Maintenance Decision Generator for Electrical Equipment Based on Reinforcement Learning

Ruifan Li*
Beijing University of Posts and Telecommunications, 100876, Beijing China
rfli@bupt.edu.cn

Zeyuan Wang
Beijing University of Posts and Telecommunications, 100876, Beijing China
wangzeyuan@bupt.edu.cn

Yifan Du
Beijing University of Posts and Telecommunications, 100876, Beijing China
yifandu@bupt.edu.cn

Zepeng Zhai
Beijing University of Posts and Telecommunications, 100876, Beijing China
zepeng@bupt.edu.cn

Yongping Xiong
Beijing University of Posts and Telecommunications, 100876, Beijing China
ypxiong@bupt.edu.cn

Ziqun Liu
State Grid Jiangsu Electric Power Co., Ltd. Research Institute, 211103, Nanjing, China
liuzq1@js.sgcc.com.cn

ABSTRACT

The maintenance of electrical equipment aims to guarantee a good working condition of the equipment and to increase the stability of the power grid operation. In this paper, we propose a novel maintenance decision model for both single and multiple electrical equipment, based on the Markov hypothesis of equipment states and reinforcement learning. Specifically, the cut set of the power grid is incorporated to calculate the weights of different equipment in multiple equipment mode, which are then applied to the dynamic programming solution to make the learned strategies focus on the differences between equipment. Moreover, the current cut set is used to recalculate action rewards to indirectly implement the communication between action sequences with the knowledge of that maintenance actions can lead to the changes of cut set. Experimental results demonstrate that the proposed method significantly improved the effectiveness of decision making.

CCS CONCEPTS

• **Computing methodologies** → *Planning for deterministic actions.*

KEYWORDS

Electric equipment maintenance, Reinforcement learning, Dynamic decision making.

ACM Reference Format:
Ruifan Li, Zeyuan Wang, Yifan Du, Zepeng Zhai, Yongping Xiong, and Ziqun Liu. 2021. Maintenance Decision Generator for Electrical Equipment Based on Reinforcement Learning. In *2021 4th International Conference on Signal Processing and Machine Learning (SPML 2021), August 18–20, 2021, Beijing, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3483207.3483233

*Corresponding Author.

SPML 2021, August 18–20, 2021, Beijing, China

ACM ISBN 978-1-4503-9017-0/21/08…$15.00
https://doi.org/10.1145/3483207.3483233

1 INTRODUCTION

Electrical equipment maintenance refers to the maintenance decision of equipment according to the running state of equipment and power grid in the course of power grid operation, so as to avoid equipment damage and subsequent effects. Traditionally, most of the maintenance methods for electrical equipment are based on manual decision [5, 8] and rule maintenance strategy [11, 17]. Manual decision usually requires much more professional knowledge of the power grid, and the rule maintenance strategy depends more on manually set weight, with limited generalization. In addition, electrical equipment maintenance can also be understood as a sequential decision-making problem, whose relationship between maintenance strategy and reward can be modeled by reinforcement learning. However, current popular reinforcement learning based on deep learning uses neural network to represent the state and model the strategy, which requires a lot data and is hard to be applied in professional domain.

Liu et al. [4] establish multiple states of electric equipment, and carries out first-order Markov hypothesis for the state of the equipment. Based on this assumption, this paper defines the reward of the equipment maintenance and uses reinforcement learning based on dynamic programming solution to find the best time to maintain the equipment and reduce the risk of equipment damage. In the maintenance decision of single equipment, the maintenance strategy can be obtained by the value matrix obtained by the traditional dynamic programming solution. In the case of multiple equipment, there are two problems: 1) The importance of each equipment to the overall operation of the power grid is not consistent. 2) The maintenance decision of equipment may influence the decision of other equipment. In order to solve the above problems, as shown in Fig. 1, we introduce the cut set to calculate the importance of equipment for power grid, and then apply it to dynamic programming solution. In addition, the maintenance of the equipment will cause the change of the cut set. In the maintenance decision of a certain device, a new cut set can be obtained according to the maintenance actions of other devices, and then the reward of each action can be recalculated according to the new cut set, so as to realize the

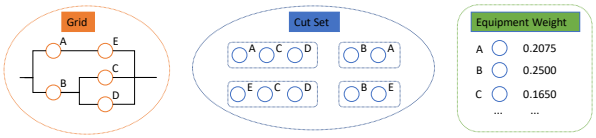


Figure 1: According to the connection of power grid, the cut set of power grid is manually obtained through the definition of cut set, and then the weight of each device is calculated with the contribution to the cut set.

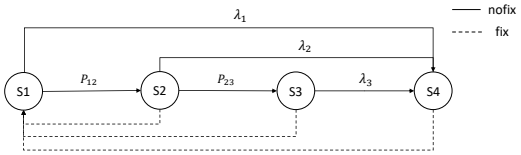


Figure 2: State transition of equipment.

decision-making communication of different equipment. Through simulation experiments, we verify the effectiveness of the proposed maintenance decision method and dynamic strategy generator.

In summary, our contributions can be highlighted as follows. 1) We design a maintenance decision model for power equipment based on reinforcement learning. 2) We employ the cut set to calculate the importance of power equipment, and integrate the obtained importance into reinforcement learning. 3) We propose a dynamic strategy generator to implement the communication among equipment according to the influence of action on the cut set.

2 RELATED WORK

2.1 Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning means that the whole system contains more than two agents, with a certain relationship between each other. Multi-agent reinforcement learning is applied in many fields such as games [1, 15], multi-robot collision avoidance [10, 14], autonomous driving [2, 18] and so on, though which we can get strategies beyond human experience. The deep-Q-network is applied to design a routing scheduling framework for multi-task agent[3], which corroborates the ability of Q-Learning to handle complex vehicle routing problems with several constraints. A scalable multi-agent reinforcement training platform was built to improve reinforcement learning effectiveness for heterogeneous collective cooperative decision making [7]. A multi-agent reinforcement learning approach is proposed to implement load frequency control without the need of a centralized authority in the power system [12]. And a multi-agent deep deterministic policy gradient algorithm is proposed to control multi-intersection traffic lights and the results of the contrast experiment imply the efficiency and stability [16].

2.2 Electric Equipment Maintenance

At present, the strategy of power equipment maintenance often depends on manual decision making. Technical personnel often have rich experience in power equipment management, and a lot of professional knowledge, who can judge subjectively whether the equipment need maintenance by scoring the status of the equipment. Methods in [9] depends mainly on on-line detection, off-line detection and periodic disassembly detection. In addition, [6] analyzes The irrationality of power maintenance and puts forward some strategies. However, the above strategies are all designed for people, requiring technicians to have rich experience, which are not only inefficient but also poor in generalization. A deep recursive q-network Multi-Agent Reinforcement Learning model is designed by applying deep reinforcement learning to power equipment maintenance, which has better optimization and decision-making abilities and a lower maintenance cost. This shows that it is reasonable to use reinforcement learning to make decision for equipment maintenance, which can often obtain a good decision.

3 METHOD

In this section, we will describe our proposed maintenance decision model for electric equipment. We first describe the Markov hypothesis and reward function of equipment in Section 3.1. Then we propose the maintenance decision model of single equipment in Section 3.2. In Section 3.3, we propose the maintenance decision model of multiple equipment.

3.1 Markov Hypothesis and Reward

3.1.1 Markov Hypothesis of Device. The equipment condition often transfers through multiple states to the fault state during the transition process. Fig. 2 shows the equipment state transfer process in the normal way. Here, we have defined four states, including S1, S2, S3 and S4, which ranges from a good state, a mild deterioration state, a severe deterioration state, to a damage state. P_{ij} denotes the transfer probability of state Si to state Sj, λ_i denotes the probability of state Si to damaged state $S4$. When the *nofix* operation is given, the state transition probability matrix $\mathbf{P}$ of the equipment is defined as:

$$\mathbf{P} = \begin{bmatrix} 1-P_{12}-\lambda_1 & P_{12} & 0 & \lambda_1 \\ 0 & 1-P_{23}-\lambda_2 & P_{23} & \lambda_2 \\ 0 & 0 & 1-\lambda_3 & \lambda_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (1)$$

The element state transition matrix refers to the probability of a device at time t to be transferred to different states at time $t+1$. The time unit is days. When the operation is *fix*, according to the goal of maintenance, the maintenance equipment certainly returns to S1 state.

3.1.2 Reward of Maintenance Action. The entire risk of power network is divided into maintenance risk R_M and the grid failure risk R_F. The maintenance risk can be further divided into two losses: the equipment maintenance loss $R_{M,1}$ and the equipment damage loss $R_{M,2}$. The grid failure risk R_F can also be divided as the maintenance risk. Therefore, the entire risk R_{sum} of the power grid can be given as:

$$R_{sum} = R_{M,1} + R_{M,2} + \sum_{t=1}^{N_T} \left(R_{F,1}(t) + R_{F,2}(t)\right), \quad (2)$$

where N_T is the number of operation days. The power failure loss of the power grid will constantly change with the operation power of the power grid. Thus, the reinforcement learning aims to minimize the overall loss of the power grid.

3.2 Maintenance Decision Method for Single Equipment

3.2.1 State & Action Definition. For single equipment, both time and the running state of the equipment constitute the state in reinforcement learning. Therefore, when the time length is T and the number of operating states of the equipment is N, there are $N \times T$ states in reinforcement learning. The action set is [*fix*, *nofix*]. In order to be more consistent with the actual situation, only the *nofix* action can be selected in the S1 state of the equipment, only the *fix* action can be selected in the S4 state of the equipment, and no restrictions can be imposed on the actions in other states of the equipment.

3.2.2 Dynamic Programming Solution. When there is only one device in the power grid, the failure of it will cause the loss of both power failure of the power grid and equipment maintenance. However, maintenance of equipment working in S2 or S3 state, will also lead to power failure loss and maintenance equipment loss. The current action has a global view with the help of evaluation of the risks by reinforcement learning optimized by dynamic programming. The dynamic programming optimization algorithm consists of two processes, strategy evaluation and strategy improvement. Strategy evaluation valuates the value of each state using the Bellman equation[13] with first randomly initialized strategy and the state transition probability of the device,

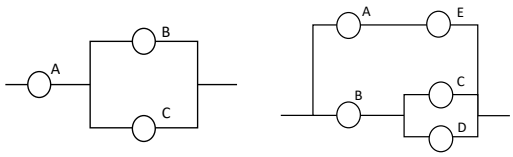
$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(\mathcal{R}^a_{s_{device_i}}(cut) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} v_k\left(s'\right) \right) \quad (3)$$

where π refers to the corresponding policy and R^a_s is the reward of action. $P_{ss\prime}$ is the probability of the current state transferring to $s\prime$ with action a. v_k represents the value of the current value matrix and γ is the discount factor. Then strategy improvement selects the action with the largest reward to improve the strategy. By continuous repetition of the two processes of strategy evaluation and strategy update, the value of different states gradually stabilizes, and finally the maintenance strategy of the equipment in different states is obtained. Here, R^a_s refers to Reward calculated in the following way.

$$R^a_s = -\left(R_M(s,a) + R_F(s,a)\right) \quad (4)$$

3.3 Multi-equipment Maintenance Decision Generator

The multi-equipment maintenance decision is usually considered as a multi-agent problem. In the process of modeling, we need to consider the relationship between equipment and global reward, and the maintenance action of equipment will cause the change of power grid state. Correspondingly, cut set is introduced to realize the function in a simple way.



(a) Grid simple connection (b) Grid complex connection

Figure 3: The dataset contains two types of grid connection.

3.3.1 State & Action definition. For multiple equipment, the state defined in reinforcement learning consists of equipment index, time and the running state of the equipment. Therefore, when the number of equipment in power gird is S, the time length is T and the number of operating states of the equipment is N, there are $S \times N \times T$ states in reinforcement learning.

3.3.2 Cut Set. The cut set of the power grid refers to the minimum combination of equipment that causes a power failure in the power grid, that is, if all equipment in the cut set is damaged, the power grid will be out of power. A grid usually contains multiple cut sets. Taking grid 1 as an example in Figure 3, the cut set combination of the power grid is $\{\{A\}, \{B, C\}\}$. When A fails, or B and C fail simultaneously, power failure will be caused. We introduce cut sets in order to model the importance of equipment to the change of power grid operating state in reinforcement learning. The operation of the cut set is the same as the operation of the set. Moreover, the cut set of the current grid is obtained by the subtraction between the original cut set and the set of damaged and fixed equipment.

3.3.3 Weighted Dynamic Programming Solution. The weight of each equipment in the power grid can be calculated through the cut set of the grid. With the weight applied to the reward of reinforcement learning, the strategy can focus on the importance of different equipment to the power grid. The weight of the equipment is calculated as follows:

$$W_{device_i}(cut) = \frac{1}{N} \sum_{m=1}^{N} \frac{1}{T_m} I(cut_m, device_i) \quad (5)$$

where T_m is the number of equipment in mth cut set, and $I(cut_m, device_i)$ is calculated as follows:

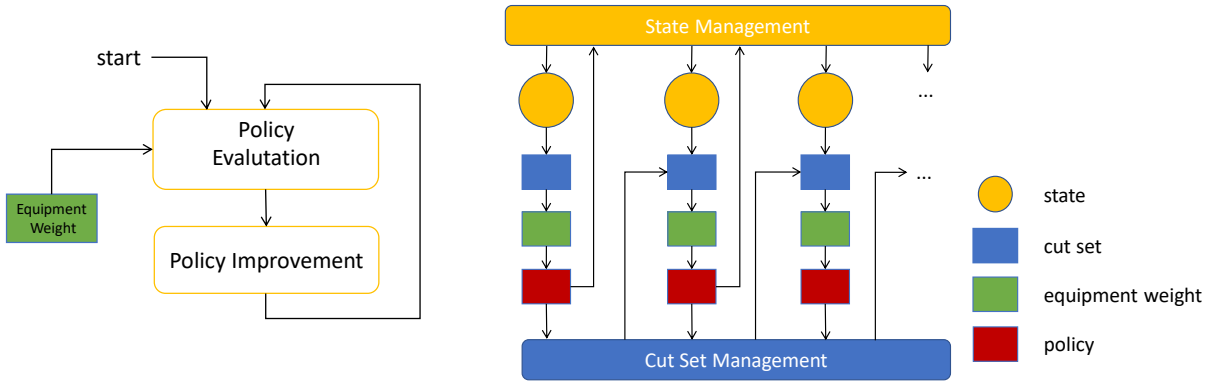
$$I(cut_m, device_i) = \begin{cases} 0 & device_i \in cut_m \\ 1 & device_i \notin cut_m \end{cases} \quad (6)$$

Therefore, when calculating the grid failure losses caused by equipment, the weights obtained by the cut set can be used to calculate the power grid risks.

$$R^a_{s_{device_i}}(cut) = -\left(R_M(s,a) + W_{device_i}(cut) \times R_F(s,a)\right) \quad (7)$$

The equipment's aware of grid connections can be obtained and reflected in decision-making by applying weight of equipment to dynamic programming.

3.3.4 Dynamic Policy Generator. The maintenance action of the equipment will also change the cut set. Taking the grid 1 as an example, when the equipment B is already in maintenance state, the maintenance of the equipment C will cause more grid failure



(a) Policy for single equipment

(b) Dynamic policy generator

Figure 4: The weight of the equipment can be weighted into the strategy evaluation process in Fig. (a). The decision-making process of Dynamic policy generator is shown in Fig. (b).

risks than before. As shown in (b) in Fig. 4, the maintenance of the equipment will cause the change of the cut set of the power grid. In the next state, the weight of each equipment needs to be calculated according to the current cut set. State management refers to the maintenance decision of the equipment in S2 and S3 state in turn. The dynamic strategy generation module can explicitly calculate the new reward for each action with the Value matrix and equipment weight as follows:

$$\pi\left(a \mid s_{device_i}\right) = \arg\max_{a} \left(\mathcal{R}^{a}_{s_{device_i}}(cut) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^{a}_{ss'} v_k\left(s'\right)\right) \tag{8}$$

The value matrix is obtained by dynamic programming solution in any way, and the cut-set state is based on the current power grid which the equipment node in maintenance state is regarded as open circuit. The action with the largest reward is selected as the strategy of the current state. By changing the cut set and integrating change into the decision-making process, indirect communication between multiple equipment can be realized.

4 EXPERIMENT

4.1 Dataset

The experimental result is obtained by simulation, where some parameters need to be defined as follows: the operating power curve, the equipment connection state, the state transition probability of the equipment, and the maintenance loss of the equipment. We use a variety of function curves for the operation power of the power grid, as shown in Fig. 3 and Fig. 5. And the equipment connection of power grid adopts two grid connection modes in the case of multiple equipment. The state maintenance loss of the equipment is shown in Table 1. Moreover, unit price of grid power is 1053 Yuan/Mkw. As for transition probability, P_{12} is 0.01, P_{23} is 0.01, λ_1 is 0.005, λ_2 is 0.01 and λ_3 is 0.05. In order to show the connection of different equipment in the grid, different equipment shares the same parameters.

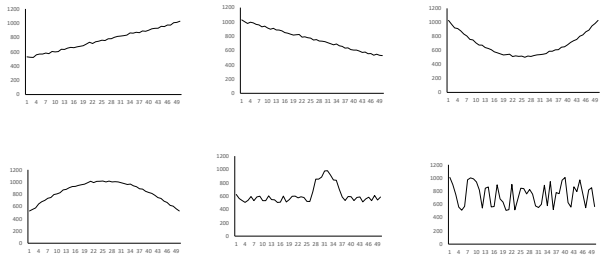


Figure 5: The dataset contains six types of grid power curves. The data simulation time interval is fifty days.

Table 1: The Maintenance Information of the Equipment

Parameter	Unit	S1	S2	S3	S4
Maintenance Cost	10000 Yuan	0	1	4	20
Maintenance Time	Day	0	1	2	4

4.2 Experiment Settings

4.2.1 Comparison method. According to the grid situation, the experimental method is divided into two parts. As for single equipment, several methods are as follows. *Template*1: repair when the equipment is in S4 state; *Template*2: repair when the equipment is in S3 and S4 state *RL_DP*: strategies obtained from reinforcement learning. For multiple equipment, some methods are as follows. *RL_DP*(*AVG*): dynamic programming method weighted by average value; *RL_DP*(*weight*): dynamic programming solution method weighted by cut set; *RL_DP*(∗) + *DD*: the dynamic programming method integrated with dynamic strategy generator.

4.2.2 Implementation details. The attenuation coefficient γ is 0.99. For testing, we simulated 10,000 episodes of data for

Table 2: Experimental Comparison Results of Single Equipment

Method	State1	State2	State3	State4	State5	State6
Template1	-238.56	-245.43	-270.17	-211.23	-193.55	-230.59
Template2	-190.91	-211.17	-214.64	-194.17	-164.52	-201.2
RL-DP	**-163.58**	**-208.35**	**-172.54**	**-191.87**	**-149.19**	**-163.51**

Table 3: Experimental Comparison Results of Multiple Equipment

Grid	Method	State1	State2	State3	State4	State5	State6
Grid 1	Template1	-694.2	-726.4	-647.73	-644.46	-590.36	-689.34
	Template2	-604.72	-638.14	-578.9	-581.83	-494.55	-601.65
	RL-DP(AVG)	-489.01	-630.35	-566.63	-563.25	-447.09	-487.87
	RL-DP(Weight)	-491.63	-624.73	-568.33	-562.22	-446.56	-490.53
	RL-DP(AVG)+DD	**-481.25**	-626.2	**-562.24**	-560.94	-447.21	-487.02
	RL-DP(Weight)+DD	-487.71	**-621.19**	-564.93	**-558.36**	**-442.04**	**-482.43**
Grid 2	Template1	-1192.29	-1219.57	-1340.88	-1069.31	-1147.45	-1156.87
	Template2	-951.03	-1063.99	-1054.06	-961.19	-1010.44	-1008.83
	RL-DP(AVG)	**-807.41**	-1046.55	-861.06	-941.38	-816.19	-816.17
	RL-DP(Weight)	-815.25	-1050.06	-855.87	-951.13	-812.85	-813.76
	RL-DP(AVG)+DD	-812.53	-1050.36	-857.29	-940.02	-811.06	**-810.43**
	RL-DP(Weight)+DD	-813.93	**-1041.73**	**-852.1**	**-937.99**	**-805.62**	-810.67

each method under each grid condition, based on the equipment transfer matrix and the corresponding overhaul strategy, and used the average value as the final experimental result. The original cut sets in Grid1 and Grid2 are $\{\{A\},\{B,C\}\}$ and $\{\{A,B\},\{A,C,D\},\{B,E\},\{E,C,D\}\}$, respectively.

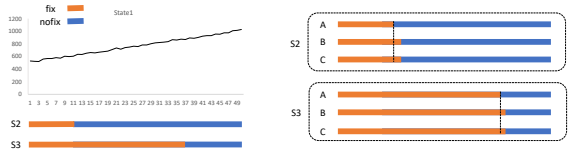
4.3 Results and Discussion

As shown in Table 2, for single equipment, compared to the policy templates, the policy obtained by reinforcement learning has a significant improvement, which is relatively stable across multiple grid power curves. For multiple equipment, as shown in Table 3, the method using a dynamic policy generator has better results compared to other methods in the vast majority of cases. At the same time, the method of solution for weighted dynamic routing outperforms the average weighted dynamic routing in general. This suggests the feasibility of equipment weight modelling approach and state communication by cut sets.

4.4 Further Analysis

In this part, we will discuss both the relationship between the learned policy and the grid power and the relationship between the policy and the grid connection. Here, we take state1, where the grid power increases linearly over time, as an example to analyze the two relationships.

4.4.1 Policy and the Grid Power. In the case of single equipment grid, the strategy learned by reinforcement learning is shown in Fig. 6 (a). In the early time of the operation of power grid, S2 and S3 states all adopt the fix strategy. With the increase of power grid operation power, the strategy in S2 state is first converted to nofix, and the strategy in S3 state is then converted to nofix. This is



(a) Policy for single equipment (b) Policy for multiple equipment

Figure 6: The policy representation for single and multiple equipment in state1.

due to the fact that reinforcement learning is carried out with the minimum loss of the whole, and the power grid operation power in the later time is larger, and the loss of equipment maintenance will also increase, so it becomes nofix. In addition, the probability from S2 to S4 is smaller than that from S3, so the strategy of S2 is changed to nofix earlier. The learned strategy is also in line with people's understanding, which can be obtained quickly by reinforcement learning instead of abundant annual experience.

4.4.2 Policy and the Grid Connection. In the case of multiple equipment, the policy learned with weighted DP solution method in grid 1, is shown in Fig. 6 (b). It can be seen that the strategy for different equipment is different. Compared with equipment B and equipment C, equipment A is more likely to adopt the strategy of nofix, because the maintenance of equipment A alone will lead to the failure loss of power grid, while the maintenance of both the equipment B and equipment C at the same time will lead to the failure loss. The strategy we learned also proves the rationality of weighted reward by cut set. In addition, the result in Table 3

proves that the dynamic strategy method can further enhance the strategy's aware of power grid state.

5 CONCLUSION

In this paper, we employ reinforcement learning to construct equipment maintenance strategy. For single equipment, dynamic programming is directly used to solve the problem. For multiple equipment, the importance of different devices is calculated using the cut set in power grid, and then applied to the dynamic programming solution. The communication between maintenance decisions is realized with the change of the cut set produced by maintenance action. We simulate the strategy through simulation data and evaluate the strategy in a variety of power grid, which verifies the effectiveness of our proposed method.

ACKNOWLEDGMENTS

This work was supported by the Science and Technology Program of the Headquarters of State Grid Corporation of China, titled as Research on Knowledge Navigation and Decision Optimization for Transformation Equipment Maintenance Based on Artificial Intelligence and Its Applications, under Grant No. 5200-201918255A-0-0-00.

REFERENCES

[1] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *CoRR* abs/1912.06680 (2019). arXiv:1912.06680 http://arxiv.org/abs/1912.06680

[2] Sushrut Bhalla, Sriram Ganapathi Subramanian, and Mark Crowley. 2020. Deep Multi Agent Reinforcement Learning for Autonomous Driving. In *Advances in Artificial Intelligence*, Cyril Goutte and Xiaodan Zhu (Eds.). Springer International Publishing, Cham, 67–78.

[3] Omar Bouhamed, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. 2019. Q-learning based Routing Scheduling For a Multi-Task Autonomous Agent. In *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE Press, 634–637. https://doi.org/10.1109/MWSCAS.2019.8885080

[4] Dongyan Chen and Kishor S Trivedi. 2005. Optimization for condition-based maintenance with semi-Markov decision process. *Reliability engineering & system safety* 90, 1 (2005), 25–29.

[5] J. Endrenyi, S. Aboresheid, R.N. Allan, G.J. Anders, S. Asgarpoor, R. Billinton, N. Chowdhury, E.N. Dialynas, M. Fipper, R.H. Fletcher, C. Grigg, J. McCalley, S. Meliopoulos, T.C. Mielnik, P. Nitu, N. Rau, N.D. Reppen, L. Salvaderi, A. Schneider, and Ch. Singh. 2001. The present status of maintenance strategies and the impact of maintenance on reliability. *IEEE Transactions on Power Systems* 16, 4 (2001), 638–646. https://doi.org/10.1109/59.962408

[6] Ding Feng, Sheng Lin, Zhengyou He, Xiaojun Sun, and Wei-Jen Lee. 2018. Optimization Method With Prediction-Based Maintenance Strategy for Traction Power Supply Equipment Based on Risk Quantification. *IEEE Transactions on Transportation Electrification* 4, 4 (2018), 961–970. https://doi.org/10.1109/TTE.2018.2863550

[7] Fang Gao, Si Chen, Mingqiang Li, and Bincheng Huang. 2019. MaCA: a Multi-agent Reinforcement Learning Platform for Collective Intelligence. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*. 108–111. https://doi.org/10.1109/ICSESS47205.2019.9040781

[8] Qi Gao, You-Wei Li, Yang Ge, and Bing Hao. 2011. Availability model of equipment based on three-levels maintenance system. In *2011 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*. 637–641. https://doi.org/10.1109/ICQR2MSE.2011.5976692

[9] Chenxi Guo, Ming Dong, Xiaoxi Yang, and Wensen Wang. 2019. A Review of On-line Condition Monitoring in Power System. In *2019 IEEE 8th International Conference on Advanced Power System Automation and Protection (APAP)*. 634–637. https://doi.org/10.1109/APAP47170.2019.9225022

[10] Yiying Li, Wei Zhou, Huaimin Wang, Bo Ding, and Kele Xu. 2019. Improving Fast Adaptation for Newcomers in Multi-Robot Reinforcement Learning System. In *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation*. 753–760. https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00162

[11] Hang Liu, Youyuan Wang, Liwei Zhou, Yufeng Chen, and Xiuming Du. 2016. An optimization method of maintenance strategy for power equipment. In *2016 International Conference on Condition Monitoring and Diagnosis (CMD)*. 940–943. https://doi.org/10.1109/CMD.2016.7757979

[12] Sergio Rozada, Dimitra Apostolopoulou, and Eduardo Alonso. 2020. Load Frequency Control: A Deep Multi-Agent Reinforcement Learning Approach. In *2020 IEEE Power Energy Society General Meeting (PESGM)*. 1–5. https://doi.org/10.1109/PESGM41954.2020.9281614

[13] John Seiffertt, Suman Sanyal, and Donald C. Wunsch. 2008. Hamilton–Jacobi–Bellman Equations and Approximate Dynamic Programming on Time Scales. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38, 4 (2008), 918–923. https://doi.org/10.1109/TSMCB.2008.923532

[14] Alvaro Serra-Gómez, Bruno Brito, Hai Zhu, Jen Jen Chung, and Javier Alonso-Mora. 2020. With whom to communicate: learning efficient communication for multi-robot collision avoidance. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, IEEE Press, Las Vegas, Nevada, 11770–11776.

[15] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (01 Nov 2019), 350–354. https://doi.org/10.1038/s41586-019-1724-z

[16] Wei Wu, Geng Haifei, and Jiang An. 2009. A Multi-agent Traffic Signal Control System Using Reinforcement Learning. In *2009 Fifth International Conference on Natural Computation*, Vol. 4. 553–557. https://doi.org/10.1109/ICNC.2009.66

[17] Di Zhou, Zizhan Wang, You Zhou, Yurong Mao, and Minqi Zhou. 2020. Research on Automatic Test Technology for Field Operation and Maintenance of Intelligent Substation. In *2020 5th Asia Conference on Power and Electrical Engineering (ACPEE)*. 11–15. https://doi.org/10.1109/ACPEE48638.2020.9136373

[18] Ming Zhou, Jun Luo, Julian Villela, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, Aurora Chongxi Huang, Ying Wen, Kimia Hassanzadeh, Daniel Graves, Dong Chen, Zhengbang Zhu, Nhat M. Nguyen, Mohamed Elsayed, Kun Shao, Sanjeevan Ahilan, Baokuan Zhang, Jiannan Wu, Zhengang Fu, Kasra Rezaee, Peyman Yadmellat, Mohsen Rohani, Nicolas Perez Nieves, Yihan Ni, Seyedershad Banijamali, Alexander Imani Cowen-Rivers, Zheng Tian, Daniel Palenicek, Haitham Bou-Ammar, Hongbo Zhang, Wulong Liu, Jianye Hao, and Jun Wang. 2020. SMARTS: Scalable Multi-Agent Reinforcement Learning Training School for Autonomous Driving. *CoRR* abs/2010.09776 (2020). arXiv:2010.09776 https://arxiv.org/abs/2010.09776