IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 14, 2021, accepted August 4, 2021, date of publication August 6, 2021, date of current version August 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3103100

Joint Optimization of Quota Policy Design and Electric Market Behavior Based on Renewable Portfolio Standard in China

DUNNAN LIU[1], WEIYE WANG[1], HUA LI[1], MENGSHU SHI[1], GANG CHEN[2], ZHONGHUA XIE[2], AND QIN YUE[2]

[1]School of Economics and Management, North China Electric Power University, Beijing 102206, China
[2]Jiangxi Power Exchange Center, Jiangxi 330002, China

Corresponding author: Weiye Wang (wangweiye012@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 72001078, and in part by the Power System State Key Laboratory under Grant SKLD20M12.

ABSTRACT Under the perspective of carbon neutrality, the green electricity absorption target constrained by the quota system policy plays a crucial role in reducing the carbon emission of the power industry. However, the current green certificate policy has not achieved good results. On the premise of reducing the additional market burden as much as possible, the policy parameters should take into account the influence of market behavior to formulate better policy parameters in line with China's carbon emission peak goal. This paper constructs a combined hierarchical reinforcement learning with off-policy correction and multi-agent deep deterministic policy gradient algorithm (HIRO-MADDPG). It realizes the benefit analysis of the existing policy parameters joint with the solution of the optimal policy parameters. The algorithm solves the problem that benefit analysis and parameter formulation cannot be jointly trained and improves the precision. The results indicate: 1) HIRO-MADDPG algorithm can reach the highest policy benefits on the premise of maintaining market fairness; 2) under the new optimal policy parameters, the income per kilowatt hour of thermal power generator(TPG) and renewable power generator(RPG) can be maintained at 10% under the condition of abolishing subsidies; 3) with the help of the new policy parameters, China's power sector will reach the peak of carbon emissions from coal-fired power plants in 2026 ahead of schedule, and reduce carbon emissions by a further 11% by 2030.

INDEX TERMS Quota system policy, hierarchical multi-agent reinforcement learning, tradable green certificate, carbon neutrality.

I. INTRODUCTION

Under the goal of carbon neutrality, China's economic transformation and structural adjustment have entered a critical period [1]. Promoting the diversification, cleanness and low-carbon energy supply, and the high efficiency, reduction and electrification of energy consumption [2] are important ways for the power industry to achieve the goal of "carbon peak" [3] and "carbon neutral" [4]. A consensus has been reached that wind and solar energy and other renewable energy sources should continue to develop on a large scale and with high quality [5]. Scientific top-level design of relevant systems has become a key scientific issue in promoting clean and low-carbon energy supply in China.

Quotation system is a mandatory regulation on the market share of renewable energy generation made by a country or region. The implementation of quota system needs the green certificate trading system to complement it. At present, China is implementing renewable portfolio standard(RPS) and tradeable green certificate(TGC) mechanisms, and the main operation process is shown in the Figure 1:

At present, relevant scholars have carried out positive research on the RPS system, which mainly includes three aspects. 1) the impact of the RPS system on a country's quota structure and social welfare level. G. Liu et al evaluated the conditional value-at-risk level of RPS system using different ratios of the environmental, social, and governance index [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Wang.

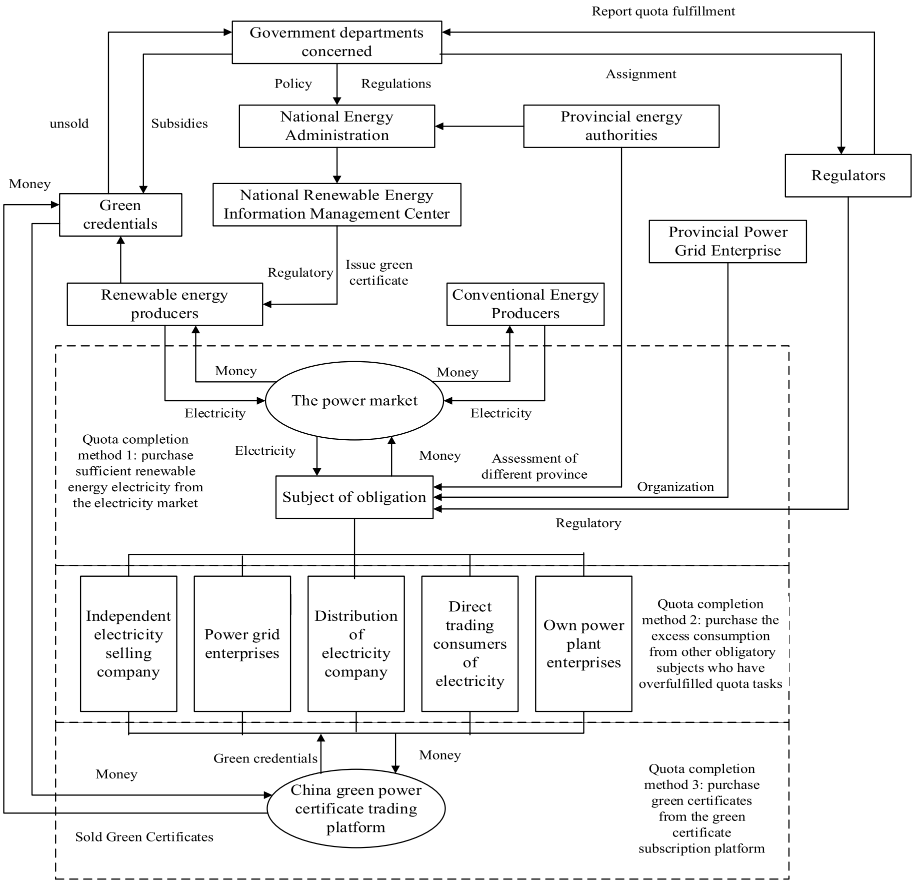
Report quota fulfillment
Government departments concerned
Assignment
Policy
Regulations
unsold
Subsidies
National Energy Administration
Provincial energy authorities
Regulators
Money
Green credentials
National Renewable Energy Information Management Center
Provincial Power Grid Enterprise
Regulatory
Issue green certificate
Renewable energy producers
Conventional Energy Producers
Money
Money
The power market
Electricity
Electricity
Quota completion method 1: purchase sufficient renewable energy electricity from the electricity market
Electricity
Money
Assessment of different province
Subject of obligation
Organization
Regulatory
Independent electricity selling company
Power grid enterprises
Distribution of electricity company
Direct trading consumers of electricity
Own power plant enterprises
Quota completion method 2: purchase the excess consumption from other obligatory subjects who have overfulfilled quota tasks
Green credentials
Money
Money
China green power certificate trading platform
Sold Green Certificates
Quota completion method 3: purchase green certificates from the green certificate subscription platform

FIGURE 1. RPS and TGC mechanisms in china.

Zhao et al constructed the social welfare function under RPS system and simulates the social welfare under the basis of the real quota situation of China [7]. Their findings suggest that Reasonable RPS quota structure can effectively reduce risks in the energy structure while improving social welfare.; 2) the influence of RPS system on the strategic behavior of power producers. W. Chen et al examined the impacts of cap-and-trade mechanisms on the decisions of a utility firm when it invests in renewable energy and has an existing conventional energy source [8]. X. Yu established system dynamic simulation model and scenario design method to improve sustainable development of Chinese power industry considering the integration impact of the green certificate market and the carbon emissions trading market [9]. X. Song introduced data envelopment analysis (DEA) model and the DEA-Malmquist index to measure the operational efficiency of TGC markets [10]. Their synthesis of research show that power plant business market behavior will produce tremendous changes under different policy parameters which could have significant impacts on the effect of RPS policy that cannot be ignored; 3) The influence of RPS system on electricity market cost. S. Shayegh analyzed the impact of market structure on RPS effectiveness by calculating the amounts of subsidies needed to achieve RPS mandates [11]. S. Shen deeply analyzed unit comprehensive cost of various energy under tradable green certificates market and sensitivity of key factors [12]. The results show that a rising of market costs is inevitable side effect of RPS policies. But the adverse consequences caused by market behavior can be greatly diminished by effective policy parameter setting.

To sum up, the importance and problem facing by RPS are well documented. In order to improve the consumption of green power under the premise of reducing the additional market burden as much as possible, it is necessary to comprehensively consider the different influence behaviors of the electricity market on different policy parameters to formulate the optimal RPS parameters. However, the existing researches mainly focus on the influence of established policy parameters on the comprehensive factors of all parties. Sensitivity analysis is often used for policy-making suggestions. The conclusions are too macroscopic and not detailed enough to fully guide the development of future policies. The results of policy benefit evaluation need to better serve to refine the formulation of policy parameter.

In order to solve the above problems, the algorithm should be able to achieve the joint optimization of policy making and market behavior. The algorithm can be divided into two parts: the first part is the reasonable response of market behavior under different policy parameters, the second part the iteration of policy parameters after the feedback of current market behavior.

For the first part, the subject can be simplified into the non- cooperative game model of green power manufacturer, thermal power manufacturer and power utility. The most important difficulty is how to get the game solution under different policy parameters which best accord with reality. The existing strategies for solving this problem can be divided into: 1) completely rational game [13], [14], that is, it is assumed that the players of the game participate in the decision that has the greatest benefit to their side from the beginning to the end, such as Cournot model [15], [16], Stark model [17], [18], etc.; 2) The bounded rational game [19], [20], that is, the player of the game needs to continuously learn the strategy rather than realize the optimal choice at one time. Evolutionary game theory (EGT) [21], [22] and neural network simulation [23], [24] are widely used classical algorithms.

The completely rational game method can accurately capture the theoretical optimal economic solution of each game subject, but the result is too ideal. And it faces the conflict between process rationality and result rationality [25]. The strategy learning process from the game subject to the stable state under bounded rationality is more consistent with the reality. However, the existing bounded rational games have great constraints on the action space and cannot get the high-precision game solution, which will affect the iteration of policy parameters.

For the second part, to determine the optimal policy parameters of quota system, the feedback of policy benefits should be iterated continuously. In the relative stable environment which can be mathematically expressed, the traditional optimization problem can be optimized by many different ways to get the optimal solution.

M. K. AlAshery et al established stochastic models to ensure that programming for the risk of the selected objective function distribution does not exceed a certain limit [26]. Bilevel optimization method is usually a good choice when faced with a comprehensive solution of multi-level or multi-party interests [27], [28]. Traditional back propagation particle swarm optimization (BPPSO) and reinforcement learning algorithm with action-reward incentive method [29] are also widely used optimization method to realize the process. However, there are few researches in this field at home and abroad, which are mainly faced with two difficulties: computational complexity and feedback accuracy. Since there is some uncertainty in the game result under bounded rationality, more than 10,000 times of repeated iteration are often needed in the optimization process to find the law. However, the neural network itself needs a long fitting time, and the computation complexity is too high in the case of two-layer model nesting, which is not feasible in practice. Although the evolutionary game algorithm converges faster, it can only solve the problem of the group's preference for the discrete decision but cannot realize the precise continuous decision. The policy parameters fed back by the algorithm have defects in the efficiency precision of the optimization, which constraints the upper limit of the optimization of the policy parameters.

In order to solve the above problems, HIRO-MADDPG algorithm based on hierarchical multi-agent reinforcement learning method is proposed in this paper. The MADDPG algorithm is composed of neural networks representing different market players, and continuous behavior decisions can be realized through the output layer of the neural network. The updating of neural network parameters approximates the bounded rational process of the game and ensures the game process is close to the reality. The improved hierarchical reinforcement learning method uses MADDPG to replace the ordinary reinforcement learning network as the lower layer, accepts the policy parameters output by TD3 reinforcement learning algorithm from the upper layer, and carries out the market game under the fixed policy parameters. The improved HIRO algorithm can update the upper and lower neural networks at the same time, so that the game process and policy parameters can be learned simultaneously. While the computational complexity of the algorithm is greatly reduced, the accuracy of the lower layer game results is improved, which helps to find out the policy parameter solutions that are most in line with the policy objectives of quota system.

II. TRADING MECHANISM MODEL

A. THE EVOLUTION MECHANISM OF QUOTA SYSTEM AND MARKET AGENT STRATEGY

There are four participants in this study: the government, RPG, TPG, and electricity power utilities [30]. The framework of the interaction is shown in Figure 2: the government establishes scientific and reasonable system parameters and rule constraints, which mainly include: quota target, transaction cost, unit penalty for failing to fulfill quota target [31]–[33], etc. In addition, the government carries out dynamic supervision on the electricity and energy market and the green certificate market and adjusts the parameters and rules of the system through the feedback of market information and the implementation of quota targets.

As shown in Figure 2. RPG, TPG and electric power utility decide their participation strategies in the electric energy market and green certificate market under the rules and parameters established by the government. Electricity utility, who are obligated to bear the quota, must consume a minimum proportion of electricity generated from renewable sources. According to the historical transaction prices of renewable energy, conventional energy and green certificates, power utilities dynamically decide the ratio of renewable energy to conventional energy, green certificate and penalty, so as to achieve the goal of minimizing the electricity cost and quota completion cost.

RPG determines the prices of electricity it can sell in the electricity market, which in turn determines the upper limit of its supply of green certificates in the green certificate market.

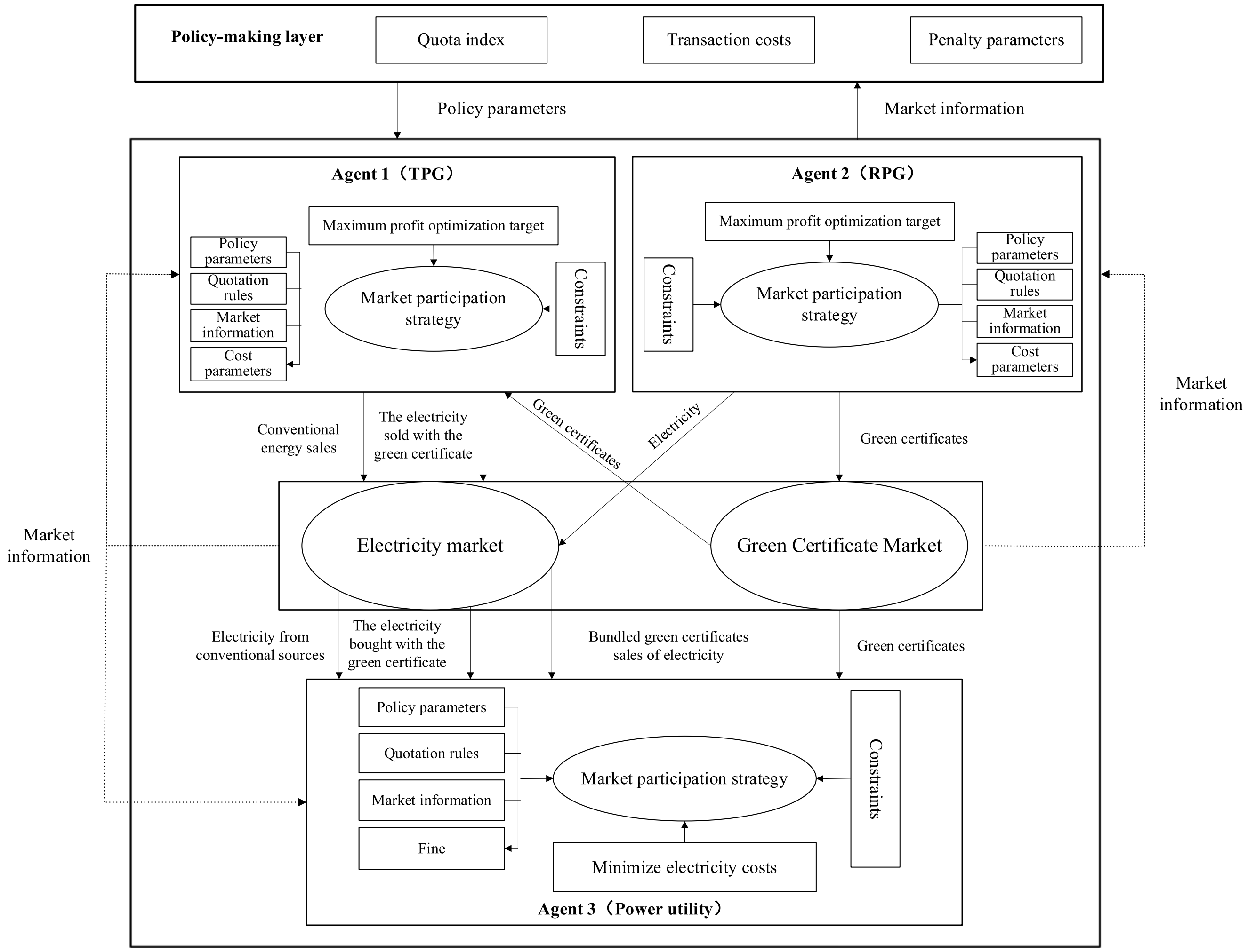

FIGURE 2. Operation principle of market participants.

RPG decides the supply of renewable energy electricity and green certificates in the market, and then influences the price of green electricity and green certificates, so as to maximize the income from electricity sales and green certificates.

TPG decides the supply of the conventional energy electricity and the price of it and accepts the green certificate price information of the market. When the market price of green certificates is low and the price of renewable energy electricity is high, TPG can purchase certain green certificates to endow the electricity with green properties and sell them bundled to power utility.

B. ASSUMPTIONS

Hypothesis 1: The market participants are all bounded rational. Due to the limited cognitive ability of the market subject, the incomplete information of the market transaction and the large amount of uncertainty in the transaction process, the market subject cannot be completely rational.

Hypothesis 2: When both parties choose to trade green certificates, both parties will share the transaction cost equally, and the cost allocation coefficient is 50%. To simplify the operation, the transaction cost of the total TGC is set as a single transaction volume multiplied by the unit transaction cost.

Hypothesis 3: The electricity transaction in this paper is mainly carried out through bilateral negotiation, and the user has full autonomy to choose the electricity of different nature. The electricity energy market and the green certificate market will organize a monthly transaction, and the validity of the green certificate is one year.

Hypothesis 4: there is no difference in power generation quality between different types of power generators, and they are connected to the grid with their respective electricity prices.

C. TWO-LAYER MODEL

a: UPPER-LAYER TARGETS

(1) In order to realize the competition between RPG and TPG and promote the healthy operation of the market, one of the objective functions in this paper is to minimize the difference

between RPG and TPG's earnings per kWh.

$$min\frac{\left|\pi^{re}/\sum_{i=1}^{12}Q_i^r-\pi^f/\sum_{i=1}^{12}Q_i^f\right|}{\pi^{re}/\sum_{i=1}^{12}Q_i^r+\pi^f/\sum_{i=1}^{12}Q_i^f} \tag{1}$$

Q_i^r is the electricity sold by RPG in i_{th} month. Q_i^f is the electricity sold by TPG in i_{th} month.

(2) In order to ensure that TPG does not lose too much due to the implementation of quota system, the objective function is to minimize the change of TPG's income per kilowatt hour before and after the implementation of quota system.

$$min\left|\left(\pi^f/\sum_{i=1}^{12}Q_i^f-r_{kwh}\right)/r_{kwh}\right| \tag{2}$$

r_{kwh} is the income per kWh of TPG before the implementation of the quota system

b: LOWER-LAYER TARGETS

i) REVENUE OF RPG

In the i_{th} step, the RPG maximizes its benefits by deciding the prices of electricity and green certificates.

$$max\pi^{re}=\sum_{i=1}^{12}\left(Q_i^r\left(p_i^r-c_r\right)+q_i^rp_i^g-\partial\lambda q_i^rp_i^g\right) \tag{3}$$

p_i^r is the transaction price of renewable energy in the i_{th} month; p_i^g is the sale price of green certificate in the i_{th} month; λ is the transaction cost rate; c_r is the renewable energy generation cost; ∂ is the transaction cost allocation coefficient.

ii) REVENUE OF TPG

In the i_{th} round of the game, TPG maximizes its benefits π^f by making decisions on the price of electricity to be sold p_i^f and the amount of green certificate to be purchased q_i^f.

$$max\pi^f=\sum_{i=1}^{12}\left(Q_i^f\left(p_i^f-c^f\right)-q_i^fp_i^g-\partial\lambda q_i^fp_i^g\right) \tag{4}$$

p_i^f is the transaction price of TPG in the i_{th} month; p_i^g is the purchase price of green certificate in the i_{th} month; c^f is the power generation cost of TPG.

iii) THE COST OF POWER UTILITIES

In the i_{th} round of the game, power utilities minimize their electricity cost C^u by making decisions on the amount of electricity purchased from renewable energy Q_i^r, conventional energy Q_i^f and green certificates q_i^u.

$$minC^u=C_B+C_P \tag{5}$$

$$C_B=\sum_{i=1}^{12}\left(Q_i^rp_i^r+Q_i^fp_i^f+p_i^uq_i^u+\partial\lambda p_i^uq_i^u\right) \tag{6}$$

$$C_P=\chi\left(\gamma\left(\sum_{i=1}^{12}\left(Q_i^r+Q_i^f\right)\right)-\sum_{i=1}^{12}\left(Q_i^r+q_i^u\right)\right)p_p \tag{7}$$

γ is the target of the quota, p_p is a unit penalty, χ is a 0-1 decision variable, when $\gamma\left(\sum_{i=1}^{12}\left(Q_i^r+Q_i^f\right)\right)-\sum_{i=1}^{12}\left(Q_i^r+q_i^u\right)\geq 0$, $\chi=1$; Otherwise $\chi=0$.

c: CONSTRAINTS

i) CONSTRAINTS OF THE NUMBER OF GREEN CERTIFICATES

$$\sum_{j=1}^{3}\sum_{i=1}^{12}\left(\tau\alpha_iQ_i^r+e_{i,j}^bG_i^b-e_{i,j}^sG_i^s\right)\geq\tau\sum_{i=1}^{12}\gamma Q_i^r \tag{8}$$

$$\sum_{j=1}^{3}\sum_{i=1}^{12}\left(e_{i,j}^b+e_{i,j}^s\right)\leq 1 \tag{9}$$

τ is the number of green certificates obtained per unit of green electricity produced, $\tau=1\text{piece}/\left(MWh\right)$. α_i is the proportion of renewable electricity consumption in the i_{th} month. G_i^b, G_i^s are respectively the number of green certificates purchased and sold in the i_{th} month. $e_{i,j}^b$, $e_{i,j}^s$ are the state variables of the purchase and sale of certificates by the market entity j in the i_{th} month. For market entity j, $e_{i,j}^b=0$, $e_{i,j}^s=1$ represents sell green certificates in the i_{th} month; $e_{i,j}^b=1$, $e_{i,j}^s=0$ represents purchase green certificates in the i_{th} month; $e_{i,j}^b=0$, $e_{i,j}^s=0$ indicates that there is no green certificate transaction in the i_{th} month.

ii) QUANTITY OF ELECTRICITY SOLD CONSTRAINT

The amount of electricity sold by RPG should be less than the maximum generating capacity of unit.

$$0\leq G_i^s\leq G_i^{\max} \tag{10}$$

$G_i^{\max}$ is the maximum number of green certificates held by power generation enterprises in the i_{th} month.

iii) TRADE BALANCE FOR GREEN CERTIFICATES

The total amount sold by each market member in the green certificate trading market shall be equal to the total amount purchased.

$$\sum_{j=1}^{3}\sum_{i=1}^{12}G_i^s=\sum_{j=1}^{3}\sum_{i=1}^{12}G_i^b \tag{11}$$

iv) PRICE CONSTRAINTS

Because of the randomness of machine learning algorithms, when the model is not adequately trained, the quotation of the agent is easy to appear extreme value. In order to ensure the rationality of the game and accelerate the convergence of the model, the price of TPG and RPG is constrained between 0.5 times and 1.5 times of the average price of a month in the market history.

The price constraint is released after 10,000 rounds of model training to more realistically simulate the real power market game.

$$0.5p_{i,avg}^{f} \leq p_{i}^{f} \leq 1.5p_{i,avg}^{f} \tag{12}$$

$$0.5p_{i,avg}^{r} \leq p_{i}^{r} \leq 1.5p_{i,avg}^{r} \tag{13}$$

p_i^f and p_i^r are the actual electricity prices of TPG and RPG in each month, $p_{i,avg}^f$ and $p_{i,avg}^r$ are the average prices of TPG and RPG in each month.

III. PROPOSED METHDOLOGY

A. BACKGROUND

In this study, reinforcement learning (RL) model was used as the benchmark algorithm of upper and lower layers, and the model training process of each benchmark algorithm was similar. In each time step t, agents composed of strategy network μ interact to generate actions $a_t \sim \mu(s_t), a_t \in \mathbb{R}^{d_a}$ to act on the environment by observing the environment $s_t \in \mathbb{R}^{d_s}$. Agent will harvest the reward value R_t according to its unknown reward function $R(s_t, a_t)$, and obtain the state s_{T+1} of the next environment or terminate at the current environment state s_T. The goal of agents is to maximize their discounted expected return $E_{s_0:T,a_0:T-1,R_0:T-1}\left[\sum_{i=0}^{T-1} \gamma^i R_i\right]$, where $0 \leq \gamma < 1$ is a user-specified discount factor [34].

B. FRAMEWORK OF IMPROVED HIERARCHICAL-MULTI-AGENT RL

Based on the hierarchical reinforcement learning algorithm structure of HIRO, this paper extends the general reinforcement learning algorithm into a two-layer structure including the lower strategy network μ^{lo} and the upper strategy network μ^{hi}. The upper-layer policy operates at a coarser layer. The lower-layer policy interacts directly with the environment. The upper-layer policy instructs the lower-layer policy via upper-layer actions, or goals, $g_t \in R^{d_s}$ which it samples anew every c steps.

The traditional HIRO algorithm uses parameterized reward functions to specify a limitless set of lower-layer policies, each of which is trained to match its observed states to a desired goal, and trains on the premise of the unity of the overall target direction of the upper and lower layers [35].

Different from the traditional HIRO algorithm, the upper layer of this paper uses the same continuous control RL algorithm as HIRO algorithm to form the strategy network μ^{hi} to output policy parameters, but the lower layer of the strategy network μ^{lo} is replaced with the MADDPG multi-agent network model representing the tripartite game between RPG, TPG and power utilities, and the unity of the interests of the upper and lower layers is cancelled. After the lower layer receives the policy target of the upper network output, the three parties aim to gain the most from their respective electricity market transactions. The lower-layer policy will store the experience $(s_t, g_t, a_t, r_t, s_{t+1})$ for off-policy training, using a fixed parameterized reward function r, with an intrinsic reward $r_t = r(s_t, g_t, a_t, s_{t+1})$ and a bounded rationality competitive game with a month as time dimension. The HIRO-MADDPG algorithm framework is shown in Figure 3:

After each round of game at the lower layer, the upper-layer policy independently calculates the policy benefits R_t, and stores the upper-layer transition $(s_{t:t+c-1}, g_{t:t+c-1}, a_{t:t+c-1}, R_{t:t+c-1}, s_{t+c})$ for off-policy training at every c time steps.

The improved HIRO-MADDPG algorithm transforms the leadership relationship between the upper and lower layers of the traditional HIRO-MADDPG algorithm into the seduction relationship. Instead of cooperating with the upper network to achieve the goals of the upper network, the lower network considers how to maximize its own interests under the goals set by the upper network. The transformation of this model is more in line with the relationship between the government and the market transaction subject in the current Chinese electricity market. The government needs to take full account of the profit-seeking of market members in order to work out the optimal policy parameters in line with the incentive compatibility principle.

C. MULTI-AGENT GAMING FOR LOWER-LAYER TRAINING

The multi-agent algorithm of the lower layer MADDPG adopts the framework of decentralized execution and centralized training to achieve their respective game objectives [36]. As shown in Figure 4, the algorithm allows the policy to use global information to simplify training, as long as this information is not used during testing.

After receiving the policy parameters of the upper network, we expect the lower algorithms to run competitive games under the following constraints:

(1) The learned strategies can only be executed using only local information (their own observations).

(2) It does not need to know the differentiable dynamic model of the environment.

(3) There is no communication method between agents (we assume there is no discernible communication channel).

Specifically, consider a game with three agents whose policy is parameterized by $\theta = \{\theta_1, \theta_2, \theta_3\}$, and the set of all agent strategies is $\pi = \{\pi_1, \pi_2, \pi_3\}$. The gradient of the expected payoff of agent i, $J(\theta_i) = E[R_i]$ is as follows:

$$\nabla_{\theta_j} J(\theta_i) = E_{s\sim p\mu, ai\sim\pi i}[\nabla\theta_i log\pi_i(a_i|o_i)Q_i^{\pi}(x, a_1, a_2, a_3)] \tag{14}$$

$Q_i^{\pi}(x, a_1, a_2, a_3)$ is a centralized action value function, which takes the action and state information x of all agents as input and outputs Q of the agent $i, x = (o_1, o_2, o_3)$ contains the observed values of all agents. Since each Q_i^{π} learns separately, the agent can have any kind of reward.

The above ideas can be extended to deterministic strategies. If N strategies $\mu_{\theta_i}^{lo}$ are considered and the parameter is θ_i (abbreviated as μ_i^{lo}), the gradient can be written as:

$$\nabla_{\theta i} J(\mu_i^{lo}) = E_{x,a\sim D}[\nabla_{\theta i}\mu_i^{lo}(a_i|o_i)\nabla_{ai}Q_i^{u^{lo}} \times (x, a_1, a_2, a_3)|a_i = \mu_{i(oi)}^{lo}] \tag{15}$$

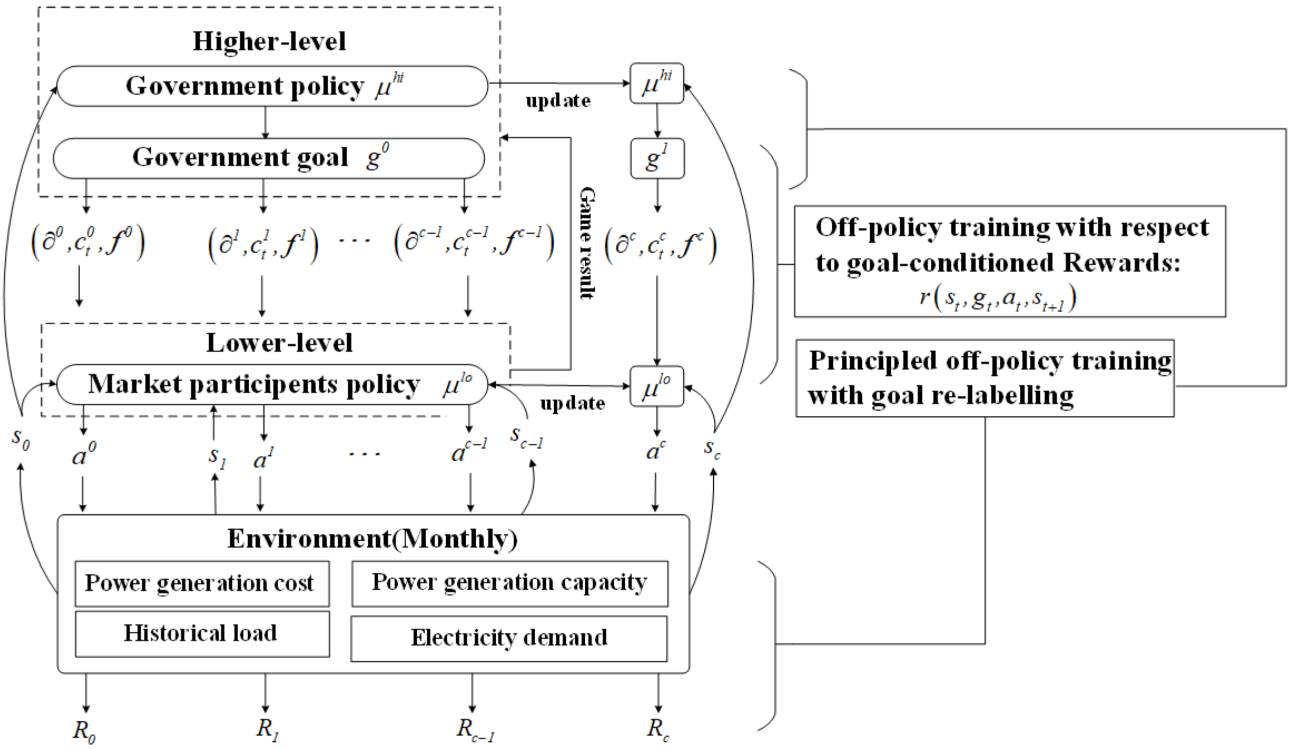
Higher-level
Government policy μ^{hi}
Government goal g^0
update
μ^{hi}
g^1
Game result
(∂^0, c_t^0, f^0)
(∂^1, c_t^1, f^1)
$\cdots$
$(\partial^{c-1}, c_t^{c-1}, f^{c-1})$
(∂^c, c_t^c, f^c)
Off-policy training with respect to goal-conditioned Rewards:
$r(s_t, g_t, a_t, s_{t+1})$
Principled off-policy training with goal re-labelling
Lower-level
Market participents policy μ^{lo}
update
μ^{lo}
s_0
a^0
s_1
a^1
$\cdots$
a^{c-1}
s_{c-1}
a^c
s_c
Environment(Monthly)
Power generation cost
Power generation capacity
Historical load
Electricity demand
R_0
R_1
R_{c-1}
R_c

FIGURE 3. The design and basic training of HIRO-MADDPG.

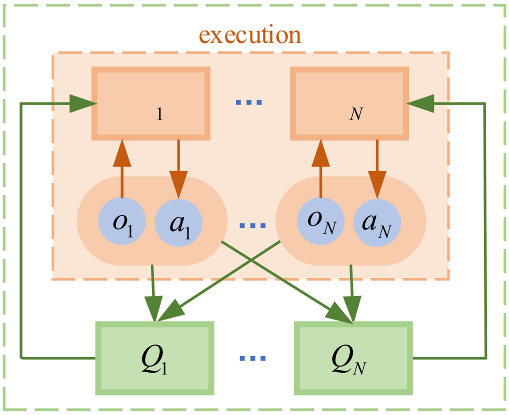

FIGURE 4. Overview of multi-agent gaming approach.

The experience replay buffer D contains tuple $(x, x', a_1, a_2, a_3, r_1, r_2, r_3)$, which records the experience of all agents. The action value function $Q_i^{u^{lo}}$ in the set is updated as follows:

$$L(\theta_i) = E_{x,a,r,x\prime}[(Q_i^{\mu^{lo}}(x, a_1, a_2, a_N) - y)^2] \tag{16}$$

$$y = r_i + \gamma Q_i^{\mu'^{lo}}(x\prime, a'_1, a'_2, a'_3)|a'_j = \mu_j^{lo\prime}(o_j) \tag{17}$$

$\mu^{lo}\prime = \{\mu^{lo}_{\theta_1\prime}, \mu^{lo}_{\theta_2\prime}, \mu^{lo}_{\theta_3\prime}\}$ is the target policy set with the delay parameter θ'_i.

In order to eliminate the assumptions of other agent strategies in(16), each agent can keep an additional approximation $\hat{\mu}^{lo}_{\phi^j}$ related to the real strategy μ_j^{lo} of agent j, ϕ is the parameter of the approximation, $\hat{\mu}^{lo}_{\phi^j}$ is abbreviated as $\hat{\mu}_j^{lo}$). This approximation strategy learns by maximizing the logarithmic probability of the actions of agent j and an entropy regularization term:

$$L(\phi_i^j) = -E_{oj,aj}[log\hat{\mu}_j^{lo}(a_j|o_j) + \lambda H(\hat{\mu}_j^{lo})] \tag{18}$$

H is the entropy of the strategy distribution. y can be replaced by the approximate value $\hat{y}$ calculated as follows:

$$\hat{y} = r_i + \gamma Q_i^{\mu\prime}(x\prime, \hat{\mu}_1^{lo\prime}(o_1), \hat{\mu}_2^{lo\prime}(o_2), \hat{\mu}_3^{lo\prime}(o_3)) \tag{19}$$

$\hat{\mu}^{lo\prime}$ is the target network of approximate strategy j. Before updating Q_i^u, the most recent sample of each agent j is fetched from the replay buffer and a gradient step is performed to update ϕ_i^j.

Relevant parameters of the model are shown in the Table 1:

D. OFF-POLICY CORRECTIONS FOR UPPER-LAYER TRAINING

Although a two-layer HRL architecture has been proposed before the HIRO algorithm, in previous work such a design usually requires on-policy training. This is because the changing behavior of the lower layer policy creates non-stationary problem policies for the upper layer policy, and the old off-policy experience may show different changes under the same goal conditions. However, for HRL methods to be applicable to real-world, they must be valid samples. Off-policy algorithms will usually show significantly better efficiency than on-policy actor-critic or policy gradient variants.

TABLE 1. Parameter list.

parameters	symbols	numerical
Upper-layer actor learning rate	α_a^{hi}	0.0001
Upper-layer critic learning rate	α_c^{hi}	0.0002
Lower-layer actor learning rate	α_a^{lo}	0.001
Lower-layer critic learning rate	α_c^{lo}	0.002
Reward discount	γ	0.99
Soft replacement	tau	0.01
Memory capacity	D	1000000
Update batch size	m	64
Max episodes	P	30000
Sub goal steps	g_t	12
Epsilon greedy value	ε	0.01

A policy may be learned efficiently from state-action-reward transition tuples (s_t, a_t, R_t, s_{t+1}) collected from interactions. TD3 learning algorithm [37] is utilized for upper-layer training, a variant of the popular DDPG algorithm for continuous control [38]. TD3 makes several modifications to DDPG's learning algorithm to yield a more robust and stable procedure. Its main modification is using an ensemble over Q-value models and adding noise to the policy when computing the target value.

In TD3, a deterministic neural network policy μ_ϕ is learned along with its corresponding state-action Q-function Q_θ by performing gradient updates on parameter sets ϕ and θ. The Q-function represents the future value of taking a specific action at starting from a state s_t. Accordingly, it is trained to minimize the average Bellman error over all sampled transitions, which is given by:

$$\varepsilon\left(s_t, a_t, s_{t+1}\right) = \left(Q_\theta\left(s_t, a_t\right) - R_t - \gamma Q_\theta\left(s_{t+1}, \mu_\phi\left(s_{t+1}\right)\right)\right)^2 \tag{20}$$

The policy is then trained to yield actions which maximize the Q-value at each state. That is, μ_ϕ is trained to maximize $Q_\theta\left(s_t, \mu_\phi\left(s_t\right)\right)$ over all s_t collected from interactions with the environment.

The upper layer transition tuples $(s_{t:t+c-1}, g_{t:t+c-1}, a_{t:t+c-1}, R_{t:t+c-1}, s_{t+c})$ is used, where $x_{t:t+c-1}$ denotes the sequence $x_t, \ldots, x_{t+c-1}$, which are collected by the upper-layer policy and convert them to state-action-reward transitions $\left(s_t, g_t, \sum R_{t:t+c-1}, s_{t+c}\right)$ that can be pushed into the replay buffer of any standard off-policy RL algorithm.

However, since transitions obtained from past lower-layer controllers do not accurately reflect the actions (and therefore resultant states $s_{t+1:t+c}$) that would occur if the same goal were used with the current lower-layer controller, a correction that translates old transitions into ones that agree with the current lower-layer controller has to be introduced.

The main observation is that the goal g_t of a past upper-layer transition $\left(s_t, g_t, \sum R_{t:t+c-1}, s_{t+c}\right)$ may be changed to make the actual observed action sequence more likely to have happened with respect to the current instantiation of μ^{lo}. The upper layer action g_t which in the past induced a lower-layer behavior $a_{t:t+c-1} \sim \mu^{lo}\left(s_{t:t+c-1}, g_{t:t+c-1}\right)$ may be re-labeled to a goal $\tilde{g}_t$ which is likely to induce the same lower-layer behavior with the current instantiation of the lower-layer policy. Thus, the off-policy issue by re-labeling the upper-layer transition $\left(s_t, g_t, \sum R_{t:t+c-1}, s_{t+c}\right)$ with a different upper-layer action $\tilde{g}_t$ chosen to maximize the probability $\mu^{lo}\left(a_{t:t+c-1} \left| s_{t:t+c-1}, \tilde{g}_{t:t+c-1}\right.\right)$ is proposed to remedy. In effect, when the lower-layer policy μ^{lo} is modified, the question should be answered: for which goals would this new controller have taken the better actions as the old one?

Most RL algorithms will use random action-space exploration to select actions, which means that the behavior policy is stochastic and the log probability $\log \mu^{lo}\left(a_{t:t+c-1} \left| s_{t:t+c-1}, \tilde{g}_{t:t+c-1}\right.\right)$ may be computed as:

$$\log \mu^{lo}\left(a_{t:t+c-1} \left| s_{t:t+c-1}, \tilde{g}_{t:t+c-1}\right.\right) \propto -\frac{1}{2} \sum_{i=t}^{t+c-1} \left\| a_i - \mu^{lo}\left(s_i, \tilde{g}_i\right) \right\|_2^2 \tag{21}$$

To approximately maximize this quantity in practice, we compute this log probability for a number of goals $\tilde{g}_t$, and choose the maximal goal to re-label the experience.

IV. CASE ANALYSIS

A. TEST SYSTEM AND IMPLEMENTATION

In this paper, the multi-agent game strategy of TPG and RPG is discussed without considering hydroelectric power, and the evolution of policy parameters is simulated. The composition of China's electricity in 2020 is shown in the Table 2:

This study is based on data from the National Energy Administration for 2020 and excludes energy types that are not affected by the weight of absorption liability.

Only coal power and grid-connected renewable power (solar, biomass, wind, etc.) are considered. According to the *Notice on Establishing and Improving the Guarantee Mechanism of Renewable Energy Consumption (2019)* issued by the National Energy Administration, the average quota of non-hydropower renewable energy consumption in China is 10.5%. In order to ensure that the penalty for failure to complete the target has a restraining effect, the indirect penalty is set at 1.5 times the current price of TGC in this study. The benchmark for transaction costs is set at about 10% of the TGC price. The above parameters are the baseline environment parameters of the game function of the lower layer of MADDPG algorithm.

According to the *State Grid Energy Research Report (2020)*, the long-term average cost of thermal power generators is 270 yuan /MWh, and the long-term average weighted cost of green power generators is 500 yuan /MWh. At the present stage, China adopts the policy of choosing a policy of alternative-green permits or subsidies. According to the data of *China Green Power Certificate Subscriptions Trading Platform*, it can be found that the transaction price of green certificate is generally not lower than the amount of the policy subsidies (the average cost difference between thermal power and green power). Based on the market supply and demand

TABLE 2. China's electricity composition in 2020 (TWh).

Type	January	February	March	April	May	June
Thermal power	391.7	388.9	389.4	397.9	413.5	432.3
Renewable energy	38.7	38.6	56.1	49.9	52.9	44.8
Market demand for electricity	333.7	359.2	394.2	403.4	382.2	472.3
Type	July	August	September	October	November	December
Thermal power	460.0	509.0	422.3	399.1	470.1	464.6
Renewable energy	40.5	40.7	38.8	46.7	47.8	51.4
Market demand for electricity	370.7	410.2	371.9	420.6	493.3	477.8

data, the three parties play the multi-agent game according to their cost parameters and public policy information, so as to achieve the goal of maximizing their respective benefits. The results of the game will be used as the reward of the upper HIRO algorithm for iterative training.

The reinforcement learning model in the experiment is implemented by Python TensorFlow 2.3. The constraints in the formula are calculated using the Python interface in Gurobi. The experimental computer was a quad-core 2.60-GHz Intel Core i7-6700HQ processor with 16GB of RAM. In order to accelerate the convergence of the model, the lower layer of MADDPG algorithm adopts China's current policy parameters to carry out 10,000 rounds of game pre-training.

B. PERFORMANCE COMPARISON OF POLICY DECISION ALGORITHM

In order to verify the effectiveness of the proposed HIRO-MADDPG algorithm, a total of 30,000 rounds of game simulation were carried out. The game cycle is 12 months in the simulation environment. The three parties of the game formulate price and electricity purchase strategies respectively and give feedback on the transaction situation and respective income results of each game cycle under different policy parameters. The results are shown in the Figure 5.

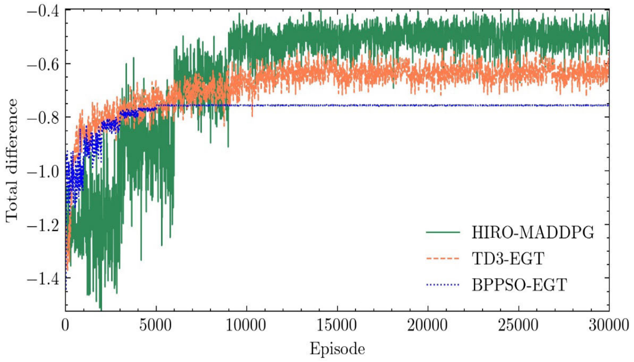

FIGURE 5. **Episodic average difference for the examined methods.**

This section is mainly used for algorithm comparison, so as to verify the effectiveness of hierarchical reinforcement learning HIRO-MADDPG algorithm proposed in this paper based on multi-agent game. This study mainly compares the policy effects of reinforcement learning TD3-EGT algorithm and traditional particle swarm neural network BPPSO-EGT algorithm based on bounded rational group evolutionary game. The EGT algorithm solves the strategy evolution trend by optimizing the annual comprehensive returns of different populations. The TD3/BPPSO algorithm at the upper layer receives the convergent strategy solutions at the lower layer to calculate the reward/loss value under the corresponding policy parameters and carries out 30,000 rounds of policy parameter iteration.

Figure 5 and Table 3 show the convergence trend of policy benefits under different algorithms and the mean value μ, variance of benefits σ and training duration $t(h)$ between different training rounds are presented. The policy benefit under different algorithms increases steadily with the training rounds, while the volatility decreases steadily. All the three algorithms can effectively improve the rationality of policy parameter formulation. However, the HIRO-MADDPG algorithm proposed in this paper obtained the highest policy benefits. The average policy benefit of HIRO-MADDPG algorithm is 38.7% higher than that of TD3-EGT algorithm, and 65.7% higher than that of traditional BPPSO-EGT algorithm. Although the volatility of the algorithm proposed in this paper is larger than that of the other two algorithms, the worst result of HIRO-MADDPG algorithm is 0.532 under the 95% confidence interval, which is still better than the optimal result under the 95% confidence interval of the other two algorithms. In order to calculate more advantageous policy parameters, HIRO-MADDPG algorithm takes about 3 days in total, 1.42 and 2.09 times longer than the other two algorithms. Since the policy parameters will not be modified within a short period of time once they are formulated, the radiation effects of the policy parameters on a long time scale need more precise calculation to determine the most appropriate policy parameters.

Compared with multi-agent reinforcement learning, EGT model, as an algorithm of the lower layer game, is superior to HIRO-MADDPG algorithm in terms of operation, convergence speed and volatility. However, due to the high speed and low volatility of the strategy convergence of the EGT algorithm, its convergence is only in the macro direction of the strategy, and it is unable to achieve the refinement and overall decision optimization of the subdivision environment in the unit of month, so the upper limit of the effect of the policy benefits it can achieve is relatively poor. TD3 reinforcement learning as a model of upper

TABLE 3. Mean (μ) and standard deviation (σ) of average difference and time spending for the examined methods.

Methods	parameters	Episode					
		5000	10000	15000	20000	25000	30000
HIRO-MADDPG	μ	-1.0799	-0.6897	-0.5132	-0.4592	-0.4566	-0.4562
	σ	0.1908	0.1534	0.0486	0.0413	0.04312	0.0389
	$t(h)$	12.3256	24.3578	35.9542	47.8652	60.1023	71.3654
TD3-EGT	μ	-0.8335	-0.7071	-0.6402	-0.6367	-0.6334	-0.6327
	σ	0.1281	0.0415	0.0341	0.0290	0.0297	0.0287
	$t(h)$	8.5321	16.8563	25.6987	34.1236	42.3654	50.1236
BPPSO-EGT	μ	-0.8639	-0.7560	-0.7560	-0.7560	-0.7560	-0.7560
	σ	0.1049	0.0016	0.0010	0.0010	0.0009	0.0009
	$t(h)$	5.7589	11.6985	17.1036	21.0365	28.1236	33.8521

layer policy parameter optimization is better than the traditional BPPSO neural network optimization algorithm. The TD3 reinforcement learning algorithm improves the exploration of the strategy space by exploring the combined optimization of the operation mechanism, multiple strategies and action networks, and under the disadvantage of the fluctuation within the acceptable range and the convergence speed. HIRO-MADDPG algorithm replaces the evolution of EGT population strategy trend with the result of multi-agent reinforcement learning game, which provides more accurate and refined reward for the upper reinforcement learning network, further improves the upper limit of algorithm solution, and obtains the optimal algorithm effect.

C. ANALYSIS OF REWARD COMPONENT FOR HIRO-MADDPG ALGORITHM

The perspective of quota policy is to enable RPG to compete fairly and healthfully with TPG in the electricity market after the policy subsidies are abolished. Therefore, the objective function of HIRO-MADDPG is to reduce the loss range of thermal power revenue due to the influence of policy while keeping the income ratio of TPG and RPG as close as possible under the policy parameters output by the algorithm.

Figure 6 shows the respective benefit ratios of the dual objectives in the algorithm training process and the variation of the income of different power producers. It can be found that in the first 5000 rounds of the game, the revenue per kWh of TPG greatly decreased and then recovered, and then stabilized at a decrease of about 35%. Meanwhile, the revenue per kWh between TPG and RPG decreased steadily in the first 10000 rounds, and the revenue per kWh fluctuated around 10% in the subsequent training, and the profitability of RPG increased with the help of the policy.

It can be seen from the changes in the profits of both parties that during the training process, TPG's early profit loss fluctuates greatly. The income of the lowest kilowatt-hours is reduced to 0, while RPG's early profit rises more, exceeding 0.2 yuan/kWh at the most, which is about three times the revenue per kWh before the thermal power quota system policy.

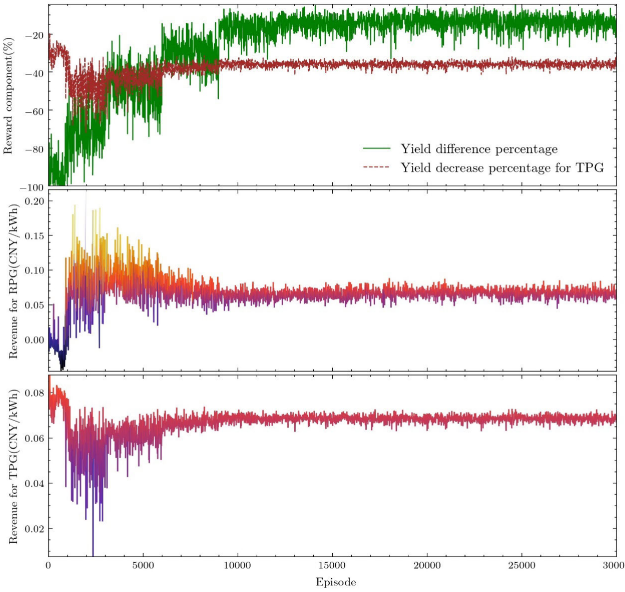

FIGURE 6. Reward component for HIRO-MADDPG algorithm and Revenue(CNY/kWh) for different generators.

Table 4 shows the changes of mean and variance of kWh earnings of TPG and RPG under different training rounds. It can be seen that the kWh earnings of TPG and RPG gradually converge to 0.069 yuan and 0.065 yuan. The average income of RPG is slightly lower than that of TPG, but the standard deviation of the income of RPG is 4.5 times that of TPG because of the income subsidy of green certificate policy. It can be seen that in the training process of the algorithm, the policy parameters avoid any party from gaining higher additional benefits or suffering losses due to excessive policy influence, thus maintaining the stable operation of the market.

D. ANALYSIS OF POLICY DECISION RESULT FOR UPPER LAYER ALGORITHM

In order to achieve the desired policy effect in the lower-layer multi-agent game, the upper-layer algorithm will

TABLE 4. Mean (μ) and standard deviation (σ) of revenue(CNY/kWh) for different generators.

participants	parameters	Episode					
		5000	10000	15000	20000	25000	30000
RPG	μ	0.06604	0.07117	0.06732	0.06663	0.06658	0.06448
	σ	0.04898	0.01631	0.00704	0.00685	0.00656	0.00638
TPG	μ	0.06131	0.06626	0.06869	0.06856	0.06853	0.06858
	σ	0.01142	0.00371	0.00148	0.00146	0.0015	0.00142

dynamically adjust the formulation of policy parameters according to the reward returned by the lower-layer multi-agent game in the current round. The result of the decision is to float up or down a certain proportion coefficient according to the benchmark policy parameters. Figure 7 shows the changes in the output results of the three policy parameters of quota, transaction cost and penalty during the training process.

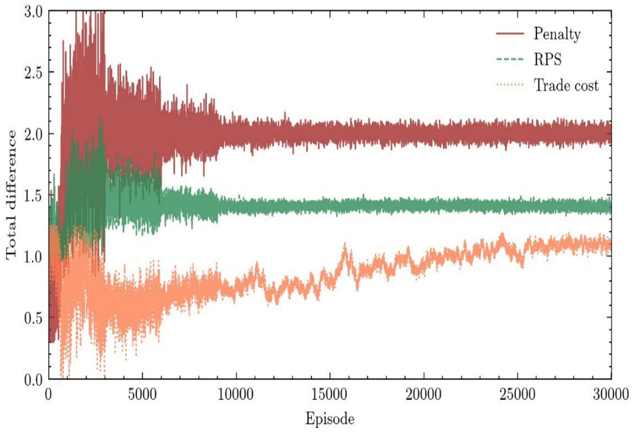

FIGURE 7. Episodic average difference for the policy decision output.

Due to the low proportion of the initial quota, there is a big gap between the income of per kWh of RPG and TPG. In order to improve the competitiveness of RPG, in the first 5000 rounds of algorithm training, the average policy parameter value is 1.6 times of the current quota and 0.7 times of the current transaction cost. Fines fluctuate wildly between about 0.5 times and 2 times. It can be seen from Figure 7 that the revenue per kWh of RPG increases rapidly under the high policy benefits, and they get much higher revenue per kWh than TPG.

The algorithm gradually corrected the excessive influence, and the policy parameters converged to 1.4 times of the current quota ratio after 20,000 rounds of iteration, so that the excess income obtained by RPG was reduced to a reasonable layer again. In order to ensure the stability of the profits of RPG, the penalty rapidly converges to twice the current penalty ratio after the high fluctuation in the early stage. The transaction cost plays a role of slightly adjusting the profit difference between TPG and g RPG per kilowatt hour. During the whole training period, the convergence rate is slow, and the transaction cost gradually increases from the low transaction cost to 1.08 times of the transaction cost.

E. ANALYSIS OF MULTI-AGENT BIDDING RESULT FOR LOWER LAYER ALGORITHM

Under the influence of different policy parameters in each round, the results are shown in Figure 8.

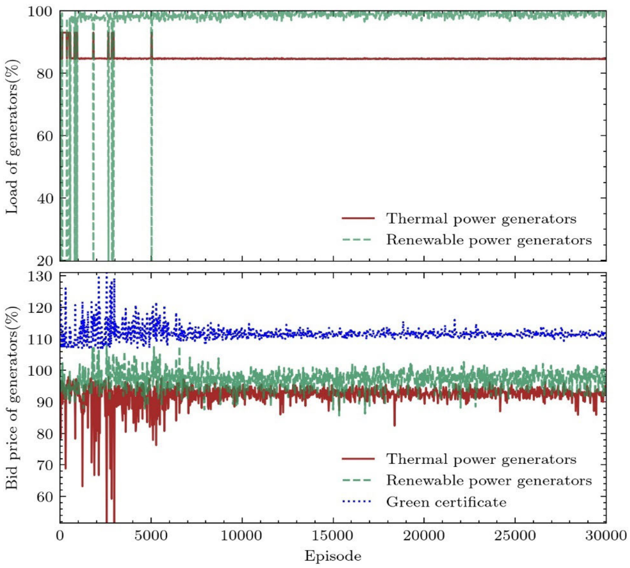

FIGURE 8. Episodic average load and price for the different generators after policy change.

Due to the small volume of RPG, when the policy parameters fluctuate greatly in the early stage, the load of RPG fluctuates sharply between 20%-100%, which is 9.4 times that of TPG. Due to the large size of TPG, the drastic fluctuation of quota keeps the fluctuation range of TPG within 10% and rapidly converges. Finally, under a reasonable quota, RPG can absorb nearly 100% of the power generation, and the load of TPG has steadily decreased by 5% from about 90% at the beginning.

In terms of pricing, TPG, in order to actively respond to the policy, continuously tested and offered different layers of low prices in the first 5000 front round game, with the average price reduced by 10%. With the stabilization of policy parameters, the price has picked up to a certain extent and finally stabilized at 92.6% of the benchmark price.

TABLE 5. Penalty condition and mean contribution for GC between TPG and power utility.

	Episode					
	5000	10000	15000	20000	25000	30000
RPS completion rates (%)	89.1	94.52	96.347	98.26	99.808	100
Max penalty(Million CNY)	265.822	98.632	25.635	10.152	1.652	0
Power utility contribution(%)	81.635	87.702	91.609	93.197	94.08	94.666
TPG contribution(%)	18.365	12.298	8.391	6.803	5.92	5.334

TABLE 6. Mean (μ) and standard deviation (σ) of load and price for the different generators after policy change.

	Month					
	1	2	3	4	5	6
RPS completion rates (%)	10.067	18.111	31.88	43.073	56.843	61.049
Average need for RPS	8.333	16.666	25	33.333	41.666	50
	7	8	9	10	11	12
RPS completion rates (%)	68.353	74.118	80.588	88.629	94.011	100.95
Average need for RPS	58.333	66.666	75	83.333	91.666	100

Although the fluctuation of electricity quantity and green license quotation strategy of green electricity generators is slightly larger than that of TPG, the average price of the quotation is relatively stable. The price of green electricity was slightly reduced by 3% in response to the reduction of thermal power, while the price of green certificates was increased by 11%, which did not affect the consumption of green certificates. Under the latest policy parameters, the quota has reached 100 percent completion rate, and the maximum penalty per round has been reduced to 0 yuan from 265 million yuan.

As shown in Figure 9, with the increase in the proportion of fines and the decrease in the price of green certificates during training, users are more inclined to purchase green certificates by themselves rather than by TPG to avoid high fines. Table 5 shows that the contribution ratio of users to the green certificate market increased from the initial 81.6% to 94.7%, an increase of 16.1%, while the contribution ratio of TPG decreased from 18.4% to 5.3%, a decrease of 71.2%.

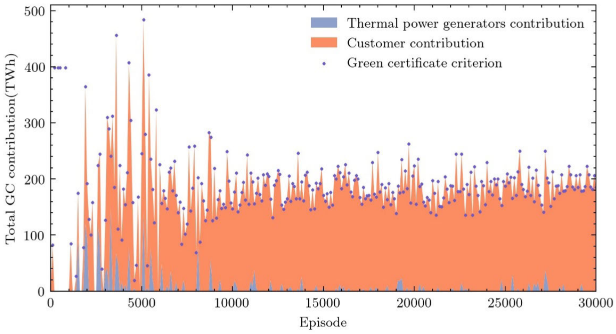

FIGURE 9. Episodic average GC criterion and contribution between TPG and power utility.

F. ANALYSIS OF MONTHLY BIDDING RESULT UNDER BEST POLICY PARAMETER

The average value of the parameter solutions in the last 5000 rounds of training was taken as the policy input of the lower layer game. Among them, the quota proportion is 14.8%, the transaction cost is 10.8%, and the fine is 690 yuan /MWh. Table 6 shows the cumulative completed RPS indicators and the average corresponding indicator demand in different months. Figure 10 shows the monthly quotation decisions and transaction results of different power producers.

As shown in Table 6, power utilities tend to guarantee a higher quota completion rate over the course of a year in order to avoid penalties and improve fault tolerance. RPS completion rates consistently exceed average monthly quota proportional allocations and exceed a maximum of 15% of linear monthly proportional allocations.

Among them at the beginning of a year and the end of the green electricity trade is the most active.

As shown in Figure 10, the highest amount of RPG winning the bid in the first 5 months of a year can reach 14% of that of TPG, and the average amount of winning the bid is 1.07 times of that in the next 7 months. The total trading volume of the first 5 months of the green certificates exceeded the sum of the next 7 months by 27%. The strategy adopted by RPG is to adopt lower prices of green electricity and green license in the first five months to guarantee the basic income, and increase the green electricity price by 2% on average and the green license price by 10% on average in the next seven months. TPG's trading strategy is the opposite, after the middle of the year to carry out appropriate price cuts. The volume of green electricity transaction and the proportion of thermal power gradually decreased after the middle of the year. At the end of the year when quota indicators need to be assessed, power

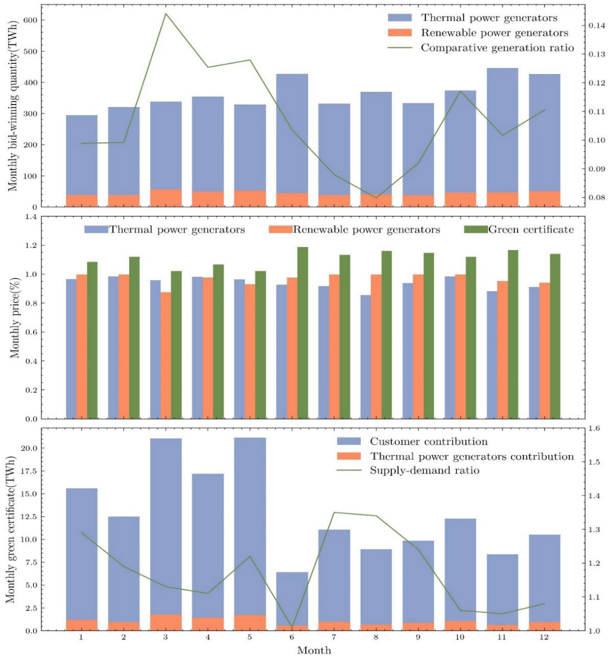

FIGURE 10. Monthly trading condition for different generators.

utilities make up for the lack of indicators, and the proportion of trading volume has increased to a certain extent.

Under the modified policy parameters, power utilities have a strong willingness to fulfill quota driven by quota index and fine. Power utilities conducted sufficient green electricity transactions at the beginning of the year and the end of the year, avoiding the low turnover in the green certificate market at the end of 2020, ensuring the consumption of green electricity and the reasonable per-kilowatt income of RPG, which can effectively promote the fair competitiveness of RPG in the electricity market.

G. ANALYSIS OF THE TREND OF CARBON EMISSION FROM THERMAL POWER UNDER OPTIMAL POLICY PARAMETERS DURING 2021-2030

The setting of RPS policy parameters will greatly affect the power generation behavior of power sector. The appropriate policy parameters, which correspond to China's future energy demand and power producer installation plans, can help China achieve the carbon peak goal in the electric power industry more efficiently and quickly.

According to *China's Energy Outlook 2030*, total energy demand will grow at an average annual rate of 2.4 percent to 5.3 billion tons of standard coal by 2030. Among them, the installed scale of non-water renewable energy will reach 1.24 billion kilowatts, accounting for 39% of the total installed capacity. The expected annual variation of thermal power and green power installed capacity is shown in Table 7:

TABLE 7. Installed capacity on the power side.

Installed capacity(Tw)	Year				
	2022	2024	2026	2028	2030
TPG	1.08	1.15	1.22	1.27	1.30
RPG	0.64	0.74	0.89	1.05	1.24

Assuming that the KWH cost of TPG and RPG remains unchanged, the installed capacity values of each year are taken as different environmental parameters for training, and the quota ratio and thermal power carbon emissions to 2030 are obtained, as shown in Figure 11.

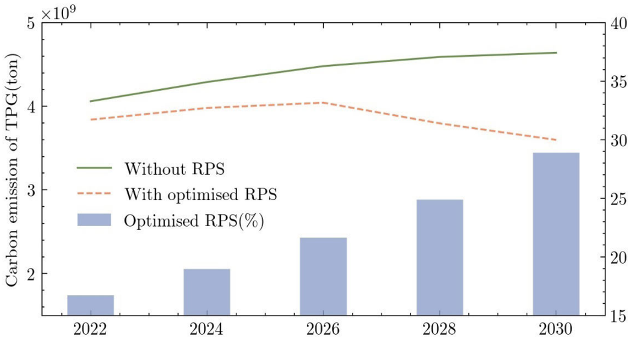

FIGURE 11. Optimal quota ratio and carbon emissions of TPG in the next 10 years.

From 2022 to 2030, the proportion of renewable energy quota will gradually increase, reaching a maximum of 28.76%. Regardless of the quota policy, carbon emissions from thermal power plants are expected to rise steadily, reaching 4.64 billion tons in 2030, in the face of continued growth in installed capacity. Under the optimal quota parameters calculated by HIRO-MADDPG algorithm, the carbon emission of thermal power units will reach the peak of 4.043 billion tons in 2026, which realizes the carbon peak of the power industry in advance. It was further reduced to 3.598 billion tons in 2030, 22.5% less than the no-quota policy, making a great contribution to achieving carbon neutrality.

V. CONCLUSION AND FUTURE IMPLICATIONS

Taking into full consideration the influence of the flexibility game results on the policy parameters under the tripartite bounded rationality of green power generators, thermal power generators and power utilities, this paper solves the problem of setting the optimal policy parameters to ensure the fair competitiveness of green power generators in the electricity market after the subsidy policy is cancelled. Different from the traditional rough policy sensitivity analysis, this paper adopts the combination of layered and multi-agent reinforcement learning, and through the synchronous update of upper and lower layer reinforcement learning network, it realizes the formulation of the optimal and refined policy combination. The results show that the benefit of policy scheme solved by HIRO-MADDPG algorithm in this paper is 38.7% and 65.7% higher than that of traditional EGT algorithm combined with TD3 single agent reinforcement learning and traditional BPPSO algorithm, respectively.

The algorithm deals with the dilemma of increasing the revenue of green electricity and reducing the new market cost in a more balanced way. In the case of reducing the net profit of thermal power by 35%, the competition result of green power from no market competitiveness to about 10% difference with the profit of thermal power per kilowatt-hour is realized, and the market competition ability of green power generators is guaranteed after the cancellation of policy subsidies.

The algorithm results show that under the policy combination of 1.4 times the current quota ratio, 1.08 times the current transaction cost and 2 times the current fine, the average and optimal policy benefits can be obtained. Under this policy, green electricity can guarantee no less than 98% of the consumption proportion and 100% of the quota completion rate. At the same time, power utilities will be more active to participate in the green certificate market driven by quotas and fines, which will avoid the phenomenon that the green certificate transaction volume is rare and the cost gap between generators is small under the current quota policy system in China in 2020.

The algorithm effectively and reasonably realizes the possibility that green power generators and thermal power generators can compete on the same stage, so that green power generators can still maintain development through reasonable profit before the technology is mature. This will further ensure that China can achieve a steady increase in the proportion of green power installed in the future, to ensure that the carbon peak target and carbon neutral vision can be successfully achieved.

In the future, we will continue to track the changes of the actual input data and make real-time adjustments to the input and training of the model, so as to ensure the true closeness between the model and the reality and provide real-time progressive reference for the actual decision-making of the government and the basis for the game of various market entities.

REFERENCES

[1] Z. Jiang, P. Lyu, L. Ye, and Y. W. Zhou, ''Green innovation transformation, economic sustainability and energy consumption during China's new normal stage,'' *J. Cleaner Prod.*, vol. 273, Nov. 2020, Art. no. 123044, doi: 10.1016/j.jclepro.2020.123044.

[2] M. T. Kohl, T. A. Messmer, B. A. Crabb, M. R. Guttery, D. K. Dahlgren, R. T. Larsen, S. N. Frey, S. Liguori, and R. J. Baxter, ''The effects of electric power lines on the breeding ecology of greater sage-grouse,'' *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0209968, doi: 10.1371/journal.pone.0209968.

[3] H. Zhang, Z. Hu, Z. Xu, and Y. Song, ''Evaluation of achievable vehicle-to-grid capacity using aggregate PEV model,'' *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 784–794, Jan. 2017, doi: 10.1109/TPWRS.2016.2561296.

[4] Y. Chi, Z. Liu, X. Wang, Y. Zhang, and F. Wei, ''Provincial CO_2 emission measurement and analysis of the construction industry under China's carbon neutrality target,'' *Sustainability*, vol. 13, no. 4, p. 1876, Feb. 2021, doi: 10.3390/su13041876.

[5] H. Can and Ö. Korkmaz, ''The relationship between renewable energy consumption and economic growth: The case of bulgaria,'' *Int. J. Energy Sector Manage.*, vol. 13, no. 3, pp. 573–589, Sep. 2019, doi: 10.1108/IJESM-11-2017-0005.

[6] G. Liu and S. Hamori, ''Can one reinforce investments in renewable energy stock indices with the ESG index?'' *Energies*, vol. 13, no. 5, p. 1179, Mar. 2020, doi: 10.3390/en13051179.

[7] Z. Ying, Z. Xin-gang, J. Xue-feng, and W. Zhen, ''Can the renewable portfolio standards improve social welfare in China's electricity market,'' *Energy Policy*, vol. 152, May 2021, Art. no. 112242, doi: 10.1016/j.enpol.2021.112242.

[8] W. Chen, J. Chen, and Y. Ma, ''Renewable energy investment and carbon emissions under cap-and-trade mechanisms,'' *J. Cleaner Prod.*, vol. 278, Jan. 2021, Art. no. 123341, doi: 10.1016/j.jclepro.2020.123341.

[9] X. Yu, Z. Dong, D. Zhou, X. Sang, C.-T. Chang, and X. Huang, ''Integration of tradable green certificates trading and carbon emissions trading: How will Chinese power industry do?'' *J. Cleaner Prod.*, vol. 279, Jan. 2021, Art. no. 123485, doi: 10.1016/j.jclepro.2020.123485.

[10] X. Song, J. Han, Y. Shan, C. Zhao, J. Liu, and Y. Kou, ''Efficiency of tradable green certificate markets in China,'' *J. Cleaner Prod.*, vol. 264, Aug. 2020, Art. no. 121518, doi: 10.1016/j.jclepro.2020.121518.

[11] S. Shayegh and D. L. Sanchez, ''Impact of market design on cost-effectiveness of renewable portfolio standards,'' *Renew. Sustain. Energy Rev.*, vol. 136, Feb. 2021, Art. no. 110397, doi: 10.1016/j.rser.2020.110397.

[12] S. Shen, D. He, Y. He, Y. Shen, F. Li, E. Xu, X. Wang, and K. Zhu, ''Research on evolution and development of power generation scale and cost under tradable green certificates market in China,'' in *Proc. IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Jun. 2020, pp. 686–690, doi: 10.1109/ITOEC49072.2020.9141868.

[13] R. J. Aumann and J. H. Dreze, ''Rational expectations in games,'' *Amer. Econ. Rev.*, vol. 98, no. 1, pp. 72–86, Feb. 2008, doi: 10.1257/aer.98.1.72.

[14] J. Gao, T. Wong, C. Wang, and J. Y. Yu, ''A price-based iterative double auction for charger sharing markets,'' *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 15, 2021, doi: 10.1109/TITS.2020.3047984.

[15] B. F. Hobbs and J. S. Pang, ''Nash-cournot equilibria in electric power markets with piecewise linear demand functions and joint constraints,'' *Oper. Res.*, vol. 55, no. 1, pp. 113–127, Feb. 2007, doi: 10.1287/opre.1060.0342.

[16] C. H. Tremblay and V. J. Tremblay, ''Oligopoly games and the cournot–bertrand model: A survey,'' *J. Econ. Surv.*, vol. 33, no. 5, pp. 1555–1577, Dec. 2019, doi: 10.1111/joes.12336.

[17] J. L. Barton and F. R. Brushett, ''A one-dimensional stack model for redox flow battery analysis and operation,'' *Batteries*, vol. 5, no. 1, p. 25, Feb. 2019, doi: 10.3390/batteries5010025.

[18] J. Kim, J. Lee, and J. K. Choi, ''Joint demand response and energy trading for electric vehicles in off-grid system,'' *IEEE Access*, vol. 8, pp. 130576–130587, 2020, doi: 10.1109/ACCESS.2020.3009739.

[19] W. Chen and L. Li, ''A new emission trading with bounded rational countries: A network game approach,'' *Social Netw. Appl. Sci.*, vol. 2, no. 3, pp. 1–12, Mar. 2020, doi: 10.1007/s42452-020-2264-8.

[20] H. Tu, J. Ma, and X. Li, ''Analysis of a dynamic triopoly game in the electricity market with heterogeneous players,'' *J. Comput. Inf. Syst.*, vol. 8, pp. 7345–7353, Sep. 2012.

[21] Z. Yi, Z. Xin-gang, Z. Yu-zhuo, and Z. Ying, ''From feed-in tariff to renewable portfolio standards: An evolutionary game theory perspective,'' *J. Cleaner Prod.*, vol. 213, pp. 1274–1289, Mar. 2019, doi: 10.1016/j.jclepro.2018.12.170.

[22] C. Lefeng, Y. Ru, L. Guiyun, W. Jianhui, C. Yang, W. Xiaogang, Z. Jie, and Y. Tao, ''Multi-population asymmetric evolutionary game dynamics and its applications in power demand-side response in smart grid,'' *Zhongguo Dianji Gongcheng Xuebao/Proc. Chin. Soc. Electr. Eng.*, vol. 40, no. 1, pp. 20–36, 2020, doi: 10.13334/j.0258-8013.pcsee.200930.

[23] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. Tb, A. Muldal, N. Heess, and T. Lillicrap, ''Distributed distributional deterministic policy gradients,'' presented at the 6th Int. Conf. Learn. Represent., (ICLR) Track, 2018.

[24] D. R. Ghica and K. Alyahya, ''Latent semantic analysis of game models using LSTM,'' *J. Log. Algebr. Methods Program.*, vol. 106, pp. 39–54, Aug. 2019, doi: 10.1016/j.jlamp.2019.04.003.

[25] L. Zarri, ''On social utility payoffs in games: A methodological comparison between behavioural and rational game theory,'' *Theory Decis.*, vol. 69, no. 4, pp. 587–598, Oct. 2010, doi: 10.1007/s11238-009-9146-2.

[26] M. K. Alashery, D. Xiao, and W. Qiao, ''Second-order stochastic dominance constraints for risk management of a wind power producer's optimal bidding strategy,'' *IEEE Trans. Sustain. Energy*, vol. 11, no. 3, pp. 1404–1413, Jul. 2020, doi: 10.1109/TSTE.2019.2927119.

[27] J. Campos Do Prado and W. Qiao, "A stochastic bilevel model for an electricity retailer in a liberalized distributed renewable energy market," *IEEE Trans. Sustain. Energy*, vol. 11, no. 4, pp. 2803–2812, Oct. 2020, doi: 10.1109/TSTE.2020.2976968.

[28] D. Xiao, M. K. AlAshery, and W. Qiao, "Optimal price-maker trading strategy of wind power producer using virtual bidding," *J. Modern Power Syst. Clean Energy*, early access, Jun. 11, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9453233, doi: 10.35833/MPCE.2020.000070.

[29] H.-N. Wang, N. Liu, Y.-Y. Zhang, D.-W. Feng, F. Huang, D.-S. Li, and Y.-M. Zhang, "Deep reinforcement learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 21, pp. 1726–1744, Oct. 2020.

[30] A. Ciarreta, M. P. Espinosa, and C. Pizarro-Irizar, "Optimal regulation of renewable energy: A comparison of feed-in tariffs and tradable green certificates in the Spanish electricity system," *Energy Econ.*, vol. 67, pp. 387–399, Sep. 2017, doi: 10.1016/j.eneco.2017.08.028.

[31] J.-L. Fan, J.-X. Wang, J.-W. Hu, Y. Yang, and Y. Wang, "Will China achieve its renewable portfolio standard targets? An analysis from the perspective of supply and demand," *Renew. Sustain. Energy Rev.*, vol. 138, Mar. 2021, Art. no. 110510, doi: 10.1016/j.rser.2020.110510.

[32] D. Young and J. Bistline, "The costs and value of renewable portfolio standards in meeting decarbonization goals," *Energy Econ.*, vol. 73, pp. 337–351, Jun. 2018, doi: 10.1016/j.eneco.2018.04.017.

[33] B. Wang, Y.-M. Wei, and X.-C. Yuan, "Possible design with equity and responsibility in China's renewable portfolio standards," *Appl. Energy*, vol. 232, pp. 685–694, Dec. 2018, doi: 10.1016/j.apenergy.2018.09.110.

[34] P. R. Montague, "Reinforcement learning: An introduction," in *Trends in Cognitive Sciences*, R. S. Sutton and A. G. Barto, Eds. London, U.K.: Bradford Books, 1999, doi: 10.1016/s1364-6613(99)01331-5.

[35] O. Nachum, H. Lee, S. Gu, and S. Levine, "Data-efficient hierarchical reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 3303–3313.

[36] M. He, B. Zhang, Q. Liu, X. L. Chen, and C. Yang, "Multi-agent deep deterministic policy gradient algorithm via prioritized experience selected method," *Kongzhi yu Juece/Control Decis.*, vol. 36, no. 1, pp. 68–74, 2021, doi: 10.13195/j.kzyjc.2019.0834.

[37] S. Fujimoto, H. Van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, International Machine Learning Society, vol. 4, 2018, pp. 2587–2601.

[38] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," presented at the 4th Int. Conf. Learn. Represent., (ICLR) Track, 2016.

DUNNAN LIU received the B.E. and Ph.D. degrees in electrical engineering from Tsinghua University, China. He is currently an Associate Professor with the School of Economics and Management, North China Electric Power University (NCEPU), China. His research interests include risk management and operation of power market.

WEIYE WANG received the bachelor's degree from the School of Economics and Management, North China Electric Power University (NCEPU), in 2019, where he is currently pursuing the master's degree. His main research interest includes electricity market.

HUA LI received the bachelor's degree from the School of Economics and Management, North China Electric Power University (NCEPU), in 2019, where he is currently pursuing the master's degree. His main research interest includes electricity market.

MENGSHU SHI received the master's degree from North China Electric Power University (NCEPU), in 2020, where she is currently pursuing the Ph.D. degree with the School of Economics and Management. Her main research interests include low-carbon and energy economy development and comprehensive energy systems.

GANG CHEN joined Jiangxi Power Exchange Center, in 2018. He was appointed as the Deputy General Manager of the Exchange Center. His main research interests include organization, operation, and management of electricity market transactions.

ZHONGHUA XIE joined Jiangxi Power Exchange Center, in 2018. He was appointed as the Deputy Director of the Exchange Center, Marketing Department. His main research interest includes power market trading.

QIN YUE joined Jiangxi Power Exchange Center, in 2018. She works with the Electrical Marketing Department. Her main research interest includes power market trading.