Deep Reinforcement Learning for Power System Applications: An Overview

Zidong Zhang, Dongxia Zhang, and Robert C. Qiu, *Fellow, IEEE*

***Abstract*—Due to increasing complexity, uncertainty and data dimensions in power systems, conventional methods often meet bottlenecks when attempting to solve decision and control problems. Therefore, data-driven methods toward solving such problems are being extensively studied. Deep reinforcement learning (DRL) is one of these data-driven methods and is regarded as real artificial intelligence (AI). DRL is a combination of deep learning (DL) and reinforcement learning (RL). This field of research has been applied to solve a wide range of complex sequential decision-making problems, including those in power systems. This paper firstly reviews the basic ideas, models, algorithms and techniques of DRL. Applications in power systems such as energy management, demand response, electricity market, operational control, and others are then considered. In addition, recent advances in DRL including the combination of RL with other classical methods, and the prospect and challenges of applications in power systems are also discussed.**

***Index Terms*—Artificial intelligence, deep reinforcement learning, machine learning, power system, smart grids.**

NOMENCLATURE

A2C	Advantage actor-critic.
A3C	Asynchronous advantage actor-critic.
ACO	Ant colony optimization.
AGC	Automatic generation control.
AI	Artificial intelligence.
ANI	Artificial narrow intelligence.
ANN	Artificial neural network.
AVC	Automatic voltagecontrol.
BFRL	Bacteria foraging reinforcement learning.
B-M	Bush-Mosteller.
CNN	Convolutional neural network.
CTQ	Consensus transferQ-learning.
DBN	Deep belief network.
DDPG	Deep deterministic policy gradient.
DDRQN	Deep distributed recurrent Q-networks.
DFRL	Deepforest reinforcement learning.
DG	Distributed generation.
DL	Deep learning.
DNN	Deep neural network.
DPG	Deterministic policy gradient.
DQL	Deep Q-learning.
DQN	Deep Q-network.
DR	Demand response
DRL	Deep reinforcement learning.
DTQ	Deep transfer network.
EH	Energy harvesting.
EI	Energy Internet.
EM	Electricity market.
EMA	Exponential moving average.
ES	Energy storage.
EV	Electric vehicle.
GA	Genetic algorithm.
GCD	Generation commanddispatch.
GRU	Gated recurrent unit.
HEV	Hybrid electric vehicle.
HVAC	Heating, ventilation and air conditioning.
LEC	Levelized energy cost.
LFC	Load frequency control.
LSE	Load serving entity.
LSTM	Long short-term memory.
MCES	Multi-carrier energy system.
MEC	Multi-access edge computing.
MILP	Mixed-integer linear programming.
MDP	Markov decision process.
ML	Machine learning.
NE	Nash Equilibrium.
NLP	Natural language processing.
OEM	Optimal energy management.
PEV	Plug-in electric vehicle.
PHEV	Plug-in hybrid electric vehicle.
PPO	Proximal policy optimization.
PSO	Particle swarm optimization.
PV	Photovoltaic.
RBM	Restricted Boltzmann machine.
RL	Reinforcement learning.
RNN	Recurrent neural network.
SDAE	Stacked denoising auto-encoders.
SG	Smart Grid.
SGC	Smart generation control.
SoC	State of charge.
SP	Service provider.

Manuscript received May 5, 2019; revised June 28, 2019; accepted August 15, 2019. Date of publication March 30, 2020; date of current version December 26, 2019. This work is supported by National Natural Science Foundation of China under Grant No. 61571296 and the National Key Research and Development Program of China under 2018YFF0214705.

Z. D. Zhang and R. C. Qiu (corresponding author, email: rcqiu@sjtu.edu.cn) are with Research Center For Big Data Engineering And Technologies, Shanghai Jiao Tong University, Shanghai 200240, China. R. C. Qiu is also with Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville, TN 38505, USA.

D. X. Zhang is with China Electric Power Research Institute, Beijing 100192, China.

DOI: 10.17775/CSEEJPES.2019.00920

I. Introduction

A. Background

A power system is a complex, dynamic, large-scale network of electrical components. Power systems have gone through many decades of development. During this time, economic, technological, environmental and political incentives have transformed conventional grids into more complex, robust, efficient and sustainable smart grids [1]–[3]. Smart grids use bi-directional energy flow accompanied by bi-directional information flow among all the participants, including producers, consumers, transmission and distribution system operators and demand response aggregators [4], [5]. Such factors have brought problems to the power system from different aspects. Firstly, a high penetration of renewable power (such as wind and solar) brings greater uncertainty to a power system. Furthermore, the deregulation of the electricity market and active participation of customers makes it more complex to find solutions that allow the incorporation of distributed energy resources [6], [7].

To solve these problems, effective methods are required for planning and operating the grid. This ongoing transformation of grids results in increased uncertainty and complexity in both the business transactions and the in physical flows of electricity [8], [9]. Moreover, the explosion of information and the fluctuation of data makes decision-making problems difficult, compared to traditional methods [10], [11]. Therefore, future smart grids need systems that can monitor, predict, schedule, learn and make decisions regarding energy consumption and production in real-time. This calls for a more efficient and intelligent solution, such as deep reinforcement learning [12]–[14].

Reinforcement learning, derived from neutral stimulus and response, is a machine learning method. It has become increasingly popular due to its success in addressing challenging sequential decision-making problems [15], [16]. Its combination with deep learning, called deep reinforcement learning, has achieved great success in games [17]–[19], robotics [20], [21], natural language processing (NLP) [21]–[23], finance and business management [24], [25]. Many papers have reported the application of deep reinforcement learning in power systems, and will be introduced in the following.

B. Methodology

Many problems in the power system can be transformed into sequential decision-making tasks. Traditional methods mainly include convex optimization methods, programming methods, and heuristic methods. Through qualitative comparison with DRL, the advantages and disadvantages of these methods are explained as follows [26].

The first is a classical mathematical method, such as the Lyapunov optimization algorithm [27]. The advantage of this method is that the mathematics are rigorous and real-time management can be realized. However, this type of method relies on explicit objective functional expressions, which are difficult to abstract from many real-world optimization decision scenarios. Moreover, the Lyapunov condition (required for the Lyapunov optimization algorithm), cannot be guaranteed in complicated, high-dimensional scenarios.

The second is the programming method, such as mixed integer programming [28], [29], dynamic programming [30], [31], and stochastic programming [32], [33]. These methods can solve a variety of optimization problems, especially sequence optimization problems. However, each iteration of this type of method needs to be recalculated from the beginning. In addition, the calculation cost is too large to realize real-time decision-making in some scenarios. Some cases using programming algorithms rely on accurate predictions of renewable energy generation and load, which are difficult to achieve in real scenarios.

Another category is heuristic methods, such as genetic algorithms (GA) [34], [35], ant colony optimization (ACO) [36], [37], or particle swarm optimization (PSO) [38], [39]. For the optimization problems, especially non-convex optimization problems, a heuristic method can achieve the local optimal solution with a certain probability, which is beneficial to solve the problem of large data scale and complicated scenarios. However, these methods are less robust and cannot be proven rigorously using mathematics.

Compared with convex optimization methods, the exact objective function is not necessary for DRL. In contrast, DRL uses the reward function to evaluate decision behavior. DRL can also handle higher dimensional data than convex optimization methods. Against the programming methods, DRL makes decisions according to the current state and thus makes real-time and online decisions. In contrast to the heuristic methods, DRL is more robust with stable convergence results and is better suited for decision-making problems.

The principles and algorithms of DRL are introduced briefly in Section II. The applications of DRL in the power system are classified in Section III. The prospect and challenges of DRL and its applications in the power system are also discussed in Section IV.

II. Deep Reinforcement Learning

Deep reinforcement learning combines the perception function of deep learning with the decision-making ability of reinforcement learning. It is an artificial intelligence method closer to human thinking and is regarded as real AI. The basic framework of DRL is shown in Fig. 1. The deep learning gets the target observation information from the environment

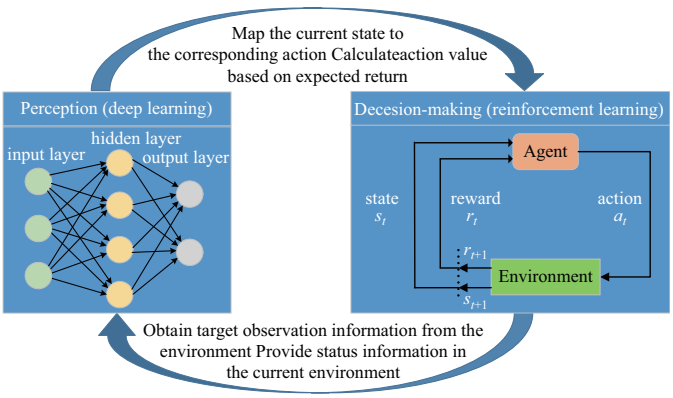

Fig. 1. Deep reinforcement learning framework.

and provides the state information in the current environment. The reinforcement learning then maps the current state to the corresponding action and evaluates values based on the expected return [40], [41]. A continuous interaction process makes decision-making behavior a step by step process. The principles and algorithms of reinforcement learning are introduced in the next section.

A. Reinforcement Learning

Reinforcement learning is applied to calculate a behavior strategy, a policy that maximizes a satisfaction criterion. Meanwhile, a long-term sum of rewardsis obtained by interacting through trial and error with a given environment. To implement these functions, a reinforcement learning framework consists of a decision-maker, called the *agent*, operating in an environment modeled by state s_t. The agent is capable of taking certainaction a_t, as a function of the current state s_t. After choosing an action at time t, the agent receives a scalar reward r_{t+1} and finds itself in a new state s_{t+1} that depends on the current state and the chosen action, just as shown in Fig. 1. The mathematical foundations and concepts of reinforcement learning are introduced in the following.

1) Markov Decision Process (MDP)

A Markov decision process, as shown in Fig. 2, which satisfies a Markov property and is a basic formalism of reinforcement learning. A Markov property is one in which the future of the process only depends on the current state, and the agent has no interest in the full history. It can be described as:

$$P(s_{t+1}|s_0, a_0, \cdots, s_t, a_t) = P(s_{t+1}|s_t, a_t) \tag{1}$$

where P is state transition probability.

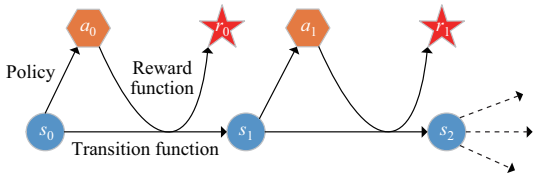

Fig. 2. Illustration of Markov decision process.

At each epoch, the agent takes an action that changes its state in the environment and provides a reward. To further process the reward value, value functions and optimal policy are proposed.

2) Value Function and Optimal Policy

To maximize the long-term cumulative reward after the current time t, in the case of a finite time horizon that ends at time T, the *return* R_t is equal to:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2}$$

where the discount factor $\gamma \in [0, 1]$, and γ can take 1 only in episodic MDPs.

In order to find an optimal policy, some algorithms are based on value function $V(s)$, which represents how beneficial it is for the agent to reach a given state s. Such a function depends on the actual policy π followed by the agent:

$$V^{\pi}(s_t) = \mathbf{E}[R_t|s_t = s] = \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s\right] \tag{3}$$

Similarly, the action-value function Q represents the value of taking action a in state s under a policy π as:

$$\begin{aligned} Q^{\pi}(s_t, a_t) &= \mathbf{E}[R_t|s_t = s, a_t = a] \\ &= \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s, a_t = a\right] \end{aligned} \tag{4}$$

In a Q-learning algorithm [42], the Q function can be expressed as an iterative form by the Bellman equation:

$$Q^{\pi}(s_t, a_t) = \mathbf{E}[r_{t+1} + \gamma Q^{\pi}(s_{t+1}, a_{t+1})|s_t, a_t] \tag{5}$$

An optimal policy π^* is a policy that achieves the largest cumulative reward in the long run:

$$\pi^* = \arg\max_{\pi} V^{\pi}(s) \tag{6}$$

At this time, the best value function and action-value function will be:

$$V^*(s) = \max_{\pi} V^{\pi}(s) \tag{7}$$

$$Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a) \tag{8}$$

B. From RL to DRL

The journey from RL to DRL has gone through a long development process. In classical tabular RL, e.g. Q-learning, state and action spaces are small enough for the approximate value functions to be represented as arrays or tables. In this case, the methods can often find the exact optimal value functions and the optimal policies [15]. However, these previous methods suffer from a difficult design issue when they come to real-world implementations. To overcome this problem, the approximate value functions are represented as a parameterized functional form with a weight vector (similar todeep neural networks), instead of a table. DRL can complete complicated tasks with lower prior knowledge thanks to its ability to learn levels of abstractions from data [43], [44]. Further details about DL and DRL are explained below.

1) High-dimensional and continuous spaces. Although a variety of real-world problems lead to high-dimensional and continuous state spaces or action spaces, it is still not possible to store them in a table or function. This phenomenon is known as the 'curse of dimensionality'. To overcome this issue, function approximation is used to obtain features from models, value functions or policies and then attempts to generalize from them to construct an approximation of the entire function by supervised learning such as deep neural networks [45], [46].

2) Exploration-exploitation dilemma. When an agent starts accumulating information about the environment, it has to navigate a tradeoff between learning more about the environment (exploration) or pursuing the most promising strategy with the experience gathered (exploitation). In tabular RL, uncertainty about the reward function and transition probabilities can be quantified as confidence intervals or posterior of environment

parameters. In DRL, different settings are applied. One is that the agent explores only when the learning opportunities are valuable enough so that it can perform well without a separate training phase. Another is that the agent follows a training policy during the first phase of interactions with the environment so as to accumulate training data and hence learn a test policy [15], [40].

3) Convergence and stability. For RL, only tables and linearly parameterized approximators can be used to guarantee convergence. When prior knowledge is not available to guide the selection of basis functions, a large number of basis functions must be defined to evenly cover the state-action space, and this is impractical in high-dimensional problems. To address this problem, non-linear approximators, such as convolutional neural networks (CNN), have been applied to obtain features of certain parts of states with replay buffer and target networks [16], [43], [45].

C. DRL Algorithms

DRL problems may be formulated as optimization, planning, management, and control problems. Solution methods may be model-free or model-based and value-based or policy-based, just as shown in Fig. 3. Model-based DRL is strongly influenced by control theory and often is explained in terms of different disciplines. In contrast, model-free DRL ignores the model and cares less about the inner workings. Model-based DRL has the advantage of being simple and efficient. For example, if it is appropriate to approximate the space as linear, it will take much fewer samples to learn the model. However, model-based methods are more complex than model-free methods by a few orders of magnitude. If sampling can be done using a computer simulation, model-free methods finish faster. Also, to simplify the computation, model-based methods have more assumptions and approximations and therefore, may limit themselves to specific types of tasks. Most policy-based, value-based methods are model-free.

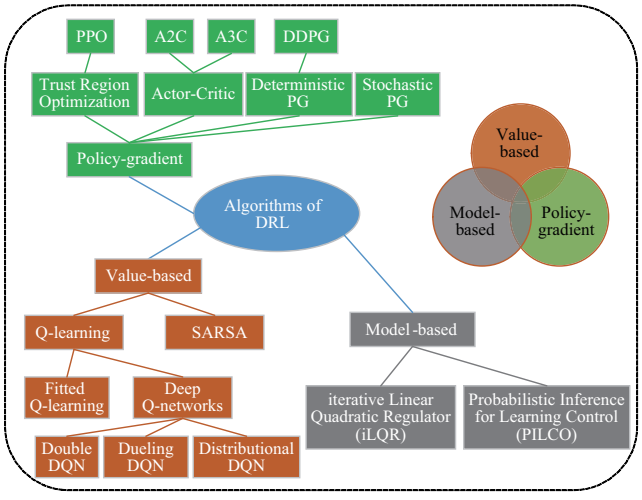

Fig. 3. Main DRL algorithms and theirinter-relationship.

Value-based methods learn from any trajectory sampled from the same environment by improving the value function at each iteration until the value-function converges. For tabular RL, e.g. Q-learning, the iteration process of Q function is as shown in (9), while in DRL it will update as shown in (10). At this time, the objective function can be defined as (11).

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_{\mathrm{a}} Q(s_{t+1},a) - Q(s_t,a_t)] \tag{9}$$

$$\theta_{t+1} = \theta_t + \alpha(r_{t+1} + \gamma \max_{\mathrm{a}} Q(s_{t+1},a,\theta) - Q(s_t,a_t,\theta))\nabla_\theta Q(s_t,a_t,\theta) \tag{10}$$

$$J(\theta) = \mathbf{E}[(r_{t+1} + \gamma \max_{\mathrm{a}} Q(s_{t+1},a_{t+1},\theta) - Q(s_t,a_t,\theta))^2] \tag{11}$$

where α is the learning rate, and θ is the collection of the parameters of function approximator [47], [48].

The policy-based methods directly optimize the quantity of interest while remaining stable under the function approximations by re-defining the policy at each step and compute the value according to this new policy until the policy converges. At first, the gradient of the objective function is calculated as policy parameters as shown in (11), and then the weight matrix will update using (12).

$$\nabla_\theta J(\theta) = \mathbf{E}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t=0}^{T} r(s_t,a_t)\right]$$
$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta) \tag{12}$$

III. Applications in Power System

After years of research, many papers have been published about the applications of DRL in power systems, and most of them were published since 2018. These applications covera wide range of decision, control and optimization problems in the power system, including energy management, demand response, electricity market, operational control and many others, as shown in Table I. This section reviews some typical application fields.

A. Energy Management

In a power system, especially a microgrid, energy management problems link source, load, storage system, and utility grid, as shown in Fig. 4. Energy management plays an essential role in several ways. Firstly, it can improve the utilization rate of renewable energy and manage household appliance-consumption. Furthermore, it can plan a storage scheduling

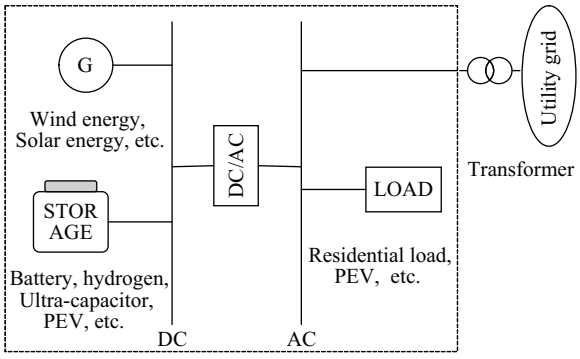

Fig. 4. Typical framework of an energy management system.

TABLE I
SUMMARY OF THE PAPERS

Paper	Fields	System	Learning algorithm	Q-function estimator	Agent	Objectives
[49]	Energy management	Residential appliances	DQN, DPG	DNN	Single agent	Energy cost, load peaks
[50]	Energy management	Residential appliances	Actor-critic, Q-learning	DNN (GRU)	Single agent	Energy cost, electricity balance
[51]	Energy management	Residential appliances	Fitted Q-iteration	Randomized trees	Single agent	Electricity cost
[52]	Energy management	Hybrid electric bus	DQN	DNN	Single agent	Fuel economy
[53]	Energy management	Electric vehicle (EV)	DQN	LSTM	Single agent	Charging/discharging cost
[54]	Energy management	Hybrid EV (HEV)	DQN	DNN	Single agent	Fuel economy
[55]	Energy management	Plug-in HEV (PHEV)	DQN	MLP	Single agent	Fuel economy
[56]	Energy management	PHEV	DQN DDQN	DNN	Single agent	Fuel economy
[57]	Energy management	PHEV	DDPG	DNN	Single agent	Fuel economy
[58]	Energy management	Stand-alone microgrid	DQN	CNN	Single agent	Levelized energy cost (LEC)
[60]	Energy management	Grid-connected microgrid	Fitted Q-iteration	Randomized trees	Single agent	Self-consumption of PV
[61]	Energy management	Stand-alone microgrid	Fuzzy Q-learning	Fuzzy inference system	Multi-agent	Electricity supply, reliability
[62]	Energy management	Grid-connected microgrid	Q-learning	Associative memory	Multi-agent	Operating cost
[63]	Energy management	Residential microgrid	Q-learning	Randomized trees	Single agent	Self-consumption of PV
[64]	Energy management	Energy Internet	A3C	RNN	Single agent	Operating cost
[65]	Energy management	EH Network	DQN	DNN	Multi-agent	Sum throughput
[68]	Demand response	Smart grid	Q-learning	DNN	Single agent	Total profits
[69]	Demand response	HVAC	DQN	DNN	Single agent	Operating cost
[70]	Demand response	Building energy	Q-learning	N/A	Multi-agent	Social cost
[74]	Demand response	Smart grid	N/A	N/A	Single agent	Nash Equilibrium (NE)
[75]	Demand response	Smart grid	Q-learning	N/A	Single agent	Social welfare
[76]	Demand response	Residential loads	Q-learning	N/A	Single agent	Total cost
[71]	Demand response	Load serving entity	N/A	N/A	Single agent	Total cost
[72]	Demand response	Residential loads	Q-learning	N/A	Single agent	Total cost
[77]	Demand response	Distribution network	Q-learning	N/A	Multi-agent	Social welfare
[78]	Demand response	Smart grid	DTQ	DBN	Multi-agent	Total payoff
[73]	Demand response	Plug-in EV (PEV)	Fitted Q-iteration	Kernel function	Single agent	Charging/discharging cost
[79]	Electricity market	Distribution network	DQN	DNN	Single agent	Prosumers' benefit
[80]	Electricity market	Distribution network	Q-learning	N/A	Single agent	Customers' benefit
[81]	Electricity market	Smart grid	B-M scheme	N/A	N/A	Own averaged utility
[82]	Electricity market	Energy storage syatems	PPO	RNN	Single agent	Total profit
[83]	Electricity market	Microgrid	Q-learning	N/A	Multi-agent	System cost
[84]	Electricity market	Smart microgrid	Learning automaton	N/A	N/A	Average revenue
[85]	Electricity market	Microgrid	DQN	CNN	Single agent	Nash Equilibrium (NE)
[86]	Electricity market	Microgrid	Q-learning	N/A	Single agent	Nash Equilibrium (NE)
[87]	Electricity market	Microgrid	DQN	DNN	Single agent	Trading profits
[88]	Electricity market	Microgrid	Fuzzy Q-learning	N/A	Single agent	Social welfare
[89]	Electricity market	Microgrid	DDPG	DNN	Single agent	LSE's profit
[91]	Operational control	Interconnected grid	Policy hill-climbing	N/A	Multi-agent	Utilization of new energy
[92]	Operational control	Interconnected grid	DDRQN	DNN	Multi-agent	Utilization of new energy
[93]	Operational control	Interconnected grid	DQL	DNN	Multi-agent	Smart generation control
[94]	Operational control	Smart grid	DQL	RBM	Multi-agent	Smart generation control
[95]	Operational control	Interconnected grid	Q-learning	Deep forest	Single agent	Smart generation control
[96]	Operational control	Interconnected grid	CTQ	N/A	Multi-agent	Automatic generation control
[97]	Operational control	Windturbine	Q-learning	ANN	Single agent	Optimal yaw control
[101]	Operational control	HVAC	Q-learning	N/A	Single agent	Energy consumption
[102]	Operational control	HVAC	A3C	DNN	Single agent	Energy consumption
[98]	Operational control	stochastic power system	DRL	SDAE	Single agent	Frequency deviation
[99]	Operational control	Residential appliances	Fitted Q-iteration	CNN, LSTM	Single agent	Energy consumption
[100]	Operational control	Residential appliances	Fitted Q-iteration	CNN	Single agent	Electricity cost
[103]	Operational control	Smart grid	DQN, DDQN	CNN	Single agent	Reliability
[104]	Operational control	IEEE system	DQN	DNN	Single agent	Grid security and resiliency
[105]	Operational control	Power grid	Q-learning	ANN	Single agent	Expected profit
[106]	Operational control	Distribution grid	DQN	DNN	Single agent	Voltage stability
[107]	Operational control	IEEE system	DQN	DNN	Single agent	Voltage stability
[108]	Cyber security	Smart grid	Q-learning	N/A	Single agent	Transmission line outages
[109]	Cyber security	Smart grid	Q-learning	N/A	Single agent	Generationloss
[110]	Cyber security	AVC system	Q-learning	N/A	Single agent	Security
[111]	Economic dispatch	DG units, ES devices	Q-learning	DNN	Single agent	Operating cost
[112]	Economic dispatch	IEEE system	Q-learning	N/A	Multi-agent	Multi-objective
[113]	System optimization	Smart grid	DQN, DPG	DNN	Single agent	Total profits
[114]	Edge computing	Microgrid	DQN	DNN	Single agent	Energy consumption
[115]	Energy routing	Energy internet	Q-learning	ANN	Single agent	Efficiency, operating cost

strategy and respond to real-time electricity prices. Note that most energy management problems can be transformed into sequential decision-making problems and can be solved well by using DRL.

Residential appliances need optimal energy management strategies, all of which DRL is capable of doing. Reference [49] proposes the use of DRL in conceiving an online optimization for the scheduling of electricity consumption in residential load and aggregations of buildings. This energy management strategy can be used to provide real-time feedback to consumers to use electricity more efficiently. An optimal strategy based on DRL is proposed in [50] to min-

imize total electricity cost of residential energy management problems without knowledge about real-time household load and electricity price. In reference [51], a residential multi-carrier energy system (MCES) including a PV array, battery, heat pump and gas boiler is built, and DRL is used to develop a control strategy using real-world data to plan optimal battery scheduling.

Electric vehicle energy management problems have grown rapidly in recent years, and have caught the attention of DRL researchers. For electric vehicles, to maximize fuel economy over a specific time horizon and keep battery state of charge (SoC) stable, an energy management strategy based on deep Q-learning (DQL) is proposed in [52], with the DQL showing better performance than Q-learning in training time and convergence rate. Similarly, reference [53] proposes a model-free DRL approach to determine an optimal strategy for real-time electric vehicle (EV) charging/discharging scheduling without any system model information. A DRL-based data-driven control approach is developed in [54]–[56], and has a real-time learning architecture for a hybrid electric vehicle without any prediction or predefined rules. For the most part, it achieves substantial energy savings compared to traditional control methods. A continuous control strategy for PHEV is proposed in [57] based on DDPG and the algorithm exhibits performance close to the global optimal for dynamic programming.

Energy management is one of the main issues of a microgrid, and DRL has achieved great success in solving such problems. Some workers apply DRL to address the task of efficiently operating a hybrid storage system in a microgrid featuring photovoltaics (PV), batteries and hydrogen [58], [59]. Mbuwir, *et al*. [60] propose a batch reinforcement learning application in microgrid energy management by using state-action value functions to plan optimal scheduling for batteries. Kofinas, *et al.* [61] contribute to the application of a multi-agent reinforcement learning system to control stand-alone microgrids in order to guarantee electricity supply and operation reliability. A two-layer optimization is proposed in [62] for real-time optimal energy management (OEM) of a grid-connected microgrid. The top-layer is a model-free Q-learning for decision making and knowledge learning, the bottom-layer is a conventional convex optimization (interior point method). To maximize the self-consumption of PV production, a data-driven RL method is applied in [63] for battery energy management in a residential microgrid, by planning battery operation scheduling. In a paper by Hua *et al.* [64], an energy management problem of an Energy Internet is formulated as an optimal control problem and is solved using a DRL approach with better performance than the optimal flow method. A mean-field multi-agent DRL framework is proposed in [65] to obtain online energy control policies in large energy harvesting networks. It does not require the state information of other nodes and can achieve a performance close to the state-of-the-art centralized policies.

As mentioned above, DRL has many advantages for the problem against traditional methods: 1) DRL can achieve online optimization as well as real-time control and feedback of energy management. 2) DRL can improve efficiency of energy utilization, reduce operating costs, and increase profits. 3) DRL can complete complicated tasks with lower prior knowledge thanks to its ability to learn different levels of abstractions from data. However, energy management still has the following difficulties: 1) Wind power generation and PV have large fluctuations and many uncertainties. When considering energy storage systems and curtailable loads, the model is complex with high data dimensions. 2) Different energy storage systems have different generations, capacity, efficiency and costs, so coordinated control is difficult. 3) The charging and discharging status of electric vehicles and residential appliances is random and the information is incomplete. 4) Energy management communicates multiple energy circulations including power generation, transmission, substation, distribution and load.

Due to the above problems, DRL should focus on the following issues: 1) Transforming problems in real-world scenarios into sequence decision problems based on historical data and physical models. 2) Constructing appropriate reward functions according to the objectives and constraint conditions of the real-world issues. 3) Considering classical models and methods while using data-driven techniques.

B. Demand Response

Demand response (DR) is a typical problem in a smart grid, which keeps the balance between the electricity demand of customers and supply of utility companies by price or incentive. To improve grid stability and shift peak demand, demand response needs to incorporate consumer feedback and consumption in the control loop. DRL is thus an effective optimal control approach with data-driven support models to solve such problems [66], [67].

For power customers, minimizing costs is the primary objective, while for utility companies, it is maximizing profits. To solve a real-time incentive-based demand response problem, a DRL approach is proposed in [68] by assisting the service provider in purchasing electricity from various consumers to balance power fluctuations and keep grid reliability. This is an alogous to that by Zhang [69], which puts forward a DRL approach to make sequential optimal decisions in heating, ventilation and air conditioning (HVAC) systems under demand response. In [70], an autonomous and optimal HVAC electricity consumption scheduling is planned by multi-agent RL to minimize the social cost of a game-theoretic methodology. An optimal pricing scheme for a demand response program based on RL is developed in [71], and the balance between exploration and exploitation in learning processes leads to better performance in load serving entity (LSE). An optimal model of residential load scheduling solving by RL is presented in [72] and considers consumer satisfaction, stochastic renewable energy and cost. Furthermore, the model can be made more general. Reference [73] proposes a novel demand response approach to reduce the long-term charging/discharging cost of plug-in electric vehicles by batch RL and a Bayesian neural network.

It is possible for DRL to build a game model between power companies and customers considering demand response. A two-stage game model between power companies and cus-

tomers is proposed in [74] and solved by RL. At the first stage, the customers' optimal power consumption is obtained and at the second stage, the power companies' prices are determined. A dynamic pricing demand response model is proposed in [75], considering the service provider's profit and customers' costs. The retail price is determined by RL according to electricity demand and wholesale electricity prices. A multi-agent RL is used for a decentralized control method in [76] to determine an optimal bidding strategy between power companies and customers considering demand response. Similarly, Babar *et al.* [77] propose an applied data-driven RL methodology by complex bidding rules for agile demand response in an unbundled electricity market. A virtual leader-follower Stackelberg game model based on deep transfer Q-learning (DTQ) is proposed in [78] to maximize the total payoff of smart grid agents.

From the above references it is clear that DRL has advantages since: 1) DRL can make decisions based on incomplete information, and such decisions can be online. 2) Through game theory, DRL can achieve maximum system benefits and reduce transaction costs. 3) DRL has a stronger transfer capability and can be applied to many different scenarios. At the same time, the challenges of demand response are reflected in the following aspects: 1) Incentivemeasures are diverse in form and include economic, technological, environmental and political incentives (different users respond differently to incentives.) 2) Demand response is usually accompanied by changes in various factors such as load and electricity price, and different factors change the results differently. 3) The control methods and constraint conditions of the electrical equipment involved in the response are different, thus making the model more complicated. 4) The process of demand response is often accompanied by the game process between consumers, service providers and power companies, so that the optimization objectives are different.

To overcome the above challenges, DRL approaches need to considerthe following issues: 1) Using DNN and other methods to extract consumers' behavior characteristics and predicting their behavior as the basis for optimal control. 2) Choosing the appropriate state space, including price, load, SoC of the storage system, etc. 3) Making full use of historical data and consumers' feedback to compensate for the lack of models.

C. Electricity Market

A hierarchical electricity market can be divided into a wholesale electricity market and a retail electricity market. It combines service providers with power companies and customers by information and power, as shown in Fig. 5. The wholesale electricity market exists when competing power companies offer their electricity to retailers which then sell the electricity to the service provider. Meanwhile, the retail electricity market exists when customers choose their suppliers from competing electricity retailers. Trading among these elements is a complex game problem and DRL can be used to obtain optimal strategies under incomplete information.

For service providers, power companies or customers, an optimal bidding strategy means more benefit and lower cost.

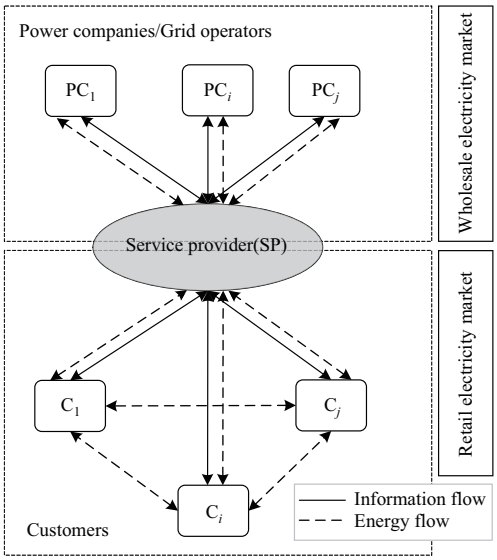

Fig. 5. Hierarchical electricity market model.

An event-driven electricity market is proposed in [79] for energy trading in local distribution networks, and prosumers' trading strategies are determined by DRL to maximize their benefit. Chen and Su [79], [80] study an indirect customer-to-customer electricity market in a distribution network, and RL is used to determine energy trading strategies. A constrained energy trading game among end-consumers is proposed in [81] by adaptive RL with incomplete information, and finally, bidding strategy converges to Nash equilibrium. A DRL based algorithm is proposed in [82] for ESSs to arbitrage in real-time electricity markets under price uncertainty, and electricity price information is extracted by EMA filter and RNN. An electricity market model with dynamic pricing and energy consumption in a microgrid is studied in [83], and in the model, RL is applied to reduce system cost for the service provider.

For the whole system, equilibrium and social welfare are objectives in game theory. Amulti-leaders and multi-followers Stackelberg game model for energy trading is developed in [84], and RL is used to obtain equilibrium under a privacy policy. Xiao *et al.* [85], [86] propose a microgrid energy trading game model considering renewable energy generation and demand, battery level and trading history, and the Nash equilibrium is obtained by DRL approach. Furthermore, a continuous real-time electricity market in a microgrid is built in [87], and a DRL approach with discrete high-level actions of action spaces is applied to obtain optimal trading cost. An hour-ahead electricity market model in continuous renewable power penetration is proposed in [88], and an IEEE 30-bus test system is implemented by fuzzy Q-learning. A hierarchical electricity market with bidding and pricing of load serving entity (LSE) is proposed in [89], dynamical bid and price response functions are learned by DNN and state transition samples are generated by a deep deterministic policy gradient (DDPG) algorithm.

To sum up, compared to conventional methods, DRL has the following strengths: 1) Most DRL algorithms are model-

free and suitable for scenarios that cannot be formulated into models. 2) Function approximators such as neural networks can extract more data features that are not considered by the models. 3) Nash equilibrium can be achieved by DRL between supply and demand sides in electricity markets. On the other hand, the electricity market has the following major difficulties: 1) There are multiple entities in the hierarchical electricity market, and their objectives are different, making reward functions difficult. 2) In addition to the energy flow, there is also incomplete information between the entities, so data-driven methods are needed. 3) Energy trading is a continuous decision problem, unlike typical discrete decision problems and requires real-time decision making.

Facing these problems, the following aspects should be the research objectives of DRL: 1) Using game theory models to construct different market entities as different game entities. 2) Using multi-agent RL, and agents corresponding to different game entities. 3) Due to the complexity of the game process, research should start from small-scale scenarios and gradually expand the scale of the scenarios. 4) Improving the ability to integrate and extract information such as price and energy.

D. Operational Control

The operational control problem is a classical power system problem. As renewable energysource become more prevalent, operational control becomes more complex and challenging. Control strategies and optimization decisions can be online learning by DRL under large-scale scenarios and with limited information [90].

Power generation side control is an essential aspect of operational control. A smart generation control scheme of multi-area interconnected grids is proposed in [91], [92]. DRL can obtain an optimal strategy facing complex operating environments, which cannot be solved by conventional centralized automatic generation control (AGC). Reference [93] and [94] present a new architecture DQL algorithm for the first time and use the proposed algorithm to design a smart generation controller for multi-agent systems with high-level robustness. Similarly, a preventive strategy for smart generation control (SGC) under large continuous disturbances is proposed in [95]. Deep forest reinforcement learning (DFRL) proves more effective than traditional AGC. Reference [96] develops a two-layer decentralized generation commanddispatch (GCD) scheme of automatic generation control (AGC). The first layer is to obtain generation command and the second layer is to exploit prior knowledge for optimal control by consensus transfer Q-learning (CTQ). A novel optimal yaw control method based on RL is proposed in [97], and ANN is used to avoid large matrix quantification problem. In order to achieve the quantity of instant wind turbine power and orientation variation, this method also considers the constraints of the mechanical limitations in the yawing system and the mechanical loads.

Load control is another critical aspect of operational control. A model-free load frequency control (LFC) approach in continuous action space by DRL is proposed in [98]. Faster response speed and stronger adaptability are obtained to minimize frequency deviation. While in reference [99], a load near-optimal control problem under sparse observations is proposed. The problem is solved by DRL with function approximators CNN and LSTM, with LSTM having a higher performance than CNN. Similarly, Claessens *et al.* [100] use CNN as a function approximator to extract hidden state features of the residential load. The high-dimensional load control problem is solved by fitted Q-iteration to reduce electricity cost.

Apart from generation and load control, appliances and system control are also complex decision-making problems. To utilize natural ventilation for an HVAC system, Chen *et al*. [101] introduce an RL control approach which works better than conventional rule-based heuristic control. Similarly, Zhang *et al*. [102] develop building a HVAC system optimal control model for energy efficiency and thermal comfort. A3C algorithmis applied for DRL training under high-dimension data and continuous space. A smart grid emergency control strategy is proposed in [103] during the transient process, DQN and double DQN are applied under a limited information scenario. Similarly, Huang *et al*. [104] present a DRL-based emergency control scheme. Grid security and resiliency are improved online with the non-linear generalization and high-dimensional feature extraction abilities. A power grid's operation and maintenance problemsare developed in [105]. ANN is used to replace the tabular representation of value function and RL exploits the information about the state and components of the grid. A real-time two-timescale voltage control strategy is proposed in [106]. Active power generation dynamic and load consumption are modeled as MDPs and optimized by DRL to minimize bus voltage deviations. Diao *et al.* [107] propose a novel Grid Mind framework to mitigate voltage issues effectively. Autonomous grid operational control policies can be learned by DRL through interactions with offline simulations.

In conclusion, DRL has the following advantages on operational control: 1) Continuous control can be achieved by DRL under continuous state and action spaces. 2) DRL can make the control system more automated with incomplete information. 3) DRL can handle some unpredictable emergencies, which is beyond the competence of most traditional methods. There remain however some problems of operational control in the following respects: 1) Device control needs to be combined with the device's physical structure and operating conditions. 2) Operational control should consider both steady state and transient stability, so different time scales need to be considered. 3) For the power generation system, it is necessary to consider the synchronous operation state of the unit. For the power generation system and the power consumption system, it is also necessary to consider the voltage and frequency stability.

To solve these problems, the following directions should be considered: 1) Combining DRL with classic control methods and strategies to avoid system failures. 2) Adoptinga hierarchical strategy, with one layer using control strategy and one layer using optimization strategies. 3) Convergence of grid data features and device model features for more intelligent and flexible control.

E. Others

As research continues, DRL is also being increasingly applied in other areas of power systems, such as cyber security, economic dispatch and system optimization etc. When solving these problems, the DRL method is highly suitable for analyzing dynamic behaviors, complicated scenariosand uncertain constraints.

A multi-stage game between cyber attacker and defender based on RL is proposed in [108]. To identify the optimal attack sequences, the attacker learns the sequence of attackactions and the defenderlearns to defend against it. Meanwhile, Paul and Ni [109] compare RL and linear programming for smart grid security problems. Linear programming determines the attacker's and defender's mixed action strategies and probability, while RL obtains the optimal attack action in the presence of a static defender's action for the one-shot game. In reference [110], RL is used to obtain the optimal attack strategy for the attacker with incomplete information, which can help maintain the security of an automatic voltage control (AVC) system.

Furthermore, a cooperative RL approach which can avoid the complexity of model and dimensionality is utilized in [111] to solve a distributed economic dispatch problem. This approach minimizes operation cost while considering power balance, capacity, and operational constraints. Similarly, a bacteria foraging RL method is proposed in [112] to deal with economic dispatch problems under system uncertainties and knowledge transfer is implied to increase efficiency. A smart grid system optimization problem with arigorous physical model is solved in [113] by DQN, DPG and mixed-integer linear programming (MILP). Furthermore, a multi-access edge computing (MEC) problem of a microgrid in [114] is divided into an energy-efficient assignment problem and energy supply plan problem. Model-based DRL is used to solve the second subproblem with the input of the first subproblem. An optimal routing problem in an Energy Internet is proposed in [115], and a DRL algorithm is used to improve energy efficiency and reduce the operating and environmental cost.

According to the previous discussion, DRL has following characteristics: 1) DRL can be combined with many traditional methods to achieve better results. 2) DRL is suitable for scenarios that are not easy to model, such as cyber security and emergency control. 3) DRL can deal with data with larger scale and higher dimension under a complicated system. In addition, the optimization and control problems of a power system are omnifarious and have at least the following difficulties: 1) The situation facing cyber security is complex, and the attacks received by the power grid are diverse and unpredictable. 2) Economic dispatch and unit commitment problems need to be considered in parallel with planning and operation, making the optimization process more complicated. 3) In system optimization and edge calculation problems, variables and constraints are complex and calculation is large, so the algorithms are more demanding.

To solve the problems, the following points need attention: 1) Combining model-driven approaches with data-driven approaches, extracting features from historical data, and determining state space and action space through models. 2) Studying the basic principles and generalization performance of DRL and updating and improving the algorithms' ability to adapt and solve problems. 3) Starting from smaller scale optimization and control issues, gradually studying more complex scenarios and developing into a smart grid. 4) Integrating domain knowledge such as cybernetics, game theory, mathematical optimization, computer science and further abstracting specific problems into general mathematical and engineering problems.

IV. Conclusion

Over the past few years, DRL has achieved rapid development in solving sequential decision-making problems, around theoretical, methodological and experimental fields. In particular, DL obtains the object's attributes, categories, or characteristics from the environment, while RL makes decisions to control strategies according to the information. Therefore, DRL can solve problems in large, high-dimensional states and action spaces.

With further research and the development of smart grids, power systems evolve and face new challenges with the integration of renewable energy and deepening of marketization. Traditional methods face many difficulties in solving these problems in the power system, so there is an increasing need for AI methods such as DRL. This paper reviews the principle, development, algorithms, and characteristics of DRL and its applications in the power system, including energy management, demand response, electricity market, and operational control. It also briefly summarizes the implements and processes of application in different scenarios and makes statistics and comparisons on the key information of the papers. There are still many issues to be discussed.

A. Application Landing Problem

So far, the application of DRL in the power system has rarely been ractical or commercially viable. On the one hand, DRL theory is still not perfect and is still in the exploration and verification stage. On the other hand, the power system has higher requirements for the reliability and stability of the control method, but the current machine learning method is still based on probability and statistical law.

In order to promote the practical applications of DRL in the power system, firstly the theory and modelling of DRL algorithmsneed to be perfected to improve DRL reliability and robustness. It is then vital to run related models in small-scale scenarios to accumulate data and experience, and gradually expand to large-scale scenarios. In addition, relevant research and industry developments require the support of market policy and promotion.

B. Main Issues and Key Technologies

Despite its successful models and applications, DRL still suffers problems and challenges when it comes to real-world implementation. This is especially the case in power systems where there are still many problems to be solved.

A Go game has strict rules, and the return of each action can be accurately calculated and evaluated. However, when compared to the Go game, a power system is more complicated, and has many uncertainties. For example, renewable energy generation cannot be described and predicted with accurate models; equipment failure usually occurs suddenly; electricity market trading behaviors and strategies are also diverse. In view of the above situation, DRL requires multiple designs and modifications to adapt to different scenarios.

Firstly, the reward function needs to be designed according to practical problems and will determine the direction and efficiency of learning. Secondly, function approximators need to be chosen carefully, especially for complicated scenarios. Thirdly, state and action spaces represent the scale and complexity of the system model, so they require thoughtful planning. Finally, the tradeoff between exploration and exploitation cannot be ignored, so initialization and parameters need to be set carefully.

C. Future Development of DRL

In terms of theory and applications, current DRL is at the first stage of artificial intelligence, or artificial narrow intelligence (ANI). It still faces theoretical, technological, economic, societal and ethical challenges, even though it has achieved success in some cases. With the development of information and communications technologies, DRL will be further developed and applied.

Regarding the future development of DRL, at least the following aspects need to be considered. The first is the generalization ability. DRL will develop popular trends of taking explicit algorithms into a specific form of the neural network so that it can be trained end-to-end and be more suitable for reasoning on an abstract level. The second is the transfer learning ability. This would allow DRL to learn complicated decision-making problems in simulations and then apply the learned information in real-world environments. The third is meta-learning and lifelong learning ability. This would improve performance and shorten training time with previous knowledge and information. Besides, multi-agent methods are also significant. This would allow DRL to deal with multiple subjects, which is closer to the real world.

In conclusion, DRL and its applications in the power system still face many opportunities and challenges. These will cause more attention and research, and there will inevitably be more surprising developments in the future.

REFERENCES

[1] D. X. Zhang, X. Q. Han, and C. Y. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids, " *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[2] M. L. Tuballa and M. L. Abundo, "A review of the development of Smart Grid technologies, " *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 710–725, Jun. 2016.

[3] R. C. Qiu and P. Antonik, *Smart Grid and Big Data*: *Theory and Practice*, New York: Wiley Publishing, 2017.

[4] X. He, Q. Ai, R. C. Qiu, W. T. Huang, L. J. Piao, and H. C. Liu, "A big data architecture design for smart grids based on random matrix theory, " *IEEE Transactionson Smart Grid*, vol. 8, no. 2, pp. 674–686, Mar. 2017.

[5] T. Yu, B. Zhou, and W. G. Zhen, "Application and development of reinforcement learning theory in power systems, " *Power System Protection and Control*, vol. 37, no. 14, pp. 122–128, Jul. 2009.

[6] E. Mocanu, "Machine learning applied to smart grids, " Ph. D. dissertation, Department, TechnischeUniversiteit Eindhoven, Eindhoven, 2017.

[7] M. F. Zia, E. Elbouchikhi, and M. Benbouzid, "Microgrids energy management systems: a critical review on methods, solutions, and prospects, "*Applied Energy*, vol. 222, pp. 1033–1055, Jul. 2018.

[8] X. He, L. Chu, R. C. M. Qiu, Q. Ai, and Z. N. Ling, "A novel data-driven situation awareness approach for future grids—Using large random matrices for big data modeling, " *IEEE Access*, vol. 6, pp. 13855–13865, Mar. 2018.

[9] L. Chu, R. Qiu, X. He, Z. N. Ling, and Y. D. Liu, "Massive streaming pmu data modelling and analytics in smart grid state evaluation based on multiple high-dimensional covariance test, " *IEEE Transactionson Big Data*, vol. 4, no. 1, pp. 55–64, Mar. 2018.

[10] R. Qiu, L. Chu, X. He, Z. N. Ling, and H. C. Liu, "Spatiotemporal big data analysis for smart grids based on random matrix theory, " in *Transportation and Power Grid in Smart Cities*: *Communication Networks and Services*, H. T. Mouftah, M. Erol-Kantarci, and M. H. Rehmani, Eds. John Wiley & Sons Ltd, 2018, pp. 591–633.

[11] R. C. Qiu, X. He, L. Chu, and Q. Ai, "Big data analysis of power grid from random matrix theory, " Institution of Engineering and Technology (IET) in Smarter Energy: From Smart Metering to the Smart GridSmarter Energy: From Smart Metering to the Smart Grid, 2016, pp. 381–425, doi:10.1049/pbpo088e_ch13.

[12] M. J. Han, R. May, X. X. Zhang, X. R. Wang, S. Pan, D. Yan, Y. Jin, and L. G. Xu, "A review of reinforcement learning methodologies for controlling occupant comfort in buildings, "*Sustainable Cities and Society*, vol. 51, pp. 101748, Nov. 2019.

[13] M. Ding and X. Q. Yin, "A review on multi-agent technology in microgrid control, " *Electronics Science Technology and Application*, vol. 5, no. 1, pp. 1–13, 2018.

[14] L. F. Cheng and T. Yu, "A new generation of AI: a review and perspective on machine learning technologies applied to smart energy and electric power systems, " *International Journal of Energy Research*, vol. 43, no. 6, pp. 1928–1973, May 2019.

[15] S. R. Sutton and A. G. Barto, *Reinforcement Learning*: *An Introduction*, 2nd ed., Cambridge: MIT Press, 2018.

[16] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, andJ. Pineau, "An introduction to deep reinforcement learning, " *Foundations and Trends in Machine Learning*, vol. 11, no. 3–4, pp. 219–354, Dec. 2018.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning, " *Nature*, vol. 518, no. 7540, pp. 529–533, Jan. 2015.

[18] 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.

[19] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. J. 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.

[20] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation, " arXiv: 1808.00177 (2018).

[21] A. A. Rusu, M. Vecerík, T. Rothörl, N. Heess, R. Pascanu, and R. Hadsell, "Sim-to-real robot learning from pixels with progressive nets, " in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017.

[22] W. Y. Wang, J. W. Li, and X. D. He.(2018). Deep reinforcement learning for NLP. [Online]. Available: https://www.aclweb.org/anthology/P18-5007.pdf.

[23] B. McCann, N. S. Keskar, C. M. Xiong, and R. Socher, "The natural language decathlon: multitask learning as question answering, " arXiv: 1806.08730 (2018).

[24] Y. Deng, F. Bao, Y. Y. Kong, Z. Q. Ren, and Q. H. Dai, "Deep direct reinforcement learning for financial signal representation and trading, " *IEEE Transactionson Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 653–664, Mar. 2017.

[25] Z. H. Hu, Y. T. Liang, J. Zhang, Z. Li, and Y. Liu, "Inference aided reinforcement learning for incentive mechanism design in crowdsourcing, " in *Advances in Neural Information Processing Systems*, 2018.

[26] H. W. Wang, C. J. Li, J. Y. Li, X. He, and T. W. Huang, "A survey on distributed optimisation approaches and applications in smart grids, " *Journal of Control and Decision*, vol. 6, no. 1, pp. 41–60, Nov. 2019.

[27] W. B. Shi, N. Li, C. C. Chu, and R. Gadh, "Real-time energy management in microgrids, " *IEEE Transactionson Smart Grid*, vol. 8, no. 1, pp. 228–238, Jan. 2017.

[28] M. Zachar and P. Daoutidis, "Microgrid/macrogrid energy exchange: a novel market structure and stochastic scheduling, "*IEEE Transactionson Smart Grid*, vol. 8, no. 1, pp. 178–189, Jan. 2017.

[29] C. Ordoudis, P. Pinson, and J. M. Morales, "An integrated market for electricity and natural gas systems with stochastic power producers, "*European Journal of Operational Research*, vol. 272, no. 2, pp. 642–654, Jan. 2019.

[30] L. Zéphyr and C. L. Anderson, "Stochastic dynamic programming approach to managing power system uncertainty with distributed storage, " *Computational Management Science*, vol. 15, no. 1, pp. 87–110, Jan. 2018.

[31] J. L. Duchaud, G. Notton, C. Darras, and C. Voyant, "Power ramp-rate control algorithm with optimal State of Charge reference via Dynamic Programming, " *Energy*, vol. 149, pp. 709–717, Apr. 2018.

[32] H. T. Nguyen, L. B. Le, and Z. Y. Wang, "A bidding strategy for virtual power plants with the intraday demand response exchange market using the stochastic programming, " *IEEE Transactionson Industry Applications*, vol. 54, no. 4, pp. 3044–3055, July-Aug. 2018.

[33] A. I. Mahmutogullari, S. Ahmed, O. Cavus, and M. S. Akturk, "The value of multi-stage stochastic programming in risk-averse unit commitment under uncertainty, " arXiv:1808.00999, (2018).

[34] Megantoro, Prisma, F. D. Wijaya, and E. Firmansyah, "Analyze and optimization of genetic algorithm implemented on maximum power point tracking technique for PV system, " in *Proceedings of 2017 International Seminar on Application for Technology of Information and Communication*, 2017.

[35] I. E. S. Naidu, K. R. Sudha, and A. C. Sekhar, "Dynamic stability margin evaluation of multi-machine power systems using genetic algorithm, " in *International Proceedingson Advances in Soft Computing, IntelligentSystemsandApplications*, M. S. Reddy, K. Viswanath, and S. P. K. M, Eds. Singapore: Springer, 2018.

[36] G. N. Nguyen, K. Jagatheesan, A. S. Ashour, B. Anand, and N. Dey, "Ant colony optimization based load frequency control of multi-area interconnected thermal power system with governor dead-band nonlinearity, " in *Smart Trends in Systems, Security and Sustainability*, X. S. Yang, A. K. Nagar, and A. Joshi, Eds. Singapore: Springer, 2018.

[37] R. Srikakulapu and V. U, "Optimized design of collector topology for offshore wind farm based on ant colony optimization with multiple travelling salesman problem, " *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 6, pp. 1181–1192, Nov. 2018.

[38] H. Li, D. Yang, W. Z. Su, J. H. Lü, X. H. Yu, "An overall distribution particle swarm optimization MPPT algorithm for photovoltaic system under partial shading, " *IEEE Transactionson Industrial Electronics*, vol. 66, no. 1, pp. 265–275, Jan. 2019.

[39] H. J. Gu, R. F. Yan, and T. K. Saha, "Minimum synchronous inertia requirement of renewable power systems, " *IEEE Transactionson Power Systems*, vol. 33, no. 2, pp. 1533–1543, Mar. 2018.

[40] Y. X. Li, "Deep reinforcement learning, " arXiv: 1810.06339 (2018).

[41] D. P. Bertsekas. Reinforcement learning and optimal control. [online]. Available: http://www.athenasc.com/.

[42] Z. R. Yang, Y. C. Xie, and Z. R Wang, "A theoretical analysis of deep Q-learning, " arXiv: 1901.00137v1 (2019).

[43] F. Agostinelli, G. Hocquet, S. Singh, and P. Baldi, "From reinforcement learning to deep reinforcement learning: an overview, " in *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State*, L. Rozonoer, B. Mirkin, and I. Muchnik, Eds. Cham: Springer, 2018, pp. 298–328.

[44] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*, Boca Raton: CRC Press, Inc., 2010.

[45] C. Szepesvari, "Algorithms for reinforcement learning, "in*Wiley Encyclopedia of Operations Research and Management Science*, J. J. Cochran, L. A. Cox Jr, P. Keskinocak, J. P. Kharoufeh, and J. C. Smith, Eds. New York: Wiley, 2011.

[46] M. Wiering and M. Van Otterlo, "Reinforcement learning, " *Adaptation, Learning, and Optimization*, vol. 12, pp. 51, 2012.

[47] M. Lapan, *Deep Reinforcement Learning Hands-On*: *Apply Modern RL Methods, with Deep Q-Networks, Value Iteration, Policy Gradients, TRPO, AlphaGo Zero and More*, UK: Packt Publishing Ltd, 2018.

[48] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, "Bridging the gap between value and policy based reinforcement learning, " in *Advances in Neural Information Processing Systems*, 2017.

[49] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning, " *IEEE Transactionson Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.

[50] Z. Q. Wan, H. P. Li, and H. B. He, "Residential energy management with deep reinforcement learning, " in *Proceedings of 2018 International Joint Conference on Neural Networks*, 2018.

[51] B. V. Mbuwir, M. Kaffash, and G. Deconinck, "Battery scheduling in a residential multi-carrier energy system using reinforcement learning, " in *Proceedings of 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids*, 2018.

[52] J. D. Wu, H. W. He, J. K. Peng, Y. C. Li, and Z. J. Li, "Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus, " *Applied Energy*, vol. 222, pp. 799–811, Jul. 2018.

[53] Z. Q. Wan, H. P. Li, H. B. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning, " *IEEE Transactionson Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[54] Y. Hu, W. M. Li, K. Xu, T. Zahid, F. Y. Qin, and C. M. Li, "Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning, " *Applied Sciences*, vol. 8, No. 2, pp. 187, Jan. 2018.

[55] X. W. Qi, Y. D. Luo, G. Y. Wu, K. Boriboonsomsin, and M. J. Barth, "Deep reinforcement learning-based vehicle energy efficiency autonomous learning system, " in *Proceedings of 2017 IEEE Intelligent Vehicles Symposium*, 2017.

[56] X. W. Qi, Y. D. Luo, G. Y. Wu, K. Boriboonsomsin, and M. Barth, "Deep reinforcement learning enabled self-learning control for energy efficient driving, " *Transportation Research Part C*: *Emerging Technologies*, vol. 99, pp. 67–81, Feb. 2019.

[57] Y. K. Wu, H. C. Tan, J. K. Peng, H. L. Zhang, and H. W. He, "Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus, " *Applied Energy*, vol. 247, pp. 454–466, Aug. 2019.

[58] V. François-Lavet, D. Taralla, D. Ernst, and R. Fonteneau, "Deep reinforcement learning solutions for energy microgrids management, " in *Proceedings of European Workshop on Reinforcement Learning*, 2016.

[59] V. François-Lavet, "Contributions to deep reinforcement learning and its applications in smartgrids, " Ph. D. dissertation, Department, Université de Liège, Liège, Belgique, 2017.

[60] B. V. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck, "Battery energy management in a microgrid using batch reinforcement learning, " *Energies*, vol. 10, no. 11, pp. 1846, Nov. 2017.

[61] P. Kofinas, A. I. Dounis, and G. A. Vouros, "Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids, " *Applied Energy*, vol. 219, pp. 53–67, Jun. 2018.

[62] Z. K. Tan, X. S. Zhang, B. M. Xie, D. Z. Wang, B. Liu, andT. Yu, "Fast learning optimiser for real-time optimal energy management of a grid-connected microgrid, " *IET Generation, Transmission & Distribution*, vol. 12, no. 12, pp. 2977–2987, Jun. 2018.

[63] B. V. Mbuwir, F. Spiessens, and G. Deconinck, "Self-learning agent for battery energy management in a residential microgrid, " in *Proceedings of 2018 IEEE PES Innovative Smart Grid Technologies Conference Europe*, 2018.

[64] H. C. Hua, Y. C. Qin, C. T. Hao, and J. W. Cao, "Optimal energy management strategies for energy Internet via deep reinforcement learning approach, " *Applied Energy*, vol. 239, pp. 598–609, Apr. 2019.

[65] M. K. Sharma, A. Zappone, M. Debbah, and M. Assaad, "Multi-agent deep reinforcement learning based power control for large energy harvesting networks, " in *Proceedings of 17th International Symposium on Modeling and Optimization in Mobile*, 2019.

[66] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: a review of algorithms and modeling techniques, " *Applied Energy*, vol. 235, pp. 1072–1089, Feb. 2019.

[67] P. Siano, "Demand response and smart grids—A survey, " *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, Feb. 2014.

[68] R. Z. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network, " *Applied Energy*, vol. 236, pp. 937–949, Feb. 2019.

[69] X. Y. Zhang, "A data-driven approach for coordinating air conditioning units in buildings during demand response events, " Ph. D. dissertation, Department, Virginia Tech, Virginia, 2019.

[70] J. Hao, "Multi-agent reinforcement learning embedded game for the optimization of building energy control and power system planning, " arXiv: 1901.07333v1 (2019).

[71] A. Ghasemkhani and L. Yang, "Reinforcement learning based pricing for demand response, " in *Proceedings of 2018 IEEE International Conference on Communications Workshops*, 2018.

[72] T. Remani, E. A. Jasmin, and T. P. ImthiasAhamed, "Residential load scheduling with renewable generation in the smart grid: a reinforcement learning approach, " *IEEE Systems Journal*, vol. 13, no. 3, pp. 3283–3294, Sep. 2019.

[73] A. Chiş, J. Lundén, and V. Koivunen, "Reinforcement learning-based plug-in electric vehicle charging with forecasted price, " *IEEE Transactionson Vehicular Technology*, vol. 66, no. 5, pp. 3674–3684, May 2017.

[74] P. A. Apostolopoulos, E. E. Tsiropoulou, and S. Papavassiliou, "Demand response management in smart grid networks: a two-stage game-theoretic learning-based approach, " *Mobile Networks and Applications*, pp. 1–14, Oct. 2018.

[75] R. Z. Lu, S. H. Hong, and X. F. Zhang, "A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach, " *Applied Energy*, vol. 220, pp. 220–230, Jun. 2018.

[76] S. Najafi, S. Talari, A. S. Gazafroudi, M. Shafie-khah, J. M. Corchado, and J. P. S. Catalão, "Decentralized control of DR using a multi-agent method, " in *Sustainable Interdependent Networks*, M. H. Amini, K. G. Boroojeni, S. S. Iyengar, P. M. Pardalos, F. Blaabjerg, and A. M. Madni, Eds. Cham: Springer, 2018, pp. 233–249.

[77] M. Babar, P. H. Nguyen, V.Ćuk, I. G. Kamphuis, M. Bongaerts, and Z. Hanzelka, "The evaluation of agile demand response: An applied methodology, " *IEEE Transactionson Smart Grid*, vol. 9, no. 6, pp. 6118–6127, Nov. 2018.

[78] X. S. Zhang, T. Bao, B. Yang, and C. J. Han, "Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid, " *Energy*, vol. 133, pp. 348–365, Aug. 2017.

[79] T. Chen and W. C. Su, "Local energy trading behavior modeling with deep reinforcement learning, " *IEEE Access*, vol. 6, pp. 62806–62814, Oct. 2018.

[80] T. Chen and W. C. Su, "Indirect customer-to-customer energy trading with reinforcement learning, " *IEEE Transactionson Smart Grid*, vol. 10, no. 4, pp. 4338–4348, Jul. 2018.

[81] H. W. Wang, T. W. Huang, X. F. Liao, H. Abu-Rub, and G. Chen, "Reinforcement learning for constrained energy trading games with incomplete information, " *IEEE Transactionson Cybernetics*, vol. 47, no. 10, pp. 3404–3416, Oct. 2017.

[82] H. C. Xu, X. Li, X. Y. Zhang, and J. B. Zhang, "Arbitrage of energy storage in electricity markets with deep reinforcement learning, " arXiv: 1904.12232 (2019).

[83] B. G. Kim, Y. Zhang, M. van der Schaar, and J. W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning, " *IEEE Transactionson Smart Grid*, vol. 7, no. 5, pp. 2187–2198, Sep. 2016.

[84] H. W. Wang, T. W. Huang, X. F. Liao, H. Abu-Rub, and G. Chen, "Reinforcement learning in energy trading game among smart microgrids, " *IEEE Transactionson Industrial Electronics*, vol. 63, no. 8, pp. 5109–5119, Aug. 2016.

[85] L. Xiao, X. Y. Xiao, C. H. Dai, M. Pengy, L. C. Wang, and H. V. Poor, "Reinforcement learning-based energy trading for microgrids, " arXiv: 1801.06285 (2018).

[86] X. Y. Xiao, C. H. Dai, Y. D. Li, C. H. Zhou, and L. Xiao, "Energy trading game for microgrids using reinforcement learning, " in *International Conferenceon Game Theory for Networks*, L. J. Duan, A. Sanjab, H. S. Li, X. Chen, D. Materassi, and R. Elazouzi, Eds. Cham: Springer, 2017.

[87] I. Boukas, D. Ernst, and B. Cornélusse, "Real-time bidding strategies from micro-grids using reinforcement learning, " in *Proceedings of CIRED Workshop 2018*, 2018.

[88] R. S. Salehizadeh and S. Soltaniyan, "Application of fuzzy Q-learning for electricity market modeling by considering renewable power penetration, " *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 1172–1181, Apr. 2016.

[89] H. C. Xu, H. B. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, "Deep reinforcement learning for joint bidding and pricing of load serving entity, " *IEEE Transactions on Smart Grid*, Vol. 10, no. 6, pp. 6366–6375, Nov. 2019.

[90] M. Glavic, R. Fonteneau, and D. Ernst, "Reinforcement learning for electric power system decision and control: past considerations and perspectives, " *IFAC-Papers OnLine*, vol. 50, no. 1, pp. 6918–6927, Jul. 2017.

[91] L. Xi, J. F. Chen, Y. H. Huang, Y. C. Xu, L. Liu, Y. M. Zhou, and Y. D. Li, "Smart generation control based on multi-agent reinforcement learning with the idea of the time tunnel, " *Energy*, vol. 153, pp. 977–987, Jun. 2018.

[92] L. Xi, J. F. Chen, Y. H. Huang, T. L. Xue, T. Zhang, Y. N. Zhang, "Smart generation control based on deep reinforcement learning with the ability of action self-optimization, " *Scientia Sinica Information is*, vol. 48, no. 10, pp. 1430–1449, Oct. 2018.

[93] L. F. Yin, T. Yu, and L. Zhou, "Design of a novel smart generation controller based on deep q learning for large-scale interconnected power system, " *Journal of Energy Engineering*, vol. 144, no. 3, pp. 04018033, Jun. 2018.

[94] L. F. Yin and T. Yu, "Design of strong robust smart generation controller based on deep Qlearning, " *Electric Power Automation Equipment*, vol. 38, no. 5, pp. 12–19, 2018.

[95] L. F. Yin, L. L. Zhao, T. Yu, and X. S. Zhang, "Deep forest reinforcement learning for preventive strategy considering automatic generation control in large-scale interconnected power systems, " *Applied Sciences*, vol. 8, no. 11, pp. 2185, Oct. 2018.

[96] X. S. Zhang, Q. Li, T. Yu, and B. Yang, "Consensus transfer Q-learning for decentralized generation command dispatch based on virtual generation Tribe, " *IEEE Transactionson Smart Grid*, vol. 9, no. 3, pp. 2152–2165, May 2018.

[97] A. Saenz-Aguirre, E. Zulueta, U. Fernandez-Gamiz, J. Lozano, and J. M. Lopez-Guede, "Artificial neural network based reinforcement learning for wind turbine yaw control, " *Energies*, vol. 12, no. 3, pp. 436, Jan. 2019.

[98] Z. M. Yan and Y. Xu, "Data-driven load frequency control for stochastic power systems: a deep reinforcement learning method with continuous action search, " *IEEE Transactionson Power Systems*, vol. 34, no. 2, pp. 1653–1656, Mar. 2018.

[99] F. Ruelens, B. J. Claessens, P. Vrancx, F. Spiessens, and G. Deconinck, "Direct load control of thermostatically controlled loads based on sparse observations using deep reinforcement learning, " arXiv: 1707.08553 (2017).

[100] B. J. Claessens, P. Vrancx, and F. Ruelens, "Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control, " *IEEE Transactionson Smart Grid*, vol. 9, no. 4, pp. 3259–3269, Jul. 2018.

[101] Y. J. Chen, L. K. Norford, H. W. Samuelson, and A. Malkawi, "Optimal control of HVAC and window systems for natural ventilation through reinforcement learning, " *Energy and Buildings*, vol. 169, pp. 195–205, Jun. 2018.

[102] Z. A. Zhang, A. Chong, Y. Q. Pan, C. L. Zhang, S. L. Lu, and K. P. Lam, "A deep reinforcement learning approach to using whole building energy model for HVAC optimal control, " in *Proceedings of 2018 Building Performance Modeling Conference and Sim Build*, 2018.

[103] W. Liu, D. X. Zhang, X. Y. Wang, J. X. Hou, and L. P. Liu, "A decision making strategyfor generating unit tripping under emergency circumstances based ondeep reinforcement learning, " *Proceedings of the CSEE*, vol. 38, no. 1, pp. 109–119, Jan. 2018.

[104] Q. H. Huang, W. T. Hao, J. Tan, R. Fan, Z. Y. Huang, "Adaptive power system emergency control using deep reinforcement learning, " arXiv: 1903.03712v1 (2019).

[105] R. Rocchetta, L. Bellani, M. Compare, E. Zio, and E. Patelli, "A reinforcement learning framework for optimal operation and maintenance of power grids, " *Applied Energy*, vol. 241, pp. 291–301, May 2019.

[106] Q. L. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning, " arXiv: 1904.09374 (2019).

[107] R. S. Diao, Z. W. Wang, D. Shi, Q. Y. Chang, J. J. Duan, and X. H. Zhang, "Autonomous voltage control for grid operation using deep reinforcement learning, " arXiv: 1904.10597 (2019).

[108] Z. Ni and S. Paul, "A multistage game in smart grid security: a reinforcement learning solution, " *IEEE Transactionson Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2684–2695, Sep. (2019).

[109] S. Paul and Z. Ni, "A study of linear programming and reinforcement learning for one-shot game in smart grid security, " in *Proceedings of 2018 International Joint Conference on Neural Networks*, 2018.

[110] Y. Chen, S. W. Huang, F. Liu, Z. S. Wang, and X. W. Sun, "Evaluation of reinforcement learning-based false data injection attack to automatic voltage control, " *IEEE Transactionson Smart Grid*, vol. 10, no. 2, pp. 2158–2169, Mar. (2018).
[111] W. R. Liu, P. Zhuang, H. Liang, J. Peng, and Z. W. Huang, "Distributed economic dispatch in Microgrids based on cooperative reinforcement learning, " *IEEE Transactionson Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2192–2203, Jun. 2018.
[112] C. J. Han, B. Yang, T. Bao, T. Yu, and X. S. Zhang, "Bacteria foraging reinforcement learning for risk-based economic dispatch via knowledge transfer, " *Energies*, vol. 10, no. 5, pp. 638, May 2017.
[113] T. Hirata, D. B. Malla, K. Sakamoto, K. Yamaguchi, Y. Okada, and T. Sogabe, "Smart grid optimization by deep reinforcement learning over discrete and continuous action space, " *Bullet in of Networking, Computing, Systems, and Software*, vol. 8, no. 1, pp. 19–22, Jan. 2019.
[114] S. Munir, S. F. Abedin, N. H. Tran, and C. S. Hong, "When edge computing meets Microgrid: a deep reinforcement learning approach, " *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7360–7374, Oct. 2019.
[115] D. L. Wang, Q. Y. Sun, Y. Y. Li, and X. R. Liu, "Optimal energy routing design in energy internet with multiple energy routing centers using artificial neural network-based reinforcement learning method, " *Applied Sciences*, vol. 9, no. 3, pp. 520, Feb. 2019.

Zidong Zhang received his B.S. degree from School of Electronic and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China in 2017. He is currently pursuing the M. S degree in Electrical Engineering with Shanghai Jiao Tong University, Shanghai, China. His current research interests include deep learning, reinforcement learning and other machine learning algorithms with their applications in the power system.

Dongxia Zhang received her M.S. degree in Electrical Engineering from the Taiyuan University of Technology, Taiyuan, Shanxi, China, in 1992 and her Ph. D. degree in Electrical Engineering from Tsinghua University, Beijing, China, in 1999. From 1992 to 1995, she was a Lecturer with Taiyuan University of Technology. Since 1999, she has been working at China Electric Power Research Institute. She is the co-author of four books, and more than 40 articles. Her research interests include power system analysis and planning, big data and AI applications in power systems. She is an Associate Editor of Proceedings of the CSEE.

Robert C. Qiu (S'93–M'96–SM'01–F'14) received the Ph.D. degree in Electrical Engineering from New York University (former Polytechnic University, Brooklyn, NY). He is currently a Professorin the Department of Electrical and Computer Engineering, Center for Manufacturing Research, Tennessee Technological University, Cookeville, Tennessee, where he started as an Associate Professor in 2003 before he became a Professor in 2008. He has also been with the Department of Electrical Engineering, Research Center for Big Data Engineeringand Technologies, Shanghai Jiao TongUniversity since 2015. He was named a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) in 2015 for his contributions to ultra-wideband wireless communications. His current interests are in wireless communication and networking, random matrix theory, artificial intelligence, and the smart grid technologies.