A Human-Machine Reinforcement Learning Method for Cooperative Energy Management

Article *in* IEEE Transactions on Industrial Informatics · August 2021

DOI: 10.1109/TII.2021.3105115

6 authors, including:

Yuechuan Tao
The University of Sydney
77 PUBLICATIONS **1,980** CITATIONS
SEE PROFILE

Jing Qiu
The University of Sydney
282 PUBLICATIONS **8,876** CITATIONS
SEE PROFILE

Xian Zhang
Harbin Institute of Technology
52 PUBLICATIONS **2,361** CITATIONS
SEE PROFILE

A Human-machine Reinforcement Learning Method for Cooperative Energy Management

Yuechuan Tao, Jing Qiu, *Member*, *IEEE*, Shuying Lai, Xian Zhang, *Member*, *IEEE*, Yunqi Wang, and Guibin Wang, *Member*, *IEEE*

***Abstract*— The increasing penetration of distributed energy resources and a large volume of unprecedented data from smart metering infrastructure can help consumers transit to an active role in the smart grid. In this paper, we propose a human-machine reinforcement learning (RL) framework in the smart grid context to formulate an energy management strategy for electric vehicles (EVs) and thermostatically controlled loads (TCLs) aggregators. The proposed model-free method accelerates the decision-making speed by substituting the conventional optimization process, and it is more capable of coping with the diverse system environment via online learning. The human intervention is coordinated with machine learning to: 1) prevent the huge loss during the learning process; 2) realize emergency control; 3) find preferable control policy. The performance of the proposed human-machine reinforcement learning framework is verified in case studies. It can be concluded that our proposed method performs better than the conventional deep Q-learning (DQN) and deep deterministic policy gradient (DDPG) in terms of convergence capability and preferable result exploration. Besides, the proposed method can better deal with emergent events, such as a sudden drop of PV output. Compared with the conventional model-based method, there are slight deviations between our method and the optimal solution, but the decision-making time is significantly reduced.**

***Index Terms*— Electric Vehicles, thermostatically controlled loads, reinforcement learning, human-machine, energy management.**

This work is supported by the ARC Research Hub Grant IH180100020, the ARC Training Centre IC200100023, the ARC linkage project LP200100056, and Sir William Tyree Foundation-Distributed Power Generation Research Fund. This work is also partially supported by the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS).

This work is jointly supported by the Foundations of Shenzhen Science and Technology Committee under Grant JCYJ20170817100412438 and Grant JCYJ20190808141019317.

This work is jointly supported by the National Natural Science Foundation of China (Grant No.72001058), General Program of Foundations of Shenzhen Science and Technology Committee (Grant No. GXWD20201230155427003-20200822103658001) and the Research Start-up Foundation for new teachers in Harbin Institute of Technology (Shenzhen)

Y. Tao, J. Qiu, S. Lai, and Y, Wang are with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia. (Email:qiujing0322@gmail.com)

G. Wang is with the College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518060, China

X. Zhang is with the School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen, 518055, China

I. NOMENCLATURE

A. Variables in the dispatch of EVs and TCLs

$E_{i,m,t}^{EV}$	Energy storage state of the EVs
I_{ij}	Branch current
$P_{i,m,t}^{EV,CHA/DIS}$	Charging and discharging power of the EVs
$P_{i,m,t}^{TCL}$	Power of the TCLs
$P_{i,m,t}^{TCL,C/D}$	Charging and discharging power of the VESS
$P_{i,t}^{DSO}$, $Q_{i,t}^{DSO}$	Active and reactive power from the sub-station
$P_{ij,t}^{B}$ $Q_{ij,t}^{B}$	Active and reactive power flow
$P_{i,t}^{DG}$	Output of distributed generators
T_t^{in}	Indoor temperature
$V_{i,t}$	Nodal voltage

B. Parameters in the dispatch of EVs and TCLs

$D_{i,m}$	Daily driving distance
$\underline{E_{i,m}^{EV}}$ $\overline{E_{i,m}^{EV}}$	Minimum and maximum energy storage state of EVs
$E_{i,m}^{EV,\text{expect}}$	Desired energy storage of EVs when departing
$\overline{P^{EV,CHA}}$ $\overline{P^{EV,DIS}}$	Maximum charging and discharging power of EVs
$P_{i,t}^{L}$ $Q_{i,t}^{L}$	Active and reactive load
$P_{i,t}^{DG}$	Output of distributed generators
R, C	Thermal resistance and heat ratio of air
r_{ij} , x_{ij}	Resistance and reactance of the branches
$\overline{S_{i,j}^{B}}$	Maximum apparent power capacity of the branches
T_t^{out}	Outdoor temperature
T^{set}	Set temperature
t_arrive/depart	Arrival and departure time of the EVs
$W_{i,m}$	Per unit energy consumption of EVs when driving
η^{cha} , η^{dis}	Charging and discharging efficiency of the EVs
δ	Acceptable temperature band
$\lambda_{i,t}$, $\overline{\lambda_t}$	Real-time and average electricity price
ε	Charge ratio of the battery degradation cost

φ^{BAT} Battery degradation cost

II. Introduction

THE conventional power systems are transforming to be more active and flexible but face new challenges such as distributed energy resources (DERs) dispatch, demand responses, and system operation strategies. On the demand side, electric vehicles (EVs) and thermostatically controlled loads (TCLs) are indispensable demand response resources that bring flexibility to the active distribution network. Traditionally, the dispatch of EVs and TCLs is solved through optimization. However, when the number of EVs and TCLs increases, the complexity of the optimization problem becomes the barrier to frequent demand responses. In recent years, the popularization of smart metering infrastructure and data-driven algorithms create new opportunities for energy management strategies.

Conventionally, the dispatch problems of EVs or TCLs are solved through optimizations that have been widely studied [1]. According to the findings in the previous literature, the uncoordinated dispatch of EVs brings extra energy losses and voltage deviations [2]. To enhance the power quality and reduce electricity bills, off-peak charging, valley charging, and other smart charging methods were considered [3]. Some references also have considered vehicle-to-grid (V2G) technologies so that the EVs owners become prosumers and can participate in the energy market actively [4]. Then, the aggregated EVs can be utilized in frequency and voltage regulation [5, 6]. With the penetration of DERs, a new energy management protocol emerges. In the work of Zhang *et al*. [7], an EV-to-EV (V2V) charging strategy was proposed to realize a cooperative charging strategy allowing communication between end-users. In conclusion, based on the current research, the energy management strategies for EVs are diverse. However, the control of EVs is usually limited by the uncertainties of the EVs' driving behaviors. Compared with EVs, TCLs are more controllable and can also realize load regulation [8, 9]. Thus, in this paper, we propose a coordinated dispatch strategy for EVs and TCLs. In the literature, Yao and Zhang focused on the coordination of heterogeneous TCLs to provide the ancillary service [10]. To protect the privacy of the end-users, distributed optimization algorithms were applied [11]. In the work of Wan *et al*. [12], the event-based control strategy was proposed, where the information exchange happened when the event-triggered condition was satisfied. However, these conventional methods are model-based and highly rely on the optimization model and the stochastic model of EVs. Hence, these conventional methods have limited capacity in rapidly adapting to time-vary changes in the distribution system.

In recent years, data-driven schemes have been identified as new opportunities and challenges to power systems. In the literature, the data-driven methods were utilized as auxiliary tools in most of the cases, such as electricity price [13] and load prediction [14], fault detection [15], load monitoring [16], etc. In 2015, the AlphaGo, developed by Google DeepMind, became the first computer program to defeat a professional human player of go game [17]. It proves that deep reinforcement learning (DRL) can solve a complex problem with massive state spaces. Motivated by this remarkable milestone, the researchers began to apply RL to the control problems in power systems [18]. The DRL can be categorized into two classes: *value-based algorithm* and *policy-based algorithm*. In a *value-based algorithm*, an action is selected based on the highest action value. A typical value-based algorithm is deep Q-learning (DQN). In the work of Qian *et al*. [19], EVs charging navigation was solved through DQN. Through DRL, the feature of the stochastic traffic condition can be extracted. According to their results, the DRL-based approach can approximate the result from optimization. In the work of Najafi *et al*. [20], a value-based algorithm was utilized to solve the charging of the EVs considering the bidding behaviors, and the proposed method can be applied to a multi-agent system. However, DQN is not good at solving continuous problems. In a *policy-based algorithm*, the probability of the action is outputted. Therefore, the strategy is learned rather than the action value. For example, Zhao *et al*. proposed a wind farm control strategy based on knowledge assist DRL by using a policy gradient algorithm [21], where the learning process is assisted by the experts' knowledge. Deep deterministic policy gradient (DDPG) is a combination of *value-based* and *policy-based algorithms* that have been utilized in the literature. In the work of Yan and Xu [22], the multi-agent DRL method was proposed for load frequency control based on DDPG. To make the algorithm easier to converge, the authors utilized the gradient of the value-action function to substitute the critic approximations. In the work of Mocanu *et al*. [23], DDPG was applied in the energy consumption schedule. Online learning and scheduling provide real-time feedback to customers so that customers are encouraged to use electricity efficiently. In the work of Ding *et al*. [24], DDPG-based DRL technology was utilized to analyze the impact of uncertainties on the charging strategy of EVs. Unlike the other machine learning algorithm, reinforcement learning can provide policies through learning, indicating that the decision-makers do not need to solve the optimization problem anymore. However, the decision made by the machine through learning is challenged because whether the DRL can converge to a satisfying result through learning cannot always be guaranteed. Besides, whether the machine can deal with an emergent event, such as a sudden drop of PV output, is questioned.

In this paper, a human-machine reinforcement learning framework based on DDPG is proposed to formulate an energy management strategy for EVs and TCLs aggregators. The main contributions can be listed as follows:

First, we propose a methodology to formulate the Markov decision process for aggregated EVs and TCLs. The EVs are molded as a battery energy storage system (BESS), and TCLs are molded as a virtual energy storage system (VESS). We have proposed a derivation process to derive the VESS model from the conventional thermal behavior model of TCLs. Hence, the BESS and VESS are jointly managed according to the Markov decision process (MDP) in the smart grid context.

Second, the DRL is the first time applied in the coordinated dispatch for aggregated EVs and TCLs. Compared with the

conventional model-based method, a decision can be made within a millisecond in a DRL-based method. The online-learning approach can be utilized to cope with the complex changes of the system states. The capabilities of different RL algorithms are compared in case studies.

Third, the human-machine is introduced to improve the traditional RL algorithm through emergency control, ramp-up reward, and human exploration. Through human intervention, the learning process is accelerated, the preferable solution is found, and the emergent situation can be better dealt with.

III. Coordinated Dispatch of EVs and TCLs

A. EVs Energy Storage Model

In this paper, the V2G technology is considered in the EVs dispatch. The concept of V2G allows EVs to provide power to help balance loads by "valley filling" (charging during the valley) and "peak shaving (discharging during the peak)". Besides, EVs can also provide frequency and voltage regulation service through V2G technology [5, 6]. When the EVs arrive home, there is no discharging resulted from driving anymore, and all the discharging behaviors occur between the EVs and the grids. Then, the EVs can be modeled as BESS as (1)-(3).

$$E_{i,m,t+1}^{EV} = E_{i,m,t}^{EV} + P_{i,m,t}^{EV,CHA}\eta^{cha}\Delta t - P_{i,m,t}^{EV,DIS}\Delta t / \eta^{dis} \tag{1}$$

$$0 \le P_{i,m,t}^{EV,CHA} \le \overline{P^{EV,CHA}}, 0 \le P_{i,m,t}^{EV,DIS} \le \overline{P^{EV,DIS}} \tag{2}$$

$$\underline{E_{i,m}^{EV}} \le E_{i,m,t}^{EV} \le \overline{E_{i,m}^{EV}} \tag{3}$$

where $E_{i,m,t}^{EV}$ is the energy storage state of m^{th} EV of i^{th} aggregator at time t; $P_{i,m,t}^{EV,CHA}$ and $P_{i,m,t}^{EV,DIS}$ are the charging and discharging power of EVs; η^{cha} and η^{dis} are the charging efficiency of the battery; $\overline{(\bullet)}$ and $\underline{(\bullet)}$ are the upward and downward limits.

However, the owners of the EVs have their driving behaviors. EVs can be dispatched only when the EVs are connected to the charging post. The constraints of the driving behaviors can be expressed as:

$$P_{i,m,t}^{EV,CHA} = 0, P_{i,m,t}^{EV,DIS} = 0 \quad \text{When } t < t\_arrive \tag{4}$$

$$E_{i,m,t\_arrive}^{EV} = \overline{E_{i,m}^{EV}} - D_{i,m} \times W_{i,m} \tag{5}$$

$$E_{i,m,t\_depart}^{EV} \ge E_{i,m}^{EV,\text{expect}} \tag{6}$$

where t_arrive and t_depart represent the arrival and departure time of each EV; $E_{i,m}^{EV,\text{expect}}$ is the expected energy storage state of EVs when departing; $D_{i,m}$ is the daily driving distance; $W_{i,m}$ is the electricity consumption per unit distance.

B. TCLs Virtual Energy Storage Model

As for the TCLs, the cooling mode and the heating mode can be chosen based on the users' needs. In this paper, the heating operation mode is taken as an example in the mathematical derivation. According to [25], conventional thermodynamic behavior for a heating system can be described as:

$$T_{t+1}^{in} = T_t^{in}e^{-1/RC} + \left(RP_t^{TCL}\Delta t + T_t^{out}\right)\left(1-e^{-1/RC}\right) \tag{7}$$

where T_t^{in} and T_t^{out} are the indoor and outdoor temperature; R refers to the thermal resistance; C refers to the heat ratio of the air; P_t^{TCL} is the working power of TCLs.

To ensure the comfort of the users, it is assumed that the acceptable temperature band is 2δ. So,

$$T^{max} = T^{set} + \delta, T^{min} = T^{set} - \delta \tag{8}$$

where T^{set} is the set temperature; T^{max} and T^{min} are the maximum and minimum indoor temperature.

Thus, for a VESS, when the indoor temperature reaches the largest value, the energy storage state for the VESS is the highest and vice versa. Then, we define the state of virtual energy storage σ_t as:

$$\sigma_t = \frac{T^{set} - T_t^{in}}{\delta}, -1 \le \sigma_t \le 1 \tag{9}$$

To model the TCLs as a VESS, the thermodynamic behavior model (7) can be rewritten as:

$$\frac{T^{set} - T_{t+1}^{in}}{\delta} = \frac{\left(T^{set} - T_t^{in}\right)e^{-1/RC}}{\delta} - \frac{\left(RP_t^{TCL}\Delta t + T_t^{out}\right)\left(1-e^{-1/RC}\right)}{\delta} + \frac{T^{set}\left(1-e^{-1/RC}\right)}{\delta} \tag{10}$$

Thus, combining equations (9) and (10), the energy storage balance of the VESS can be derived as:

$$\sigma_{t+1} = \kappa\sigma_t + \gamma P_t^{TCL,C/D}\Delta t \tag{11}$$

where

$$\kappa = e^{-1/RC}, \gamma = \frac{\left(1-e^{-1/RC}\right)}{\delta} \tag{12}$$

$$P_t^{TCL,C/D} = RP_t^{TCL} - \left(T_t^{out} + T^{set}\right) \tag{13}$$

where $P_t^{TCL,C/D}$ is the charging/discharging power of the VESS.

When the VESS is neither charging nor discharging, i.e. $P_t^{TCL,C/D} = 0$, the working power of TCL is $\left(T_t^{out} + T^{set}\right)/R$, which maintains the indoor temperature at the set value. When $P_t^{TCL,C/D}$ is negative, VESS will discharge, the indoor temperature will drop, and vice versa.

C. Formulation of EVs and TCLs Coordinated Energy Management

For the coordinated energy management problem, the objective function is to minimize the total cost of the system, including the energy purchase price from the sub-stations, system loss, and the battery degradation cost of EVs.

$$\min F = \sum_t \sum_{i\in\Omega^{ST}} \lambda_{i,t}P_{i,t}^{DSO} + \sum_t \sum_{ij\in\Omega} \overline{\lambda_t} I_{ij} r_{ij} + \sum_t \sum_{i\in\Omega^{EVA}} (1+\varepsilon)\varphi^{BAT} P_{i,t}^{EV,DIS} \tag{14}$$

where Ω, Ω^{ST} and Ω^{EVA} is the set of electricity buses, sub-stations, and EVs aggregators; $\lambda_{i,t}$ is the electricity price; $P_{i,t}^{DSO}$ is the energy purchased from sub-stations; $\overline{\lambda_t}$ is the average electricity price; I_{ij} is the line current; r_{ij} is the branch

resistance; φ^{BAT} is the battery degradation cost.

The constraints include Eqs.(1)-(13), the active power and reactive power balance, voltage, and power flow constraints.

$$P_{ij,t}^{B}-\sum_{k\in\Omega_j^{+}}P_{jk,t}^{B}-r_{ij}\frac{\left(P_{ij,t}^{B}\right)^2+\left(Q_{ij,t}^{B}\right)^2}{V_{i,t}^2}=P_{i,t}^{DG/DSO}-\sum_{m\in\Omega_i^{EV}}P_{i,m,t}^{EV,CHA}+\sum_{m\in\Omega_i^{EV}}P_{i,m,t}^{EV,DIS}-\sum_{m\in\Omega_i^{TCL}}P_{i,m,t}^{TCL}-P_{i,t}^{L},\ \forall i,j\in\Omega,\forall t \quad (15)$$

$$Q_{ij.t}^{B}-\sum_{k\in\Omega_j^{+}}Q_{jk.t}^{B}=Q_{i,t}^{DSO}-Q_{i,t}^{L}+x_{ij}\frac{\left(P_{jk,t}^{B}\right)^2+\left(Q_{jk,t}^{B}\right)^2}{V_{i,t}^2},\ \forall i,j\in\Omega,\forall t \quad (16)$$

$$V_{i,t}^2-V_{j,t}^2=2\left(r_{ij}P_{ij,t}^{B}+x_{ij}Q_{ij,t}^{B}\right)-\left(r_{ij}^2+x_{ij}^2\right)\frac{\left(P_{ij,t}^{B}\right)^2+\left(Q_{ij,t}^{B}\right)^2}{V_{i,t}^2},\ \forall i,j\in\Omega,\forall t \quad (17)$$

$$\underline{V_i}\le V_{i,t}\le\overline{V_i},\ \forall i\in\Omega,\forall t \quad (18)$$

$$\sqrt{\left(P_{ij,t}^{B}\right)^2+\left(Q_{ij,t}^{B}\right)^2}\le\overline{S_{ij}^{B}},\ \forall ij\in\Omega,\ \forall t \quad (19)$$

where $P_{ij,t}^{B}$ and $Q_{ij,t}^{B}$ are the active and reactive power flow; $Q_{i,t}^{DSO}$ is the reactive power from the sub-station. $P_{i,t}^{DG}$ is the active output of the distributed generators; $P_{i,t}^{L}$ and $Q_{i,t}^{L}$ are the active and reactive power load; r_{ij} and x_{ij} are the resistance and the reactance of the branch; $V_{i,t}$ is the nodal voltage; $\overline{(\bullet)}$ and $\underline{(\bullet)}$ are the upward and downward limits.

IV. Reinforcement Learning Based on Deep Deterministic Policy Gradient

A. Formulation of the Markov Decision Process (MDP)

When the number of EVs and TCLs increases, solving the optimization problem (14)-(19) is time-consuming and becomes a barrier to the frequent demand response. Hence, in this paper, we propose a DRL method to solve the energy management problem. In a reinforcement learning context, the agent, i.e., EVs and TCLs in this paper, will learn to act using the MDP. The MDP is defined by a tuple containing four elements: $\langle S,A,Pr(.,.),R(.,.)\rangle$.

- S is the state space, $\forall s\in S$
- A is the action space $\forall a\in A$
- $Pr(.,.)$ is the transition function defined as the probability that choosing action *a* under the state *s*, and will reach a new state *s'*, such that $Pr(s_{t+1}=s'|s_t=s,a_t=a)$.
- $R(.,.)$ is the reward function defined as the immediate reward received by the agent when the state is transit to *s'* from *s*.

The multi-agents, i.e., EVs and TCLs, aim to optimize a stochastic policy $\pi:S\times A\times R$ that can obtain the largest reward in the whole trajectory.

1) State

The state-space for EVs agents contains the local load, local voltage magnitude, the energy storage state of EVs, and the utilization (whether EVs are connected) of EVs and time, shown as:

$$s_{i,t}^{EV}=\left[P_{i,t}^{D},V_{i,t},\mathbf{E}_{i,t}^{EV},\boldsymbol{\mu}_{i,t}^{EV},t\right] \quad (20)$$

The state-space for TCLs agents contains the local load, local voltage magnitude, the energy storage state of VESS, and time, shown as:

$$s_{i,t}^{TCL}=\left[P_{i,t}^{D},V_{i,t},\boldsymbol{\sigma}_{i,t}^{TCL},t\right] \quad (21)$$

2) Action

Given the defined states, EVs take actions, including charging and discharging. The action set of EVs is compact and continuous, which can be shown as:

$$a_{i,t}^{EV}=\left[\mathbf{P}_{i,t}^{EV,CHA},\mathbf{P}_{i,t}^{EV,DIS}\right] \quad (22)$$

The action set of TCLs is the charging/discharging power of TCLs, which is compact and discrete or continuous.

$$a_{i,t}^{TCL}=\left[\mathbf{P}_{i,t}^{TCL,C/D}\right] \quad (23)$$

3) Transition Function

State transition from *t* to *t+1* is partly affected by the agents' actions $\mathbf{P}_{i,t}^{EV,CHA},\mathbf{P}_{i,t}^{EV,DIS},\mathbf{P}_{i,t}^{TCL,C/D}$ and random exogenous data Θ_t, such as local load. The transition function can be expressed as:

$$\begin{cases}[P_{i,t+1}^{D},V_{i,t+1},\boldsymbol{\mu}_{i,t+1}^{EV}]=f^{T}\left(P_{i,t}^{D},V_{i,t},\boldsymbol{\mu}_{i,t}^{EV},a_{i,t}^{EV},a_{i,t}^{TCL},\Theta_t\right)\\ \mathbf{E}_{i,t+1}^{EV}=\mathbf{E}_{i,t}^{EV}+\tilde{\mathbf{P}}_{i,t}^{EV,CHA}\eta^{cha}\Delta t-\tilde{\mathbf{P}}_{i,t}^{EV,DIS}\Delta t/\eta^{dis}\\ \boldsymbol{\sigma}_{i,t+1}^{TCL}=\kappa\boldsymbol{\sigma}_{i,t}^{TCL}+\gamma\mathbf{P}_{i,t}^{TCL,C/D}\\ t=t+1\end{cases} \quad (24)$$

4) Reward

The immediate reward of the agents is defined from four perspectives. First, when there is no local voltage violation, a positive reward ξ_2 will be given. Otherwise, a penalty (negative reward) will be given as (25).

$$r_1=\begin{cases}-\xi_1\left(V_{i,t}-\overline{V_i}\right) & ,if\quad V_{i,t}>\overline{V_i}\\ -\xi_1\left(\underline{V_i}-V_{i,t}\right) & ,if\quad V_{i,t}<\underline{V_i}\\ \xi_2 & ,else\end{cases} \quad (25)$$

From the monetary perspective, the agents will pay for the actual energy consumption $\tilde{P}$, and EVs will obtain profits from providing energy through V2G, shown as (26).

$$r_2=\begin{pmatrix}\mathbb{E}\left(\lambda_t^{S}\sum_{m\in\Omega_i^{EV}}\tilde{P}_{i,m,t}^{EV,DIS}\right)-\mathbb{E}\left(\lambda_t^{B}\sum_{m\in\Omega_i^{EV}}\tilde{P}_{i,m,t}^{EV,CHA}\right)\\ -\mathbb{E}\left(\lambda_t^{B}\sum_{m\in\Omega_i^{TCL}}\tilde{P}_{i,m,t}^{TCL}\right)\end{pmatrix}\Delta t \quad (26)$$

Furthermore, the constraints for EVs should be met. If the energy storage state exceeds the upper and lower limits, a large penalty ξ_3 will be given.

$$r_3 = \begin{cases} 0 & ,if \quad \mathbf{E}_{i,t}^{EV} \in \left[\mathbf{E}_{i}^{EV,\min}, \mathbf{E}_{i,t}^{EV,\max}\right] \\ -\xi_3 & ,else \end{cases} \tag{27}$$

If the energy storage state of EVs cannot support the next trip, a large penalty ξ_3 will be given.

$$r_3 = \begin{cases} 0 & ,if \quad E_{i,m,t\_depart}^{EV} \geq E_{i,m}^{EV,\text{expect}} \\ -\xi_3 & ,else \end{cases} \tag{28}$$

Finally, as for the dispatch of TCLs, to ensure the comfort of the customers, the indoor temperature should be maintained within a specific range. From the perspective of VESS, the constraints $\boldsymbol{\sigma}_{i,t}^{TCL} \in [-1,1]$ should be satisfied.

$$r_4 = \begin{cases} 0 & ,if \quad \boldsymbol{\sigma}_{i,t}^{TCL} \in [-1,1] \\ -\xi_3 & ,else \end{cases} \tag{29}$$

B. Deterministic Deep Policy Gradient

In this paper, we solve the RL problem based on the DDPG algorithm. In the policy gradient, we focus on the direct learning of the policy rather than the action value and the state value [26]. The expression $\pi(a \mid s, \vartheta)$ represents a policy π with the parameter ϑ indicating the possibility to choose action a based on state s. The action set and the state set of EVs and TCLs are defined in (20)-(23). The policy function π_ϑ is a mapping from the current state to the action. The mapping function is represented by a deep neural network in this paper, known as the actor network. Hence, the parameter ϑ to be optimized is the weights and bias in the convolutional neural network. Through updating ϑ, the largest reward of the policy π_ϑ is expected.

First, we consider the MDP of one step from the initial state. The initial state follows the distribution $s \sim d(s)$. Then the expected reward of one step can be expressed as:

$$J(\vartheta) = \mathbb{E}_{\pi\vartheta}\left[R_{s,a}\right] = \sum_{s\in S} d(s) \sum_{a\in A} \pi_\vartheta(s,a) R_{s,a} \tag{30}$$

The gradient can be derived according to:

$$\begin{aligned} \nabla_\vartheta J(\vartheta) &= \sum_{s\in S} d(s) \sum_{a\in A} \pi_\vartheta \nabla_\vartheta \log \pi_\vartheta(s,a) R_{s,a} \\ &= \mathbb{E}_{\pi\vartheta}\left[\sum_{a\in A} \pi_\vartheta \nabla_\vartheta \log \pi_\vartheta(s,a) R_{s,a}\right] \end{aligned} \tag{31}$$

However, to reach a globally optimal solution, the total reward of a whole trajectory is expected. We define the trajectory as τ which donates a state-action sequence from the start of the time horizon to the end, i.e. $a_1, s_1, ..., a_T, s_T$. Then the reward of the trajectory can be expressed as $\sum_{t=0}^{T} R(s_t, a_t)$. The expected reward of the policy π_ϑ can be written as:

$$J(\vartheta) = \mathbb{E}\left[\sum_{t=0}^{T} R(s_t, a_t); \pi_\vartheta\right] = \sum_{\tau} \Pr(\tau; \vartheta) R(\tau) \tag{32}$$

Our goal is to learn the parameter ϑ in the neural network that can maximize the expected reward.

$$\max_{\vartheta} J(\vartheta) = \max \sum_{\tau} \Pr(\tau; \vartheta) R(\tau) \tag{33}$$

In the same way, the gradient can be expressed as:

$$\nabla J(\vartheta) = \frac{1}{m} \sum_{i=1}^{m} \nabla_\vartheta \log\left(\Pr\left(\tau^{(i)}; \vartheta\right)\right) R\left(\tau^{(i)}\right) \tag{34}$$

where,

$$\begin{aligned} \nabla_\vartheta \log\left(\Pr\left(\tau^{(i)}; \vartheta\right)\right) &= \nabla_\vartheta \log\left[\prod_{t=1}^{T} \underbrace{\Pr\left(s_{t+1}^{(i)} \mid s_t^{(i)}, a_t^{(i)}\right)}_{Transition\ \ Function} \cdot \underbrace{\pi_\vartheta\left(a_t^{(i)} \mid s_t^{(i)}\right)}_{Policy}\right] \\ &= \nabla_\vartheta \sum_{t=1}^{T} \log\left(\pi_\vartheta\left(a_t^{(i)} \mid s_t^{(i)}\right)\right) \end{aligned} \tag{35}$$

From (34), it can be concluded that the gradient is not related to the transition function.

The reward $R\left(\tau^{(i)}\right)$ in (34) can be obtained through Monte Carlo by calculating an average value. However, the space state of the trajectory will be very large in some cases, which brings technical barriers to convergence. Thus, the Q-value is introduced to estimate the value of the policy. The Q-value is a mapping from the current action and state, which is expressed by a deep neural network known as a critic network. The structure of the critic network is similar to the DQN. Finally, the action is determined by the Q-value and the policy with the largest probability. The gradient can be further transferred to (36).

$$\nabla J(\vartheta) = \mathbb{E}\left[\nabla_\vartheta \pi_\vartheta(s) \nabla_a Q_\pi(s,a)|_{a=\pi_\vartheta(s)}\right] \tag{36}$$

As for the critic network, the parameter ϖ is updated based on the loss function (37), which is the mean square error between the real reward and the estimated reward.

$$L(\varpi) = \frac{1}{M} \sum_{i=1}^{M} \left(y^{(i)} - Q\left(s^{(i)}, a^{(i)} \mid \varpi\right)\right)^2 \tag{37}$$

C. Structure of the Deep Neural Network

As discussed in the previous part, there are two types of networks, namely the actor network and the critic network. The actor network is responsible for the action determination of the EVs and TCLs. The input for the actor network is the observed states, while the output is the action. The critic network is responsible for evaluating the performance of the charging or discharging behaviors. The input of the critic network has two channels, including states and action, while the output is the Q-value.

For each type of network, it contains the main network and target network. Thus, there are four networks in the proposed structure as Fig. 1, namely, actor main network, actor target network, critic main network, and the critic target network. The training process of these four networks is introduced briefly. For the critic network, the training process is similar to DQN based on the loss function (37). The value $Q\left(s^{(i)}, a^{(i)} \mid \varpi\right)$ in (37) is obtained from the critic main net, where action a is transited from actor main net. The real Q-value $y^{(i)}$ is obtained from the real current reward and the estimated next reward

from the critic target net. The estimated next reward is obtained based on the next state in the experience buffer and the next action transited from the actor target net. The actor network is trained according to (34) to earn the largest Q-value evaluated by the critic network. The target networks have the same structure as the main networks. The parameter of the main network will be updated to the target network regularly [27].

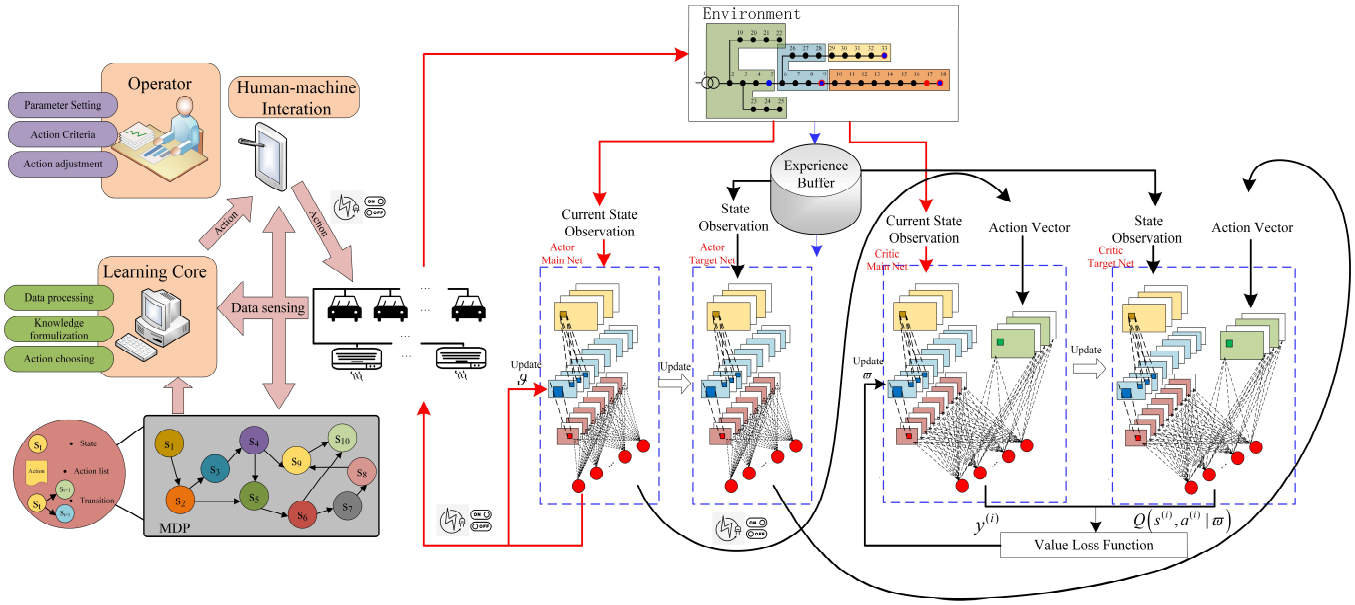

Fig. 1. Human-machine reinforcement learning framework based on DDPG

Table I. Network structure and parameters

Actor			Critic				
Index	Layer	Parameter	Index	Layer	Parameter	Layer	Parameter
				State channel		Action channel	
1	Input	S*1	1	Input	S*1	Input	1*1
2	FC	512, Sigmoid	2	FC	512, Sigmoid	FC	512, Sigmoid
3	Conv	3*1,256, ReLu	3	Conv	3*1,256, ReLu	Conv	3*1,256, ReLu
4	Conv	5*1,128, ReLu	4	Conv	5*1,128, ReLu	Conv	5*1,128, ReLu
5	Conv	1*1,64, ReLu	5	Conv	1*1,64, ReLu	Conv	1*1,64, ReLu
6	Output: FC	1, Linear	6	Output: FC		1, Linear	

The network structure is shown in Tab. I. In Tab.1, "FC" represents a fully-connection layer, where the number "512" represents that it has 512 units, and "sigmoid" is its activation function. "Conv" represents a convolutional layer, "×*×" represents the size of the convolutional kernel, followed by the number of kernels, such as "256", and the "ReLu" activation function. As for the input layer, "×*×" represents the size of the input data, where "S" represents the dimension of the states. The fully-connection layer is finally used as an output layer, where a linear function is used to make the vector become a scalar.

During the online training, the agents will first observe the current states and choose actions based on the actor main net. Then the agents will interact with the environment to transit to the new states and receive immediate rewards. The actions, states, and rewards will be saved to the experience buffer. The experience buffer is a queue, which means that the bottom elements will leave the queue when the buffer is out of memory. To update ϑ and ϖ in the deep neural network, take M samples from the experience buffer to calculate $y^{(i)}$. The actor network and the critic network are updated online, according to (34) and (37).

V. Human-Machine Learning Algorithm

Algorithm 1: Human-Machine Based DDPG

Initialize ϑ and ϖ in actor (A) and critic (Q) main network
Initialize actor (A') and critic (Q') target network according to $\vartheta' \leftarrow \vartheta, \varpi' \leftarrow \varpi$
Initialize the experience buffer R
for episode 1 to M **do**
  Initialize a random process N for action exploration
  Receive initial observation state s_1
  for t=1 to T **do**
    Select *exploration strategy* and set ε in (43)
    if *human exploration*
      then Get *criteria action* $a_t = \text{argmin}\, F(a_t, s_t)$
    else Select action a_t according to current policy and *exploration strategy*
      if action a_t from machine is rejected
        then Activate *emergency control* by (40)
    Execute action a_t and observe *ramp-up reward* r_t
    Store transition $\langle s_t, a_t, r_t, s_{t+1} \rangle$ in R
    Sample a random minibatch of N transitions $\langle s^{(i)}, a^{(i)}, r^{(i)}, s^{(i+1)} \rangle$ from R
    Calculate $y^{(i)}$ by $y^{(i)} = r^{(i)} + \gamma Q'\left(s^{(i+1)}, a^{(i+1)} \mid \varpi'\right)$
    Update main critic network by minimizing loss (37)
    Update actor policy by the sampled policy gradient (34)
    Update the target network every τ episodes
  end for
end for

In the traditional RL framework, the agents will choose a policy to interact with the environment without any constraints,

which will cause huge losses during the exploration process. The machine is superior to the human in computation, information storage, and big data analysis, while human shows advantages in reasoning and the reactions to the changes. Hence, human-machine learning is introduced to help the machine accelerate the learning process and find a preferable policy [28]. The human-machine learning framework is shown in Fig. 1. The human interacts with the machine through an interface where humans can intervene in the machine learning and adjust the action signal in real-time. In this section, the human-machine learning process will be reflected from three perspectives: emergency control, reward ramp-up, and human exploration. The Pseudo-code of the human-machine based DDPG algorithm is given.

A. Emergency control

When the emergency happens, the agents may choose to give up the opportunity of the voltage regulation to sacrifice the immediate rewards since the agents consider the rewards of the whole trajectory. An emergency refers to but is not limited to a sudden drop of PV output, a sudden increase of load, an outage of a feeder, etc. These unexpected situations in the power grids will lead to changes in the states of the grids, such as voltage violation or frequency fluctuation. Under this situation, the bad quality of electricity is unexpected, and hence human is required to intervene in the decisions made by machines through emergency control. The simplest idea of emergency control is to reject the actions decided by the machine and produce the control signal manually. There are many ways to regulate the voltage in an emergency. In this paper, we have utilized a control method based on the *V/P* sensitivity, which is a rule-based method.

The *V/P* sensitivity is calculated according to:

$$\begin{bmatrix}\mathbf{P}\\ \mathbf{Q}\end{bmatrix}=\begin{bmatrix}\mathbf{J}_{\mathbf{P\theta}} & \mathbf{J}_{\mathbf{PV}}\\ \mathbf{J}_{\mathbf{Q\theta}} & \mathbf{J}_{\mathbf{QV}}\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{\theta}\\ \mathbf{V}\end{bmatrix} \tag{38}$$

$$\Xi_{VP}=\frac{\partial V}{\partial P}=\left(\mathbf{J}_{\mathbf{PV}}-\mathbf{J}_{\mathbf{P\theta}}\mathbf{J}_{\mathbf{Q\theta}}{}^{-1}\mathbf{J}_{\mathbf{QV}}\right)^{-1} \tag{39}$$

If an emergency happens, the operator can choose to reject the decisions from the machine and input orders to control the agents. For example, when the downward violation of voltage happens, the human can activate emergency control to enforce the EVs discharging more energy through V2G according to:

$$\widetilde{\mathbf{P}}_{i,t}^{EV,DIS*}=\widetilde{\mathbf{P}}_{i,t}^{EV,DIS}+\left[\left(V^{e}-\underline{V_i}\right)\Big/\Xi_{VP}\right] \tag{40}$$

B. Reward ramp-up

During the beginning of the training, the agents select the actions randomly based on the initial random policy. The training process of RL is through "trial and error", which will result in a low reward at the beginning. One method to solve this problem is to pre-train the networks in an offline manner, and the network can be fine-tuned itself in real-time. With the development of digital twin technology, it will become a promising method. However, the limitation is that it relies highly on offline training data. In this paper, we proposed another method to speed up the online training process, which is reward ramp-up through human intervention. At the beginning of the training, the human can provide criteria actions that are based on the experience or optimization result. Then in the reward function, a penalty will be given between the action decided through reinforcement learning and the criteria actions shown as (41).

$$\begin{aligned}r_5=&-\xi_3\left(\sum_{m\in\Omega_i^{EV}}\widetilde{P}_{i,m,t}^{EV,DIS}-\sum_{m\in\Omega_i^{EV}}P_{i,m,t}^{EV,DIS}\right)^2\\&-\xi_3\left(\sum_{m\in\Omega_i^{TCL}}\widetilde{P}_{i,m,t}^{TCL}-\sum_{m\in\Omega_i^{TCL}}P_{i,m,t}^{TCL}\right)^2\end{aligned} \tag{41}$$

When the RL is closed at convergence, the ramp-up reward (41) can still be retained. If there is a deviation between the control signal from the machines and the real, a large penalty will be given. The deviation may be caused by system congestion, EVs driving behaviors, etc.

Another reward ramp-up method that can improve the performance of the machine is introducing an individualized risk parameter Λ. The parameter Λ is a positive number donating the degree of risk aversion. The more risk-averse the agent is, the larger Λ will be. Then the reward function (25) will be changed to:

$$r_1=\begin{cases}-\xi_1\left(V_{i,t}-\overline{V_i}\right)\Big/\exp(-\Lambda) & ,if \quad V_{i,t}>\overline{V_i}\\ -\xi_1\left(\underline{V_i}-V_{i,t}\right)\Big/\exp(-\Lambda) & ,if \quad V_{i,t}<\underline{V_i}\\ \xi_2 & ,else\end{cases} \tag{42}$$

C. Human exploration

The exploration strategy is an essential part of the model-free algorithm. In fact, the entire space is very large for some problems and can hardly be fully explored. As for machine exploration, there are mainly three kinds of exploration strategies, i.e., random exploration, greedy exploration, and increased greedy exploration strategy. In the DDPG algorithm, the exploration strategy is deterministic, which is based on the greedy algorithm. However, this method may cause the machine to be trapped in local optimal. Thus, in the human-machine framework, the operator is allowed to enforce the machine to conduct a random exploration or changed the parameter ς for an increased greedy exploration strategy shown as (43). The random exploration probability will be gradually decreased with the increase of episodes and is affected by ς. Another human exploration strategy is using inputs from human experts. This exploration method can not only accelerate the learning process but also help the machine find the optimal.

$$\Pr(Random\_explore)=1\Big/\log\left((episodes+9)\times\varsigma\right) \tag{43}$$

where ς is a positive number larger than 1.

VI. Case Studies

A. Experiment Setting

The proposed human-machine reinforcement learning for energy management strategies of EVs and TCLs is tested on the IEEE 33-bus system. First, simulations were conducted for

three aggregators located at buses 18, 22, 33. To show the scalability of the proposed method, we increase the number of aggregators. Further, the proposed method is tested on the IEEE 123-bus system. The main parameters utilized in the simulation are given in Tab. II. The simulations were completed by a PC with an Intel Core (TM) i7-9750 CPU @ 2.60 GHZ with 16.00 GB RAM, GTX GeForce 1660Ti.

Table II. Main Parameter Settings

EVs parameters					
η^{cha}, η^{dis}	0.92	$\overline{P^{EV,CHA/DIS}}$	6-12 kW	$\overline{E_{i,m}^{EV}}$	20-30 kWh
$W_{i,m}$	12.3-16.2 kWh/100km	φ^{BAT}	0.125 $/kWh	ε	0.2
TCLs parameters					
R	1.2-2.5 ºC/kW	C	1.7-3.4 kJ/ºC	δ	3-6°C
T_{set}	24-26°C				
Electricity Network and DRL parameters					
V_{max}	1.05	V_{min}	0.95	ξ_1	10
$\xi 2$	1	ξ_3	999	Λ	3

B. Comparison between Proposed Human-machine DDPG with Conventional DRL Algorithms

To verify the performance of the proposed human-machine DDPG algorithm, the training process is repeated 10 times, shown in Fig. 2. It can be concluded that the algorithm reaches convergence with 50 to 100 episodes. The unstable situations may occasionally happen during the training process, such as the third test. But it can be adjusted rapidly.

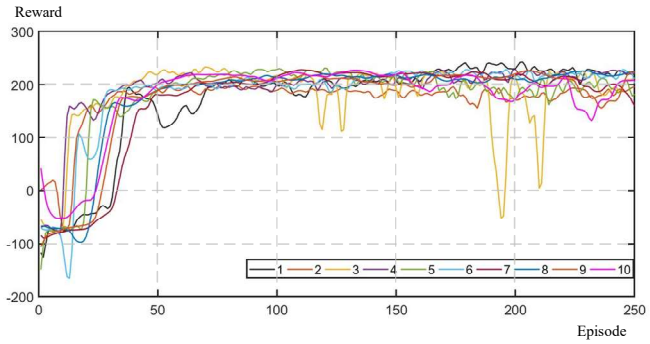

Fig. 2. Ten times training process of human-machine DDPG algorithm

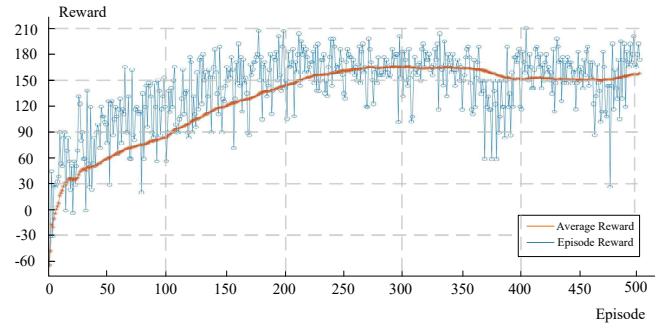

Fig. 3. Reward evolution of conventional DDPG algorithm

In this section, we compare the proposed human-machine DRL algorithm with the conventional DDPG [24] and DQN [20] algorithms utilized in the literature. Figures 3-5 show the convergence process of DDPG, DQN, and the proposed human-machine DDPG algorithm, respectively. It can be concluded that the proposed method has the quickest convergence (73 episodes) speed because of the ramp-up reward where human provides action criteria at the early stage of the training. Besides, the proposed method has the largest reward after convergence. That is because human exploration helps the machine travel more solution space and prevent from being trapped in local optimal.

The percentage of human intervention during the training process is shown in Fig. 6. During the first one hundred iterations, the human intervention percentage is very high, because human is providing action criteria to accelerate the training in this stage. During the second one hundred iterations, the human intervention percentage decreases dramatically, and at this stage, most intervention comes from human exploration, which enriches the solution space in the experience buffer. During the rest iteration, the algorithm is closed to convergence, and the human intervention percentage is relatively low. At this stage, most of the interventions are emergency controls.

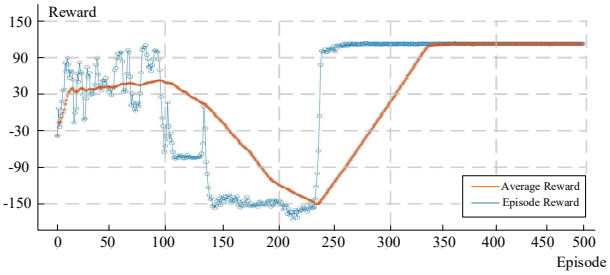

Fig. 4. Reward evolution of conventional DQN algorithm

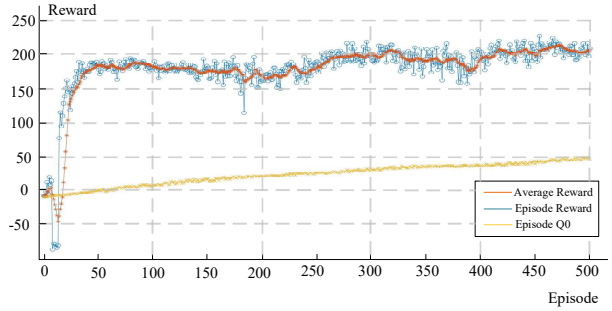

Fig. 5. Reward evolution of human-machine DDPG algorithm

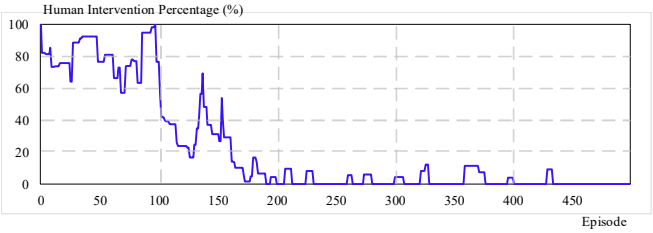

Fig. 6. Human intervention percentage of the proposed method

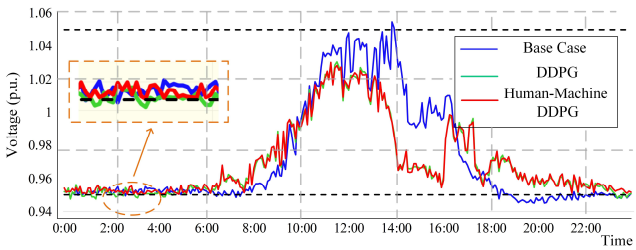

Fig. 7. Local voltage profile comparison under normal situation

To further explore the benefits of the proposed human-machine DDPG algorithm, the local voltage profile under different cases is shown in Figs. 7-9. In Fig. 7, the training episode is after 500 iterations, which means both algorithms have converged. In the base case, the coordinated dispatch of EVs and TCLs is not considered. From the result, it can be concluded that through reinforcement learning, the local voltage regulation can be achieved by the dispatch of EVs and TCLs. However, the proposed method can accomplish the task more accurately. For example, when applying traditional DDPG [24], there is a slight voltage violation from 2:00 to 4:00.

Figure 8. shows the performance of the man-machine DDPG algorithm under an emergency. There is a sudden drop in PV output at 14:00, causing the downward voltage violation. For traditional DDPG, the machine chooses to give up the immediate reward to earn the largest reward of the whole trajectory. However, in the proposed method, the human can intervene with the machine by rejecting the action signals produced by the machine and give emergency control signals manually. Thus, the proposed method is more flexible and can better cope with the emergency.

Figure 9. shows the performance of DDPG and human-machine DDPG before the convergence. During the early stage of training, voltage violation frequently happens due to the improper control signals from the machine. These bad actions not only lead to a low reward but influence the quality of the electricity. Under these conditions, the base case will be better than the conventional DDPG and the proposed method. To prevent it from happening, human exploration can adjust the control signals from the machine and provides action criteria. Hence, for human-machine DDPG, the voltage violation still happens occasionally, but the large violation can be avoided.

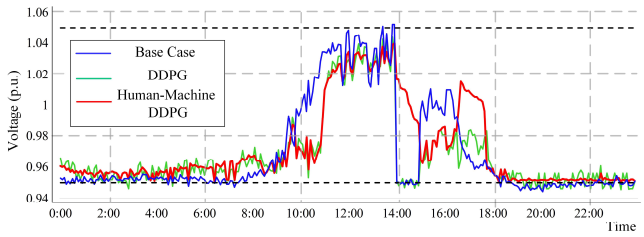

Fig. 8. Local voltage profile comparison under an emergent situation

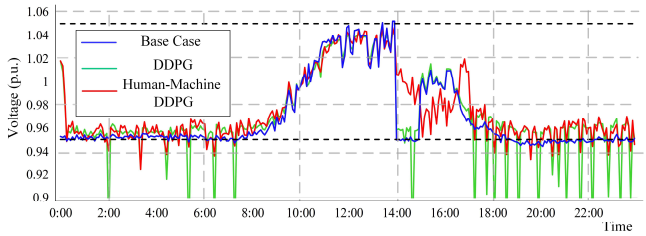

Fig. 9. Local voltage profile comparison before convergence

C. Comparison between RL results and optimization results

To investigate the performance of energy management strategies under the RL framework, the proposed method is compared with the optimization result according to the method utilized in [3]. The difference between the cost minimization solutions obtained for three aggregators correlated with their average electrical pattern is shown in Figs. 10-12. The black dotted curves are the dispatch solutions according to the optimization results. It can be concluded that the deviations between the results under the proposed human-machine RL framework and the optimal results are relatively small. The average daily cost of energy consumption for all aggregators is summarized in Tab. III. Specifically, we analyze the results from the optimization, DQN, DDPG, and the proposed human-machine DDPG method at a different scale of aggregators. The average daily cost of the proposed human-machine DDPG method is close to the optimal result and is superior to DQN and DDPG. In the case of 20 aggregators, the average daily cost from the proposed method is higher than the optimal value by 3.9%, while the average daily cost from the DQN and DDPG method is 30.1% and 20.2% higher than the optimal value.

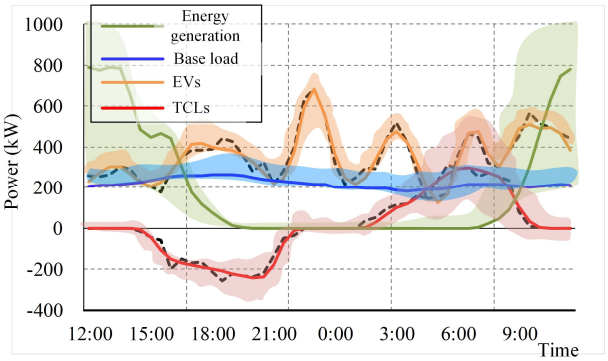

Fig. 10. Energy management profile of aggregator one

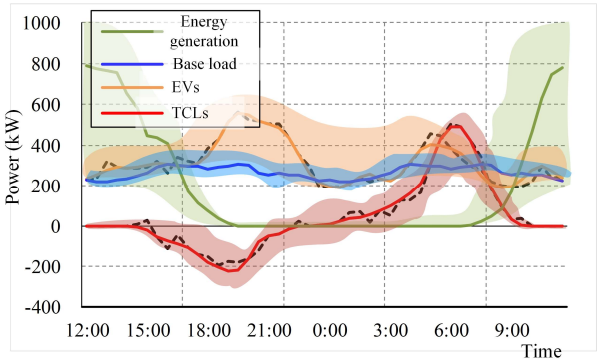

Fig. 11. Energy management profile of aggregator two

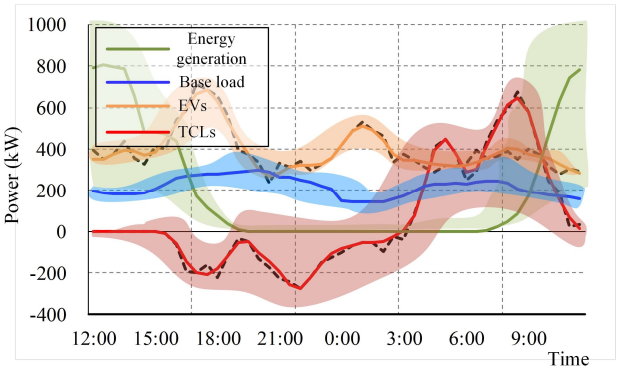

Fig. 12. Energy management profile of aggregator three

Table III. Cost comparison between different algorithms

	Method	Number of aggregators		
		3	10	20
Average Cost ($)	Optimization	3132.78	10253.46	20229.24
	DQN	4042.14	13875.25	26321.97
	DDPG	3856.12	13110.24	24312.86
	Human-Machine DDPG	3227.18	10869.45	21019.64

D. Scalability and efficiency analysis

To further analyze the scalability and the efficiency of the proposed method, the average training episode, average computation time per episode, and the average training time under a different number of aggregators are summarized in Tab. IV. There are two criteria to evaluate the efficiency of the DRL-based algorithm, i.e., sampling efficiency and computation efficiency. The sampling efficiency is a measurement of the optimality of a sampling strategy. When referring to DRL, a more efficient sampling strategy requires less amount of experience (measured in training episodes) to reach convergence and a certain level of performance. In the power system, if the sampling efficiency is low, the operation security in the real-time application can not be guaranteed. As for the computation efficiency, it is measured by the time to reach the convergence, i.e., asymptotic time. Although the energy cost of deep learning is not obvious, learning with low computation efficiency will continuously consume massive of energy. The main energy consumption mainly comes from the operation of the servers and their cooling systems. Therefore, computation costs can be saved by improving computation efficiency.

We have compared the sampling efficiency and computation efficiency of the proposed algorithm and the other algorithms in the state-of-art work, shown in Tab. IV. When the number of aggregators is increased to 20, it takes 1642 and 2865 episodes for DQN and DDPG to reach convergence, while the proposed method only takes 267 episodes. It indicates that the sampling efficiency of the human-machine DDPG is superior to the conventional DRL-based algorithms. The average computation time per episode of the proposed method is close to that of DDPG, while DQN has the lowest average computation time per episode. This is because the average computation time per episode is mostly affected by the complexity of the algorithm. From the aspect of average training time, it can be found that the proposed human-machine DDPG has the shortest average training time in all situations. When the number of the aggregators is 3, it takes 18 minutes and 51 minutes for DQN and DDPG to reach the convergence, which is more than twice and six times longer than the proposed method. When the number of aggregators is increased to 20, it takes 104 minutes and 315 minutes for DQN and DDPG to reach the convergence, which is around three times and ten times longer than the proposed method. It indicates that the proposed method has the highest computation efficiency. The average training time will be affected by both sampling efficiency and algorithm complexity.

Besides, it should be noted that the direct comparison of computation and sample efficiency between different DRL algorithms in a single task is not completely fair because there may be many different variables in the implementation. One algorithm may converge earlier, but it cannot reach the same score as other slower algorithms. Therefore, we need to comprehensively evaluate the performance of the DRL algorithm. For example, although the sampling efficiency and computation efficiency of DQN are higher than that of DDPG, DDPG can reach a higher average reward after convergence, as shown in Figs. 3 and 4. Hence, the performance of these two algorithms needs to be further evaluated. However, in our proposed method, the efficiency is higher than the conventional algorithm, and a higher average reward after convergence is also expected, as shown in Fig. 5. It indicates that human intervention can improve the convergence capability and preferable result exploration of the DRL algorithm.

Tab. V shows the decision-making time of the DRL-based method and the optimization-based method. All the methods under the RL framework take very little time. It takes only several milliseconds for each time of forward propagation. For human-machine DDPG, sufficient time should be reserved for the human reaction. However, for the optimization method, the optimization process needs to be re-run for every decision, and the computational time will increase with the problem scale.

Then, we further verified the proposed method in the IEEE 123-bus system. Figure 13 shows the average training episodes under a different number of aggregators. The training episodes of all methods rise with the increase of the aggregators. The proposed human-machine DDPG algorithm always converges faster than DQN and DDPG. It is essential in the online training process since huge losses can be avoided during the beginning of the learning.

Table IV. Efficiency Comparision between different algorithms

	Method	Number of aggregators		
		3	10	20
Average Training Episodes	DQN	312	598	1642
	DDPG	486	892	2865
	Human-Machine DDPG	73	125	267
Average Computation Time per Episode	DQN	3.5s	3.8s	3.8s
	DDPG	6.3s	6.5s	6.6s
	Human-Machine DDPG	6.6s	6.8s	7.1s
Average Training Time	DQN	18m	38m	104m
	DDPG	51m	97m	315m
	Human-Machine DDPG	8m	14m	33m

Table V. Decision marking time comparison between different algorithms

Average Decision Time	Optimization	52s	3min26s	10min07s
	DQN	<1s	<1s	<1s
	DDPG	<1s	<1s	<1s
	Human-Machine DDPG	1-5s	1-5s	1-5s

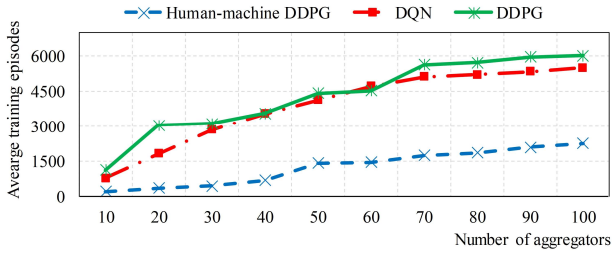

Fig. 13. Average training episodes of different methods in the IEEE 123-bus system

VII. CONCLUSION

In this paper, we have proposed a human-machine reinforcement learning framework based on DDPG for the energy management of EVs and TCLs. Through the proposed method, we aim to help the EVs and TCLs aggregator reach the lowest energy consumption cost and maintain the local voltage mag-

nitude at the same time. First, we formulate the Markov decision process for the BESS and VESS based on conventional DDPG. To further improve the online learning performance of reinforcement learning, human-machine reinforcement learning is introduced to accelerate the learning process and help the machine explore solutions widely. In the case study, we first compare the proposed method with the DQN and DDPG algorithms. It is verified that the proposed method is superior to the conventional DRL algorithm in convergence ability and the capability of dealing with emergencies. Then we compare the proposed DRL-based method with the optimization result. It can be concluded that the proposed method has slight deviations from the optimal solution but dramatically reduces the decision-making time, which is more capable of adapting to time-vary changes in the future smart grids.

REFERENCES

[1] G. Liu, Y. Tao, L. Xu, Z. Chen, J. Qiu, and S. Lai, "Coordinated management of aggregated electric vehicles and thermostatically controlled loads in hierarchical energy systems," *International Journal of Electrical Power & Energy Systems,* vol. 131, p. 107090, 2021.

[2] O. Sundstrom and C. Binding, "Flexible charging optimization for electric vehicles considering distribution grid constraints," *IEEE Transactions on Smart grid,* vol. 3, no. 1, pp. 26-37, 2011.

[3] K. Qian, C. Zhou, M. Allan, and Y. Yuan, "Modeling of load demand due to EV battery charging in distribution systems," *IEEE Transactions on Power Systems,* vol. 26, no. 2, pp. 802-810, 2010.

[4] E. Sortomme and M. A. El-Sharkawi, "Optimal scheduling of vehicle-to-grid energy and ancillary services," *IEEE Transactions on Smart Grid,* vol. 3, no. 1, pp. 351-359, 2011.

[5] L. Cheng, Y. Chang, and R. Huang, "Mitigating voltage problem in distribution system with distributed solar generation using electric vehicles," *IEEE Transactions on Sustainable Energy,* vol. 6, no. 4, pp. 1475-1484, 2015.

[6] M. Wang, Y. Mu, F. Li, H. Jia, X. Li, Q. Shi, and T. Jiang, "State space model of aggregated electric vehicles for frequency regulation," *IEEE Transactions on Smart Grid,* vol. 11, no. 2, pp. 981-994, 2019.

[7] R. Zhang, X. Cheng, and L. Yang, "Flexible energy management protocol for cooperative EV-to-EV charging," *IEEE Transactions on Intelligent Transportation Systems,* vol. 20, no. 1, pp. 172-184, 2018.

[8] N. Lu and Y. Zhang, "Design considerations of a centralized load controller using thermostatically controlled appliances for continuous regulation reserves," *IEEE Transactions on Smart Grid,* vol. 4, no. 2, pp. 914-921, 2012.

[9] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Aggregate flexibility of thermostatically controlled loads," *IEEE Transactions on Power Systems,* vol. 30, no. 1, pp. 189-198, 2014.

[10] Y. Yao and P. Zhang, "Unified control strategy of heterogeneous thermostatically controlled loads with market-based mechanism," *Journal of Modern Power Systems and Clean Energy,* vol. 8, no. 6, pp. 1231-1239, 2020.

[11] M. Song, W. Sun, Y. Wang, M. Shahidehpour, Z. Li, and C. Gao, "Hierarchical scheduling of aggregated TCL flexibility for transactive energy in power systems," *IEEE Transactions on Smart Grid,* vol. 11, no. 3, pp. 2452-2463, 2019.

[12] Y. Wan, C. Long, R. Deng, G. Wen, X. Yu, and T. Huang, "Distributed event-based control for thermostatically controlled loads under hybrid cyber attacks," *IEEE Transactions on Cybernetics,* 2020.

[13] J. H. Zhao, Z. Y. Dong, X. Li, and K. P. Wong, "A framework for electricity price spike analysis with advanced data mining methods," *IEEE Transactions on Power Systems,* vol. 22, no. 1, pp. 376-385, 2007.

[14] B. Li, J. Zhang, Y. He, and Y. Wang, "Short-term load-forecasting method based on wavelet decomposition with second-order gray neural network model combined with ADF test," *IEEE Access,* vol. 5, pp. 16324-16331, 2017.

[15] V. Le, X. Yao, C. Miller, and B.-H. Tsao, "Series DC arc fault detection based on ensemble machine learning," *IEEE Transactions on Power Electronics,* vol. 35, no. 8, pp. 7826-7839, 2020.

[16] J. M. Gillis, S. M. Alshareef, and W. G. Morsi, "Nonintrusive load monitoring using wavelet design and machine learning," *IEEE Transactions on Smart Grid,* vol. 7, no. 1, pp. 320-328, 2015.

[17] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot, "Mastering the game of Go with deep neural networks and tree search," *Nature,* vol. 529, no. 7587, pp. 484-489, 2016.

[18] L. Yin, T. Yu, L. Zhou, L. Huang, X. Zhang, and B. Zheng, "Artificial emotional reinforcement learning for automatic generation control of large-scale interconnected power grids," *IET Generation, Transmission & Distribution,* vol. 11, no. 9, pp. 2305-2313, 2017.

[19] T. Qian, C. Shao, X. Wang, and M. Shahidehpour, "Deep reinforcement learning for EV charging navigation by coordinating smart grid and intelligent transportation system," *IEEE Transactions on Smart Grid,* vol. 11, no. 2, pp. 1714-1723, 2019.

[20] S. Najafi, M. Shafie-khah, P. Siano, W. Wei, and J. P. S. Catalão, "Reinforcement learning method for plug-in electric vehicle bidding," *IET Smart Grid,* vol. 2, no. 4, pp. 529-536, 2019.

[21] H. Zhao, J. Zhao, J. Qiu, G. Liang, and Z. Y. Dong, "Cooperative wind farm control with deep reinforcement learning and knowledge assisted learning," *IEEE Transactions on Industrial Informatics,* 2020.

[22] Z. Yan and Y. Xu, "A multi-agent deep reinforcement learning method for cooperative load frequency control of multi-area power systems," *IEEE Transactions on Power Systems,* vol. 35, no. 6, pp. 4599 - 4608, 2020.

[23] 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 Transactions on Smart Grid,* vol. 10, no. 4, pp. 3698-3708, 2018.

[24] T. Ding, Z. Zeng, J. Bai, B. Qin, Y. Yang, and M. Shahidehpour, "Optimal electric vehicle charging strategy with Markov decision process and reinforcement learning technique," *IEEE Transactions on Industry Applications,* vol. 56, no. 5, pp. 5811-5823, 2020.

[25] M. Song, C. Gao, M. Shahidehpour, Z. Li, J. Yang, and H. Yan, "Impact of uncertain parameters on TCL power capacity calculation via HDMR for generating power pulses," *IEEE Transactions on Smart Grid,* vol. 10, no. 3, pp. 3112-3124, 2018.

[26] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet of Things Journal,* vol. 6, no. 5, pp. 8577-8588, 2019.

[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski, "Human-level control through deep reinforcement learning," *Nature,* vol. 518, no. 7540, pp. 529-533, 2015.

[28] S. Doltsinis, P. Ferreira, and N. Lohse, "A symbiotic human–machine learning approach for production ramp-up," *IEEE Transactions on Human-Machine Systems,* vol. 48, no. 3, pp. 229-240, 2017.

Yuechuan Tao received the B.Sc. degree in Electrical Engineering and Automation from Shanghai Normal University, Shanghai, China, in 2017, and the M.Sc. degree in Electrical Engineering from the University of Sydney, Australia in 2019, and are currently working toward the Ph.D. degree in the University of Sydney, Australia. His main fields of interest include power system operation and planning, electric vehicles, data-driven approaches, smart grid, etc.

Jing Qiu (M'14) is currently a Senior Lecturer in Electrical Engineering at the University of Sydney, Australia. He obtained his B.Eng. degree in control engineering from Shandong University, China, M.Sc. degree in environmental policy and management, majoring in carbon financing in the power sector, from The University of Manchester, U.K., and Ph.D. in electrical engineering from The University of Newcastle, Australia, in 2008, 2010 and 2014 respectively. His areas of interest include power system operation and planning, energy economics, electricity markets, and risk management. He is the Editorial Board Member of IET Energy Conversion and Economics

Shuying Lai received the B.Sc. degree in Finance, Accounting and Management from the University of Nottingham, Ningbo, China, in 2017, and the M.Sc. degrees in Finance, Accounting from the University of Sydney, Australia in 2019, and are currently working toward the Ph.D. degree in the University of Sydney, Australia. Her main fields of interest include risk management, energy sharing and pricing, electricity derivatives, and transactive energy, etc.

Xian Zhang (M'19) received the B.Sc. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2009, the M.Sc. degree in electrical engineering from Tsinghua University, Beijing, in 2012, and the Ph.D. degree in electrical engineering from The Hong Kong Polytechnic University, Hong Kong, in 2019. She is currently an Assistant Professor with the Harbin Institute of Technology (Shenzhen), Shenzhen, China. Her main fields of interest include smart grids and electric vehicles.

Yunqi Wang obtained his B.E. degree in control engineering from Northeast Electric Power University, China, M.E. degree in electrical engineering from the University of Sydney, Australia, in 2014 and 2018, respectively. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Sydney, Australia. His research interests include power system operation and planning, low-carbon energy economy, and demand-side management.

Guibin Wang (M'16) received his B.E. and Ph.D. degrees in electrical engineering from Shandong University, Jinan, China and Zhejiang University, Hangzhou, China, in 2009 and 2014, respectively. From 2011 to 2014, he was also a Research Assistant in the Department of Electrical Engineering, Hong Kong Polytechnic University, Hong Kong. He is currently an Associate Research Professor in Shenzhen University, Shenzhen, China. His main research interests lie in electric vehicles and renewable energy.