Privacy-Preserving MARL for P2P Energy Trading in Low-Voltage Communities

Yong Cui
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.
yongcui@sjtu.edu.cn

Article Info
Elaris Computing Nexus
https://elarispublications.com/journals/ecn/ecn_home.html

https://doi.org/10.65148/ECN/2026012

Received 26 May 2026
Revised from 02 July 2026
Accepted 25 July 2026
Available online 04 August 2026
Published by Elaris Publications.

Corresponding author(s):
Yong Cui, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.
Email: yongcui@sjtu.edu.cn

Abstract – In this paper, we present a privacy preserving MARL (multi-agent reinforcement learning) model for distributed P2P (peer-to-peer) energy exchange in a low voltage energy community with 55 prosumers. The model incorporates twin critic networks, multi-head attention, a counterfactual baseline, message passing and AC power-flow evaluation in a centralized-training, decentralized-execution setting. The results show that the proposed approach achieves the maximum increasing testing reward of -162.98 while maintaining the privacy of the prosumers and operational constraints of voltage and temperature. Through ablation analysis, we find that attentiveness, twin critics and counterfactual baseline improve Q-value estimation, convergence and policy stability. The framework also reduces dependence on utility, improves the decentralized trading efficiency, and does not incur an excessive computational burden, when compared to many benchmark MARL algorithms with real network constraints.

Keywords – Multi-Agent Reinforcement Learning, Peer-To-Peer Energy Trading, Energy Communities, Privacy Preservation, Energy Storage Systems, Low-Voltage Networks, AC Power Flow.

I. INTRODUCTION

Peer-to-peer (P2P) is a new energy trading model, which is based on the sharing economy concepts [1]. This includes any prosumer-to-consumer sale, whether or not there is an intermediary. This contrasts to the one-dimensional energy trading. There is a two-dimensional trade, where the movement of cash and energy is possible in a variety of ways. P2P energy trading on the other hand would enable multidirectional energy trading within a limited area. This implies that this dramatic growth of energy prosumers can enable the building of a more decentralized and accessible electrical infrastructure.

P2P energy trading has been developed to enhance economic efficiency, to stabilize grid electricity as the variable renewable energy sources are introduced, and to address the uncertainty of energy demand with the high input of sustainable energy sources. The advantages of P2P energy exchange include greater efficiency of the system, lower energy consumption and storage in P2P devices, better integration of sustainable energy sources with low-level loss in the process of energy transmission and storage, conservation of energy quality and less strain on the grid during the exchange of energy. P2P electricity markets can contribute to the decarbonization of society by implementing storage technologies and renewable energy sources, cope with inaccuracies and uncertainties in PV power forecasting, minimize losses and aid in managing the power system congestion and voltage level [2].

P2P energy trading normally brings new solutions and new daily energy transfer paradigms on the energy demand side on the power network, and the prosumers can freely decide on their energy exchange criteria, such as the amount of energy to be exchanged or the price per unit energy, for better energy efficiency or better interaction between individuals. In the P2P energy exchange system, there is no middle energy provider. People feel encouraged to actively engage in giving away their ‘extra' energy transfers with their immediate environments. The excess power can be sold at an export market rate and the increased demand for power will be encouraged by a discounted retailing rate. P2P energy offers many benefits, such as reduced power outages, increased efficiency of the power grid, increased local energy resources, the ability for local consumption of these resources, some degree of independence from energy utility providers, and the ability to select from multiple sources of energy based upon preferences of users [3].

Moreover, P2P energy trading can cater to the community's needs such as cutting electricity costs, boosting renewable energy, and sharing excess energy to those who need it, as defined by the ecosystem. Nonetheless, there is a potential issue for conventional energy providers as they may lose control of the market dynamics and price, thereby affect their earnings and possibly oppose the sharing model. **Fig. 1** displays the variations between a normal centralized energy system and the emerging decentralized network of prosumers that build demand-side P2P energy exchange networks.

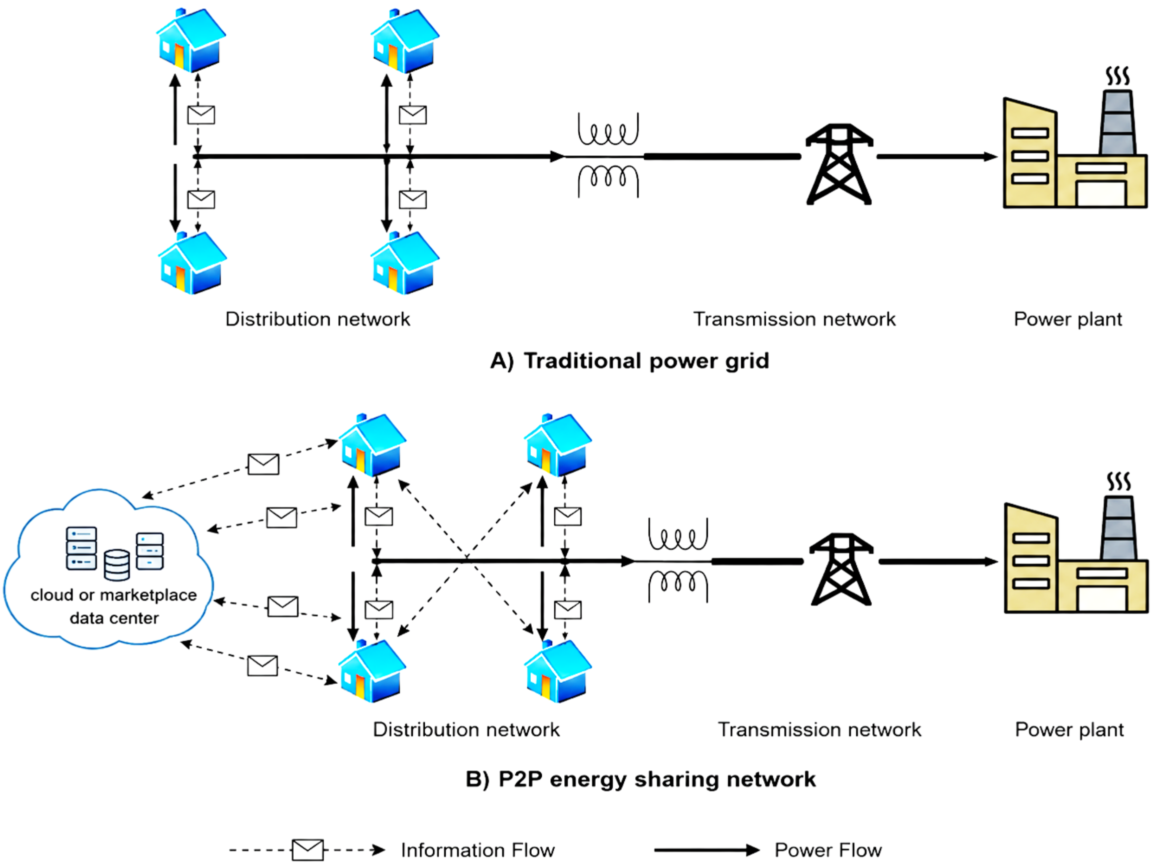

Fig 1. Electricity Trade from The Demand Side: A) Conventional Setting, B) P2P Enabled Energy Network.

There are significant problems with P2P trading. First, prosumers and consumers do not possess the advanced expertise in the continuous P2P exchange and in managing energy assets. Second, in contrast with conventional energy sources, sustainable energy sources like solar typically have zero incremental costs, so it can be hard to set competitive marketplace prices for energy transactions: a single-price auction mimicking the bulk energy marketplace could be ineffective because the market clearing prices could often be zero. Third, whereas P2P trading is only monetary, energy is transmitted by tangible distribution networks. The viability of the network within a local exchange P2P market is one of the big problems and still not solved [4].

The amount of research on MARL has drastically increased in the last few years and many more surveys exist currently. Some works focused on deep learning, combined with MARL. Zhu, Dastani and Wang [5] explored new actions: cooperation and communication learning, highlighting the interactive nature of MADRL (Multi-agent Deep Reinforcement Learning). Wang et al. [6] provide a detailed survey of MADRL and its uses in competitive, cooperative and mixed scenarios. A detailed study of MADRL and its use in different ecosystems like collaboration, competition and mixed scenarios were conducted by Hady et al. [7]. They provide a comprehensive evaluation of MADRL and its applications in different contexts such as cooperative, competitive and mixed contexts.

This was further enhanced by extensive literature evaluation done by Bao, Shi and Wang [8] including the analysis of classic techniques and providing insights on recent studies and the future scope including transfer learning and model free DRL. Liu et al. [9] categorized hurdles in MADRL and provided an overarching perspective of challenges in implementing MADRL. They also indicated more advanced approaches to addressing these issues in real life in their poll.

Additionally, Shi et al. [10] researched knowledge reuse and independence in MASs. Advanced teaching approaches may be based on previous expertise and will need a vision. Da Silva et al. [11] also gave a positive comment about how to improve the learning ability of agents through the addition of transfer training to MARL. A detailed state-of-the-art of MARL algorithms was provided by Zhang, Yang and Basar [12] who classified the different algorithms, provided challenges and answers, and listed applications and current MARL settings. Ikeda and Shibuya [13] describe the practical issues of having one central training location and multiple deployments, critical to understanding the operational characteristics of MARL models. Li et al. [14] presented a comprehensive survey of MARL; they covered the basic techniques, applications scenarios and progress of research. They also shared thoughts on the lack of practical applications of MARL and ethical issues.

Emami et al. [15] however, offer a comprehensive analysis of MARL, its various methodologies, areas of applications and research trends with a focus on limitations and ethical concerns in the implementation of the MARL. The researchers also offered a comprehensive analysis of MARL, the basic procedures, deployment scenarios, and study patterns, and pointed out the limitations and ethical issues in the practical application of MARL. As MARL continues to grow and develop in more complicated and dynamic environments, these will serve as a starting point for future research and applications.

Due to these difficulties, many researchers have investigated programming the bidding operation of agents through a decentralized model using MARL to improve the efficiency of the control of their associated solar photovoltaic and stored-energy resources. Besides that, it is also important that the architectural plans are also feasible with regards to a supply network. Kayal and Ravimaran [16] were inspired by the concept in [17] to propose a consensus-oriented actor-critic method for distributed MARL with recurrent P2P binding mechanisms, in which every agent is represented as an MDP (markov decision process). This proposed architecture facilitates knowledge sharing and agreement between agents for effective decision-making and resource assignment without having to compromise on the privacy problems and computational issues of some centralized learning approaches.

The goal of this work is to develop a privacy-protected MARL approach for P2P energy trading in low consumption energy communities. This framework aims at improving the efficiency of the trade, decreasing the dependency on the utility, coordinating the decisions of the prosumers, keeping private information private, optimizing the operation of the energy storage and ensuring that voltage and thermal limits of the network are not violated during system operation. The remainder of this study is organized as follows: In Section II, the energy-community decision problem, network constraints, goal function and multi-agent learning architecture are described. In Section III, the data, simulated environment, proposed MARL algorithm and assessment mechanism are presented. In Section IV, results on trading and network performance are presented. The main results are summarized and their implications discussed with regard to decentralized energy management in Section V.

II. PROBLEM FORMULATION

Energy Community and Decision Model

The study covers a local energy community consisting of 55 prosumer agents connected to a low-voltage distribution network. Each prosumer can use electricity, generate photovoltaic (PV) power, operate an energy storage system (ESS), participate in the capacity demand auction (CDA) market, and trade electricity with the utility. The connection between prosumers, the low voltage network operator (LVNO), the P2P market and the distribution network is shown in **Fig. 2**.

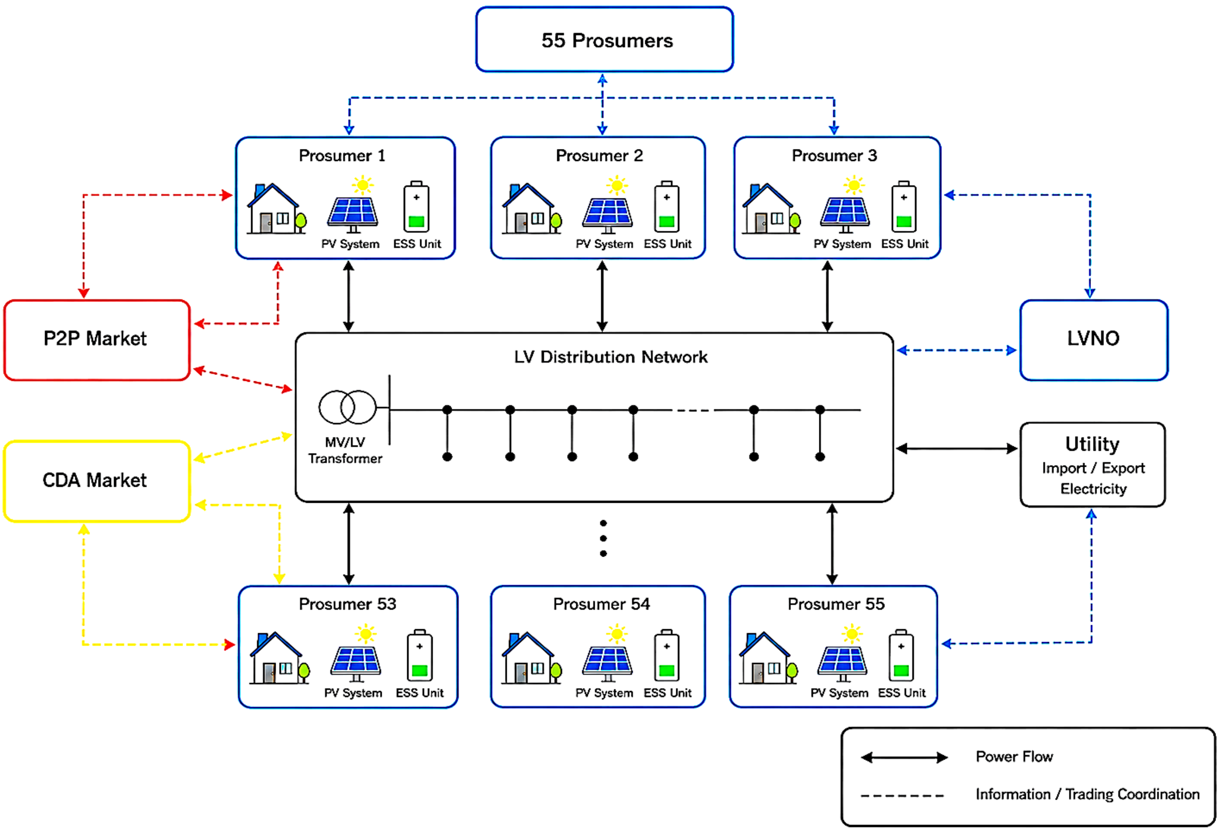

Fig 2. The Proposed P2P Energy Exchange Architecture.

The decision-making process is articulated as a POMG (partially observable Markov game). At time t, prosumer i receives a local observation including its electricity demand, PV generation, ESS state of charge, market information and current trading position. This is shown by the observation

$$o_t^i = [P_{i,t}^{\text{load}}, P_{i,t}^{\text{PV}}, S, O, C_{i,t}\lambda_t P_{i,t}^{\text{trade}}]. \tag{1}$$

Based on this insight, prosumer i chooses an action a_t^i , which can include utility trading, P2P bidding, ESS charging or discharging, and CDA-market involvement. The combined action is written as

$$a_t = [a_t^1, a_t^2, \dots, a_t^{55}, a_t^{\text{LVNO}}]. \tag{2}$$

The energy balance of prosumer i is defined as

$$P_{i,t}^{\text{grid}} + P_{i,t}^{\text{P2P}} + P_{i,t}^{\text{PV}} + P_{i,t}^{\text{ESS}} = P_{i,t}^{\text{load}}, \tag{3}$$

The ESS is partial in its charging, capacity, discharging, and state-of-charge.

Network Restrictions and Objective Function

The purpose is to exploit the estimated reduced accumulated reward:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t R_t \right] \tag{4}$$

where π is the joint policy, T is the operating horizon, and γ is the discount factor. The total reward is defined as

$$R_t = R_t^{\text{UT}} + R_t^{\text{P2P}} + R_t^{\text{LS}} + R_t^{\text{ESS}} + R_t^{\text{TD}}, \tag{5}$$

The utility-trading reward is denoted as R_t^{UT}; the P2P-trading reward as R_t^{P2P}; the load shedding reward/penalty as R_t^{LS}; ESS degradation reward/penalty as R_t^{ESS} and the thermal-discomfort penalty and network-violation penalty as R_t^{TD}. As shown in **Fig. 3**, the actions executed by the prosumer agents affect not only the energy community's economic performance, but also the physical status of the LV network. This joint action is then analyzed using the AC power flow model to determine nodal voltage magnitudes, line apparent power flows, utility exchanges, P2P transactions and other operating results.

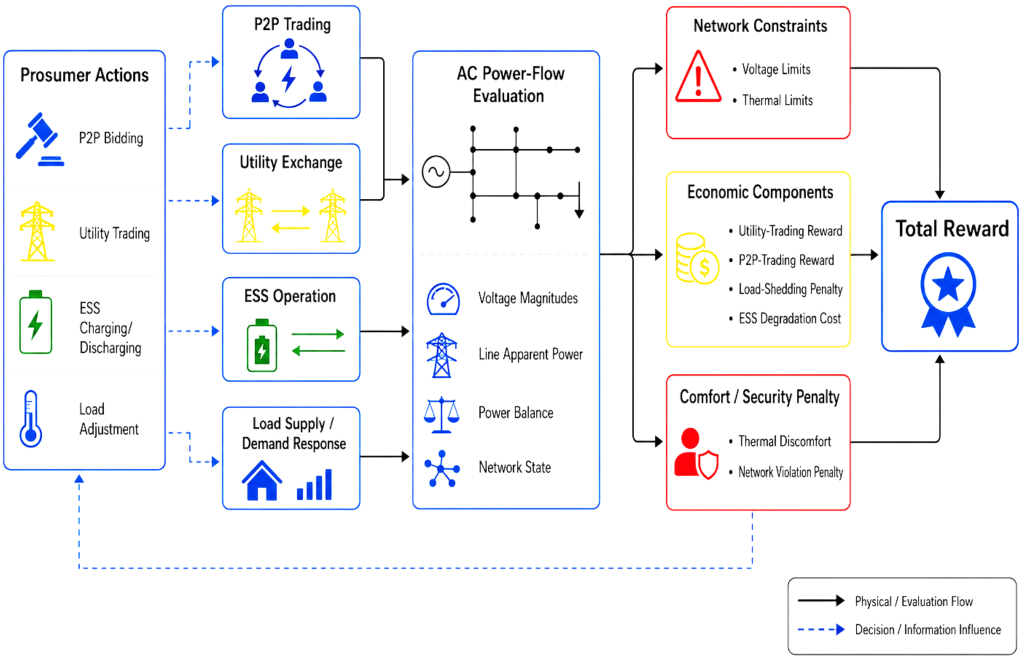

Fig 3. Economic and Network Objectives Interaction.

The main network limitations are

$$V_n^{\min} \le V_{n,t} \le V_n^{\max}, \tag{6}$$

and

$$S_{\ell,t} \le S_\ell^{\max}, \tag{7}$$

where $V_{n,t}$ is the voltage magnitude at node n, and $S_{\ell,t}$ is the apparent power flowing through line ℓ. Actions violating the acceptable voltage or thermal limitations are heavily negatively rewarded. The proposed formulation thus incorporates economic objectives and energy-balance, ESS-operation, load-satisfaction and network-security criteria.

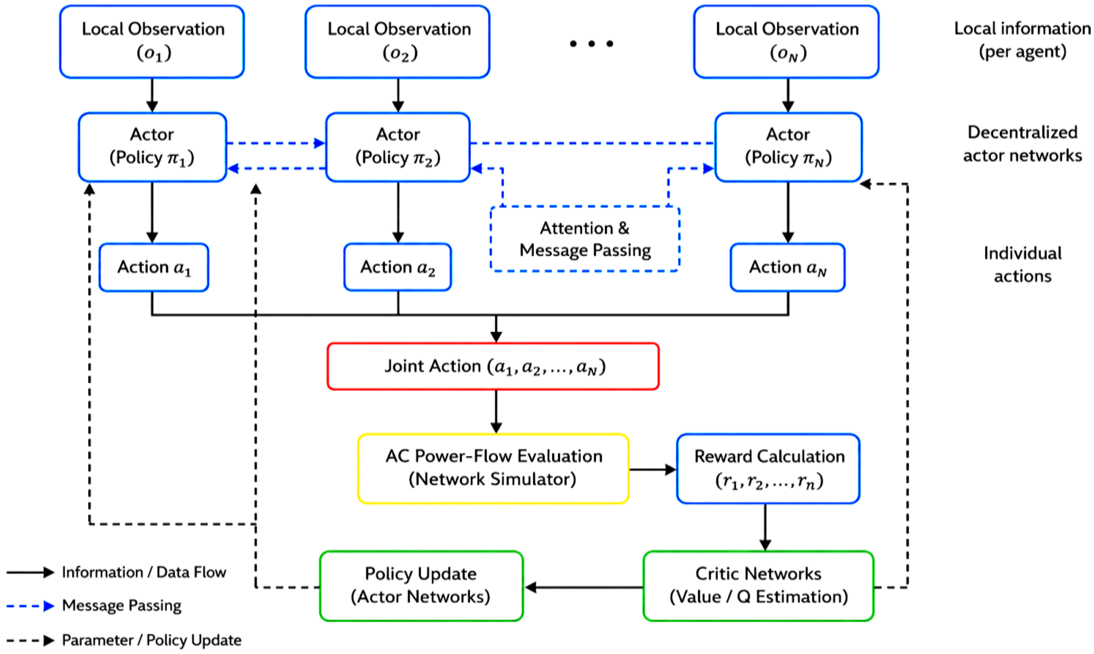

Fig 4. Decentralized Decision-Making Process in MARL.

Multi-Agent Learning Formulation

However, the global state-action space is rapidly increasing in prosumer number, making it a complex problem. Moreover, agents can only perceive fragments of the global environment and other agents' policies are subject to change during training. These features make the problem partially observable, non-stationary and make it hard to estimate individual Q-values. In this experiment, a decentralized decision-making process is used as shown in **Fig. 4**. The decision of each prosumer is made using observations from the local agent and the privacy-preserving message transfer provides encoded information for coordination among the agents without revealing all of agents' private states. The individual global maximum condition includes the individual agent-level contribution:

$$A_{\text{tot}}(s_t, a_t) = \sum_{i=1}^{55} A_i\left(o_t^i, a_t^i\right) + A_{\text{LVNO}}\left(o_t^{\text{LVNO}}, a_t^{\text{LVNO}}\right). \tag{8}$$

This factorization does not require an explicit listing of all combinations of prosumer and LVNO state-action pairings. In this way it lowers the combinatorial complexity and allows for a more precise estimation of individual Q-values. The proposed formulation is based on decentralized rules that aim to minimize dependence on utility, maximize P2P energy trading, minimize load shedding, minimize degradation of ESS, maximize prosumer privacy, and keep the LV network within acceptable operating parameters.

III. DATA AND METHODS

Data and Simulation Environment

The proposed methodology is applied to a local energy community with 55 prosumer agents (agents) that is part of a low voltage distribution network. The simulation environment includes time-varying values of electricity demand, PV generation and energy storage system (ESS) conditions, electricity prices, P2P transaction information and network operating parameters. The actual simulation setup should give the following information: actual source of the data set, sampling interval, number of running days, geographical setting, and the training–testing split.

Each prosumer receives a local observation in each operation interval, providing their electricity demand, PV generation, ESS state of charge, and market information as well as their trading position at the present time. Based on this information, the agent decides on actions relating to the utility exchange, P2P trading, charging/discharging of ESS and joining the capacity demand auction (CDA) market. The proposed joint-action is investigated based on an AC power-flow model for evaluation of the nodal voltage magnitudes, line apparent-power flows, power exchanges and network violations.

Table 1 provides an overview of the training set-up. The previous experiences are stored in the replay buffer and the mini-batches are sampled to update actor and critic networks. The discount factor is .95. There are 4 attention heads in the attention layer. The size of replay-buffer is 1*106 transitions and mini-batch size is 512. The exploration and policy noise have a standard deviation of 0.1 and 0.2, respectively.

Table 1. Training set-up

Parameter	Value
Number of prosumer agents	55
Number of attention heads	4
Discount factor	0.95
Replay-buffer capacity	1×10^6 transitions
Mini-batch size	512
Exploration-noise standard deviation	0.10
Policy-noise standard deviation	0.20
Training arrangement	Centralized training, decentralized execution
Feasibility model	AC power flow

Proposed MARL Method and Evaluation

The approach that we provide is trained in a centralized manner and then implemented locally. During training, the critic networks can approximate the value of the collaborative tasks from the input of the global or aggregated. In execution, each prosumer independently executes their action according to their local observations, which also decreases the necessity of information sharing, and thus privacy protection. We propose a framework, which combines the multi-head attention mechanism, twin critic networks, counterfactual baseline, and message-passing mechanisms. The attention mechanism aims to locate relevant interactions between the prosumers and the LVNO and the two critic networks give two estimations of the value to avoid overestimation bias. The counterfactual baseline is used to measure the individual contribution of each agent's action, and to stabilize the policy-gradient update. Message forwarding is a way to coordinate without revealing the totality of agents' private state.

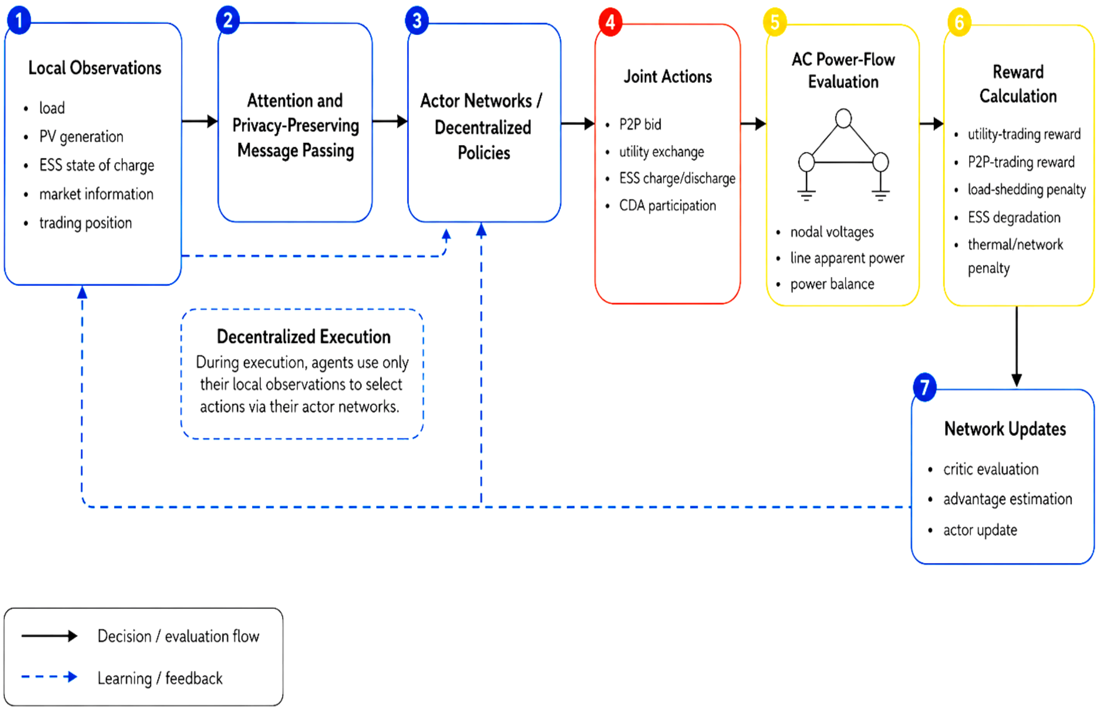

Fig 5. Training and Decentralized-Execution Procedure of Suggested MARL Technique.

The training and implementation steps are summarized in **Fig. 5**, from the local observations to the action selection, the evaluation of the AC power-flow, the computation of the reward and the actor–critic update. The energy-community, market and network conditions are the same for all competing algorithms, including Fed-MARL, MF-MARL, FACMAC, DoP, MAAC, MATD3, and MADDPG. Where possible their hyperparameters are the same or similar. The evaluation factors include cumulative reward, privacy preservation, network feasibility, adjustable parameters, execution time and training time.

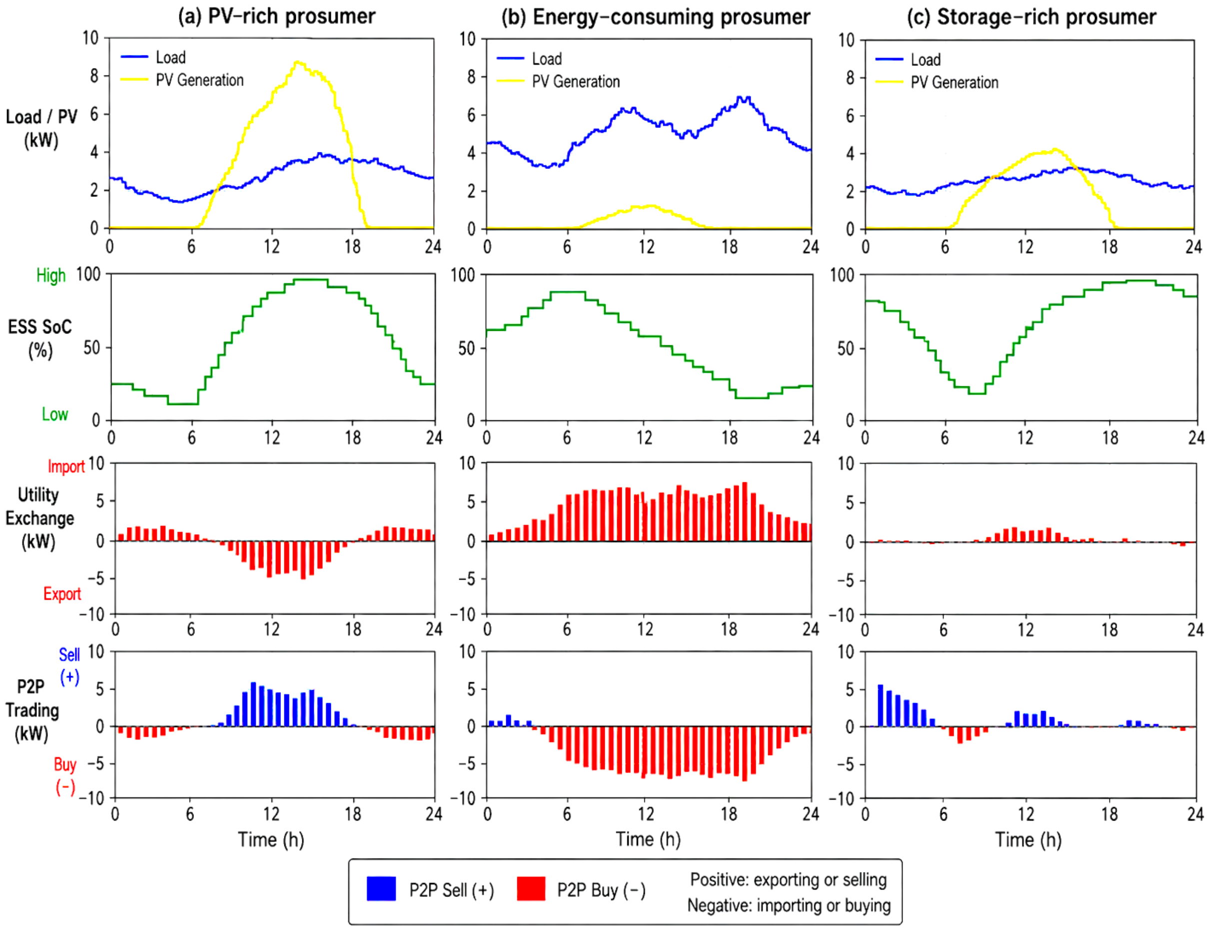

Fig 6. Representative Decentralized Energy-Trading Policies in P2P Networks.

Table 2. Evaluation Design

Evaluation element	Specification
Size of energy community	55 prosumers
Sample prosumer profiles	3
Ablation elements	Attention, twin critics, counterfactual baseline
Benchmark algorithms	7 competing MARL methods
Primary performance measure	Cumulative reward
Privacy measure	Preservation of private prosumer information
Computational measures	Parameters, execution time, training time
Network measures	Voltage magnitudes and line apparent power
Physical validation method	AC power-flow analysis

Ablation studies are performed with regard to the multiple-head focus mechanism, dual-critic architecture and hypothetical baseline. Three typical prosumers are also selected to illustrate the decentralized policies taught: PV-rich prosumer, energy-consuming prosumer who buys energy and charges its ESS, and storage-rich prosumer who sells the stored energy before sunrise. These working patterns are shown in **Fig. 6**. Evaluation design is summarized in **Table 2**.

IV. RESULTS

From our analysis, the value of the attenuation head mechanism is 4.01 and the discounting coefficient is 0.939. The reiteration batch size and buffer size were 512 and 1e6, correspondingly. The SD (standard deviation) of the stochastic perturbation ε_i is 0.12 and of the policy perturbation noise $\check{\varepsilon}_i$ is 0.21. The model tuning parameters for the different MARL approaches are equal to the ones used in the proposed technique. The total reward system developed over 5 iterations in the learning phase is illustrated in **Fig. 7**. The learning procedure starts with a small reward in the beginning (from an inefficient policy) and gradually the cumulative returns increase, showing how a high reward policy is built up step-by-step.

Discrete Q-value estimations are made possible and the combinatorial computational complexity is eliminated by the global maximum, which is introduced at the individual level for the proposed method in the decomposition of the global advantage estimation functions between LVNO and prosumer agents. Our technique is different from MAAC, MATD3 and

MADDPG, which always join a full collection of state-action pairings, and removes low Q-value correlations and larger input spaces.

Although DoP and FACMAC calculate weighted sums according to other parameters, our factorization does not introduce any unnecessary complexity [18]. Remarkably, the whole reward function includes reward for load-shedding and AC energy flow, resulting in a large non-convex optimization formulation. Saddle spots and lack of results in current techniques like FACMAC, DoP, MAAC, MATD3 and MADDPG. Our factoring solves optimization complexities and enhances results. In the proposed approach, the non-stationarity issues in learning are tackled by the information-exchange mechanisms. In particular, Fed-MARL replicates the same settings across all prosumer agents, regardless of the variety of reward functions and state transitions. MF-MARL relies on mean properties of other agents, and does not take account of the important properties of the scenario that are necessary to approximate the unique Q-value well. Experiments show that our approach maintains privacy and achieves higher growing rewards and steadier union compared to competitors.

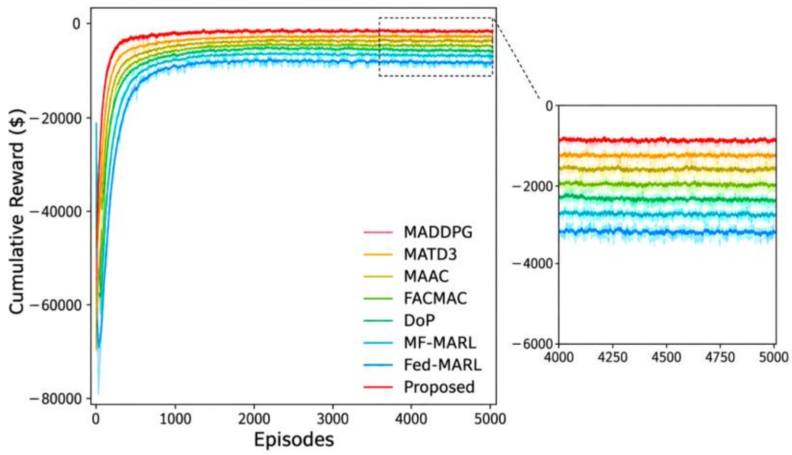

Fig 7. Cumulative Reward Evolution.

Ablation experiments are performed to make further evaluations of the effectiveness of multiple-head attention architecture, counterfactual baselines and twin-critic networks in the proposed framework as shown in **Fig. 8**. According to the simulation results, the multi-head attention architecture, counterfactual baseline network, and twin critic network, are beneficial to the entire performance and are more successful in estimating the Q-function. In addition, the counterfactual reference baseline boosts the policy optimization gradient, which results in more efficient and stable training.

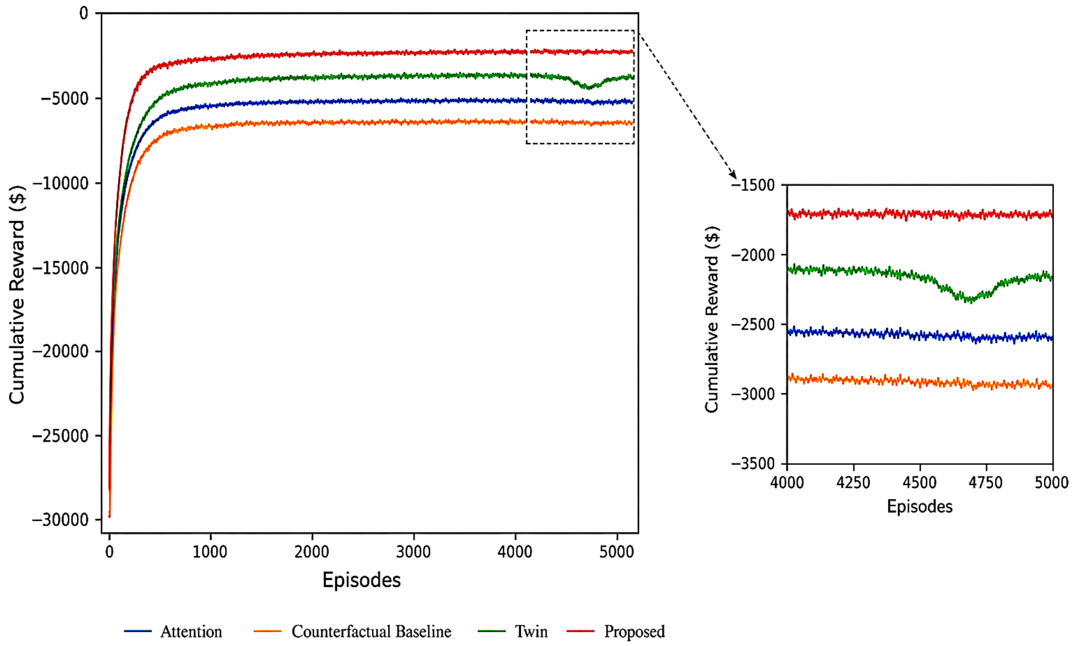

Fig 8. Simulation Results.

Table 3 and **Table 4** quantify the counterfactual baseline effect on reward enhancement in comparison to the utility-only trading scenario. The findings show that this reference estimate boosts the P2P energy exchange and mitigates the "inactive agent" problem in the MARL architecture.

Table 3. Reward Lift With/Without Counterfactual Reference Estimate

Sole reward gains compared to exchange only with utility	Mean (%)	Maximal (%)	Minimal (‰)
Without reference estimate	2.82	50.74	-85.7
Proposed	6.40	124.99	0.04

Table 4. Accumulated Reward on Testing Days

Algorithms	Privacy Preservation	Cumulative Reward ($)
Fed-MARL	✓	-249.81
MF-MARL	✓	-185.89
FACMAC	×	-240.51
DoP	×	-245.67
MAAC	×	-190.16
MATD3	×	-173.93
MADDPG	×	-207.27
Proposed	✓	-162.98

Table 5 shows the computational efficiency and tunable model variable of the framework for each agent. Please note that a negative signal means that there is a positive P2P energy exchange cost. It is observed that with an improved reward signal (low cost) it implies that an optimal P2P energy exchange scheme is investigated in the policy optimization phase. These competing methods are more efficient in the training process, but each agent still requires milliseconds to process. Since training takes place offline, it doesn't impact the overall assessment of computing efficiency significantly when compared to other competitors that train for a longer period.

Additionally, MAAC has a smaller number of tunable model variable compared to MATD3 and MADDPG, but its attention approach necessitates every prosumer to undertake computations iteratively [19]. This increases the computation burden of MAAC as it is time consuming. We propose to include attention processes; however, they are limited to a pre-learned proxy simulation inside the LVMO decision module. In this way, the computing attention load is taken out of MARL, allowing the suggested methodology to achieve a meaningfully reduced computing time to that of MAAC.

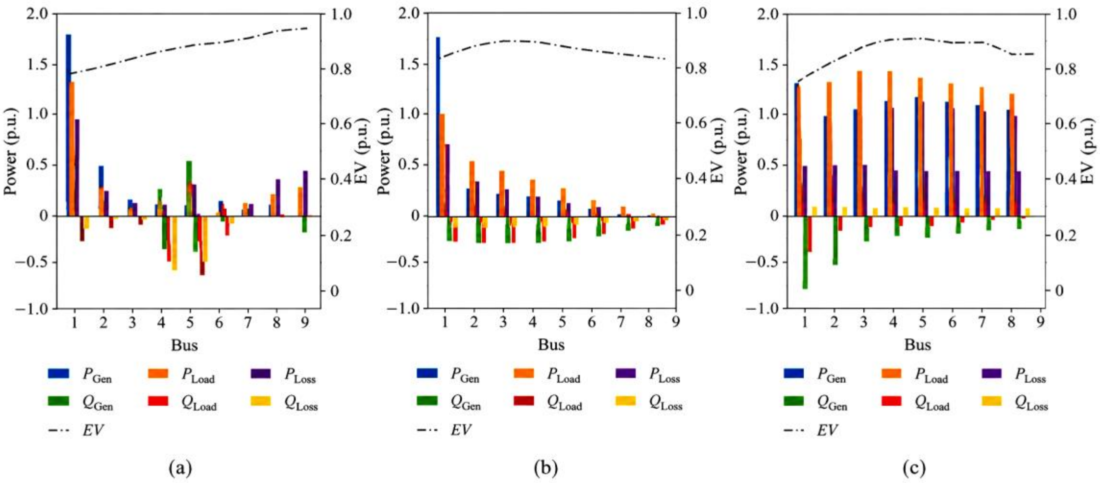

Fig 9. P2P Energy Market Simulation Output.

Table 5. Performance Analysis

MARL Algorithm	Configurable Parameters	Execution Time (ms)	Training Time (h)
Proposed	3736	12.44	3.05
Fed-MARL	1738	12.43	1.33
MF-MARL	2135	12.35	2.75
FACMAC	2779	12.45	1.65
DoP	1839	12.42	1.35
MAAC	3992	12.31	4.41
MATD3	40728	12.38	3.95
MADDPG	20791	12.33	3.25

The P2P energy exchange plan is implemented in a distributed manner according to local measurements. Since the energy ecosystem integrates 55 prosumer-based agents, we chose three archetypal decision modules to demonstrate rules trained by the MARL approach. The model findings of these modules are visualized in **Fig. 9. Fig. 9(a)** depicts a prosumer group with sufficient PV generation to cover localized demands at midday but requiring trading on the CDA marketplace or/and with the utility. **Fig. 9(b)** visualizes a cluster of energy users who are purchasers from the CDA marketplace to power the ESS (energy storage system) and address the local energy demand.

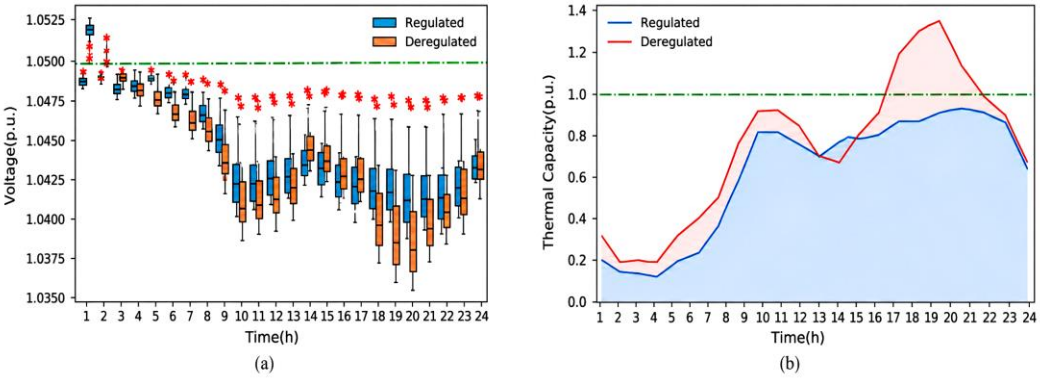

Fig 10. The Framework of Deregulated Vs. Regulated.

Before the morning, there are a number of prosumers with excess energy in the ESS who want to participate in the energy marketplace CDA as energy sellers, as shown in **Fig. 9(c).** The PVs will top up the energy store around midday, ensuring that the energy stored stays within limits. It is important to emphasize that the three prosumers' types are not fixed: the roles of these prosumers continuously vary depending on the variations of the power consumption and the photovoltaic generation characteristics and on the state of charge. The node voltage scale and the peak apparent energy distribution on a full day are shown in **Fig. 10** for the proposed scheme and the uncontrolled scheme. The results of the simulation indicate that the agents who have a considerable ESS will participate in the early morning capacity demand auction (CDA) marketplace without taking into account the voltage solidity in the de-controlled environment. These unscheduled transactions upset the system viability and violate the thermal constraint at high-load periods.

The proposed scheme, on the other hand, synergically enhances the practicability of the P2P energy exchanges and LV network, which ensures that the voltage level of every node and the thermal limit of each line are within safe operating ranges. It is worth pointing out that the framework is a soft-inclined one, and does not give a strict computational agreement of zero desecrations in all potential situations. However, a substantial cut in total compensation will be deducted if the activity is found to be in breach of the operating limits. Next, the actor is coached to avoid these desecrations in turn and the combined policy sensibly operates within the bounds of acceptability. To identify the source of efficiency optimization, the whole reward decomposition is given in **Table 6**, with positive numbers being revenues to the prosumers and negative numbers being the costs paid. A significant fact to mention is that the benefits of P2P energy exchange are being offered from various angles to suppliers and customers.

Table 6. Decomposing the Total Compensation on Testing Days

Reward Component	Value
Utility Trading Reward	-70.07
P2P Trading Reward	20.87/-20.87
Load-Shedding Reward	-6.36
ESS Degradation Reward	-64.49
Thermal Discomfort Reward	-27.40

The P2P economy is a system in which the total income of the sellers is the total expenditure of the buyers and payments are transferred between prosumers. Therefore, the overall impact of P2P trading on the overall reward is null, which means that P2P exchange is a transfer imbursement approach, and it does not create any additional benefit to the society [20]. This also shows that the suggested MARL technique does not incur load shedding as efficiency of the operation. This improvement is primarily attributed to the adoption of a more efficient P2P energy exchange framework that decreases the use of electricity from the energy grid while ensuring the system's operating restrictions.

V. CONCLUSION

Results demonstrate that decentralized energy resource trading can be successfully handled while ensuring prosumers' privacy and distribution network security. The proposed solution offers a more balanced approach to incentivize agents to take economically efficient measures and adjust to the physical network conditions, and to meet the community level requirements. The energy being taught also demonstrates flexibility with energy storage and solar energy generation, and local energy exchange to minimize unnecessary reliance on the utility. Important, network conditions can be considered in the decision-making process, so that behaviors that are economically attractive can be prevented, resulting in suboptimal voltage or heat conditions. However, due to the soft restrictions in the framework, it is not possible to ensure the elimination of all breaches in all work contexts. Larger energy communities, changing market dynamics, uncertainty and practicality should be considered for future research.

CRediT Author Statement
The author reviewed the results and approved the final version of the manuscript.

Data Availability
The datasets generated during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interests
The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding
No funding was received for conducting this research.

Competing Interests
The authors declare no competing interests.

References
[1]. E. A. Soto, L. B. Bosman, E. Wollega, and W. D. Leon-Salas, "Peer-to-peer energy trading: A review of the literature," Applied Energy, vol. 283, p. 116268, Dec. 2020, doi: 10.1016/j.apenergy.2020.116268.
[2]. J. Liu, H. Yang, and Y. Zhou, "Peer-to-peer energy trading of net-zero energy communities with renewable energy systems integrating hydrogen vehicle storage," Applied Energy, vol. 298, p. 117206, Jun. 2021, doi: 10.1016/j.apenergy.2021.117206.
[3]. A. Shrestha et al., "Peer-to-Peer energy trading in Micro/Mini-Grids for Local Energy Communities: A Review and case Study of NePAl," IEEE Access, vol. 7, pp. 131911–131928, Jan. 2019, doi: 10.1109/access.2019.2940751.
[4]. L. Ali et al., "Application of a Community Battery-Integrated microgrid in a Blockchain-Based local energy market accommodating P2P trading," IEEE Access, vol. 11, pp. 29635–29649, Jan. 2023, doi: 10.1109/access.2023.3260253.
[5]. C. Zhu, M. Dastani, and S. Wang, "A survey of multi-agent deep reinforcement learning with communication," Autonomous Agents and Multi-Agent Systems, vol. 38, no. 1, Jan. 2024, doi: 10.1007/s10458-023-09633-6.
[6]. E. Wang et al., "A MADRL-Based credit allocation approach for interactive Multi-Agents," International Journal of Information Technology & Decision Making, vol. 24, no. 07, pp. 2117–2137, Feb. 2025, doi: 10.1142/s0219622025500312.
[7]. M. A. Hady, S. Hu, M. Pratama, Z. Cao, and R. Kowalczyk, "Multi-agent reinforcement learning for resources allocation optimization: a survey," Artificial Intelligence Review, vol. 58, no. 11, Aug. 2025, doi: 10.1007/s10462-025-11340-5.
[8]. K. Bao, T. Shi, and S. Wang, "Engineering Application Progress of Multi-Agent Deep Reinforcement Learning," 2025 IEEE 19th International Conference on Control & Automation (ICCA), pp. 791–796, Jun. 2025, doi: 10.1109/icca65672.2025.11129859.
[9]. D. Liu, F. Ren, J. Yan, G. Su, W. Gu, and S. Kato, "Scaling up Multi-Agent Reinforcement Learning: An extensive survey on scalability issues," IEEE Access, vol. 12, pp. 94610–94631, Jan. 2024, doi: 10.1109/access.2024.3410318.
[10]. D. Shi, J. Tong, Y. Liu, and W. Fan, "Knowledge reuse of Multi-Agent reinforcement learning in cooperative tasks," Entropy, vol. 24, no. 4, p. 470, Mar. 2022, doi: 10.3390/e24040470.
[11]. F. L. Da Silva, G. Warnell, A. H. R. Costa, and P. Stone, "Agents teaching agents: a survey on inter-agent transfer learning," Autonomous Agents and Multi-Agent Systems, vol. 34, no. 1, Dec. 2019, doi: 10.1007/s10458-019-09430-0.
[12]. K. Zhang, Z. Yang, and T. Başar, "Multi-Agent Reinforcement Learning: A selective overview of theories and algorithms," in Studies in systems, decision and control, 2021, pp. 321–384. doi: 10.1007/978-3-030-60990-0_12.
[13]. T. Ikeda and T. Shibuya, "Centralized Training with Decentralized Execution Reinforcement Learning for Cooperative Multi-agent Systems with Communication Delay," 2022 61st Annual Conference of the Society of Instrument and Control Engineers (SICE), pp. 135–140, Sep. 2022, doi: 10.23919/sice56594.2022.9905866.
[14]. T. Li et al., "Applications of Multi-Agent Reinforcement Learning in Future Internet: A Comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 24, no. 2, pp. 1240–1279, Jan. 2022, doi: 10.1109/comst.2022.3160697.
[15]. Y. Emami et al., "Human-In-The-Loop Machine Learning for safe and ethical autonomous Vehicles: principles, challenges, and opportunities," arXiv (Cornell University), Aug. 2024, doi: 10.48550/arxiv.2408.12548.
[16]. A. Kayal and S. Ravimaran, "Experimental Evaluation of Coordinated Decision Making Strategies Based on Multi-Agent Reinforcement Learning," 2026 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1–9, Jan. 2026, doi: 10.1109/icses66558.2026.11478990.
[17]. V. Markova, V. Shopov, and M. Dimitrova, "Multi-Agent Reinforcement Learning for Pedagogical Interaction with Emotion-Informed Behaviour," 2026 22nd International Conference on Information Technology Based Higher Education and Training (ITHET), pp. 1–9, May 2026, doi: 10.1109/ithet69978.2026.11585150.
[18]. L. Cai, J. Chen, R. G. Downey, and M. R. Fellows, "On the parameterized complexity of short computation and factorization," Archive for Mathematical Logic, vol. 36, no. 4–5, pp. 321–337, Aug. 1997, doi: 10.1007/s001530050069.
[19]. C. Lou, Z. Jin, W. Tang, G. Geng, J. Yang, and L. Zhang, "LLM-Enhanced Multi-Agent Reinforcement Learning with Expert Workflow for Real-Time P2P Energy Trading," IEEE Transactions on Smart Grid, p. 1, Jan. 2026, doi: 10.1109/tsg.2026.3684885.
[20]. T. Perger, L. Wachter, A. Fleischhacker, and H. Auer, "PV sharing in local communities: Peer-to-peer trading under consideration of the prosumers' willingness-to-pay," Sustainable Cities and Society, vol. 66, p. 102634, Dec. 2020, doi: 10.1016/j.scs.2020.102634.

ISSN (Online): 3105-9082