and 1997, respectively. He is Pro Vice-Chancellor (Research Quality) and a Distinguished Professor at Swinburne University of Technology, Australia. He held various academic and management positions at Griffith University and Central Queensland University, Australia. His research interests include networked control systems, multi-agent systems, time-delay systems, smart grids, unmanned surface vehicles, and neural networks.

Professor Han was the recipient of The 2021 Norbert Wiener Award (the Highest Award in systems science and engineering, and cybernetics), The 2021 M. A. Sargent Medal (the Highest Award of the Electrical College Board of Engineers Australia), The 2021 IEEE/CAA Journal of Automatica Sinica Norbert Wiener Review Award, The 2020 IEEE Systems, Man, and Cybernetics (SMC) Society Andrew P. Sage Best Transactions Paper Award, The 2020 IEEE Transactions on Industrial Informatics Outstanding Paper Award, and The 2019 IEEE SMC Society Andrew P. Sage Best Transactions Paper Award.

He is a Member of the Academia Europaea (The Academy of Europe) and a Fellow of The Institution of Engineers Australia. He is a Highly Cited Researcher in both Engineering and Computer Science (Clarivate Analytics, 2019−2021). He has served as an AdCom Member of IEEE Industrial Electronics Society (IES), a Member of IEEE IES Fellow Committee, and Chair of IEEE IES Technical Committee on Networked Control Systems. He is Co-Editor-in-Chief of *IEEE Transactions on Industrial Informatics*, Deputy Editor-in-Chief of *IEEE/CAA Journal of Automatica Sinica*, Co-Editor of *Australian Journal of Electrical and Electronic Engineering*, an Associate Editor for 12 international journals, including the *IEEE Transactions on Cybernetics*, *IEEE Industrial Electronics Magazine*, *Control Engineering Practice*, and *Information Sciences*, and a Guest Editor for 13 Special Issues.

Jürgen Kurths received the B.S. degree in mathematics from the University of Rostock, Germany, the Ph.D. degree from the Academy of Sciences of the German Democratic Republic, Germany, in 1983, the Honorary degree from the N. I. Lobachevsky State University of Nizhny Novgorod, Russia in 2008, and the Honorary degree from Saratov State University, Russia in 2012. He was a Full Professor with the University of Potsdam, Germany, from 1994 to 2008. Since 2008, he has been a Professor of nonlinear dynamics with the Humboldt University of Berlin, Germany and the Chair of the research domain transdisciplinary concepts with the Potsdam Institute for Climate Impact Research, Potsdam. His current research interests include synchronization, complex networks, time series analysis, and their applications. Since 2009, he has been the Sixth-Century Chair with the University of Aberdeen, UK. He has authored over 680 articles, which are cited more than 40000 times (H-index: 104). He became a Member of the Academia Europaea in 2010, the Macedonian Academy of Sciences and Arts in 2012, and the Royal Society of Edinburgh in 2021. He is a Fellow of the American Physical Society. He received the Alexander von Humboldt Research Award from the Council of Scientific and International Research, India in 2015. He is named as an ISI Highly Cited Researcher in physics and engineering by Thomson Reuters. He is the Editor-in-Chief of CHAOS.