Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 2, 2026

MULTI-AGENT REINFORCEMENT LEARNING FOR COORDINATED DEMAND RESPONSE IN SIMULATED MICROGRIDS

Roshan Ali Siyal[1], Abdul Quddus Bhutto[2], Aadil Jamali[*3], Imtiaz Ali Brohi[4], Altaf Hussain Bouk[5], Najma Imtiaz Ali[*6]

[1]Civil & Electrical Projects Contracting Co. Ltd. (CEPCO), Jeddah
[2]Department of Plant & Machinery Water Maintenance, MARAFIQ
[*3]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Sindh, Pakistan
[4, *6]Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[5]Department of Computer Science & Engineering, Yanbu Industrial College, Yanbu, Saudi Arabia

[*1]roshan.ali@cepco-sa.com, [2]bhuttoaq@marafiq.com.sa, [3]aadil.jamali@usindh.edu.pk, [4]drimtiazbrohi@gmail.com, [5]bouka@rcjy.edu.sa, [*6]najma@utem.edu.my

DOI: https://doi.org/10.5281/zenodo.18721871

Keywords
Multi-Agent Reinforcement Learning; Centralized-Training Decentralized-Execution (CTDE); Demand Response; Microgrids; Safety-Aware Reinforcement Learning; Communication-Efficient Coordination; Constrained Markov Decision Processes (CMDP); Scalability and Robustness; PyPSA-based Simulation; Renewable Integration; Fairness in Demand Response; Privacy-Preserving Learning

Article History
Received: 21 December 2025
Accepted: 05 February 2026
Published: 21 February 2026

Corresponding Authors: *
Najma Imtiaz Ali
Aadil Jamali

Abstract
The accelerated growth of the distributed energy resources, as well as the associated growth of prosumer-based variability, require new control paradigms, that enable balancing economic goals, operational safety and communicative limits in distribution level. The paper presents a centralized-training-decentralized-execution multi-agent-reinforcement training platform, which combines safety conscious learning (lagrangREQ), i.e. Lagrangian-constrained critics and run time learnability filter) with communication efficient coordination (budget-aware scheduling and message compression). Testing is done to a reproducible microgrid simulator however on power-flow primitives and real-life renewable and price traces that covers across various situations in which the renewable penetration is varied as well as the population of agents (10-100 agents), forecast error, packet loss and contingency events. Empirical findings indicate that the developed approach achieves a normalized operational cost of 0.71 over 0.78 using a centralized MPC oracle, 0.79 using a leading MARL oracle (MAPPO) and 0.95 using independent agents and achieves a peak reduction of 31.6 percent and shifts more than 50 kWh/day of flexible demand in test scenarios. The results are significantly more safety performance, where the composite-voltage and SOC violations have been significantly decreased to almost single-digit value per thousand steps with statistical significance of cost and peak gains verified by paired tests and bootstrap confidence intervals at the 95 percent level. The method is gracefully degraded when forecast errors are moderate, and when it can tolerate up to approximates of 20 percent communication loss and ablation studies can isolate the efforts of safety critics and communication modules. The findings show that safety-enhanced, socially-aware CTDE MARL represents an experimental approach to coordinated demand response that is practical, scalable, and capable of ensuring reliability, and open artifacts are released to facilitate the speed at which fields may be

translated.

1. INTRODUCTION

New grid-control paradigms are being put under a pressure that has never occurred before by the ever-growing proliferation of the distributed energy resource (DER) (sorufenib) such as rooftop photovoltaics, behind-the-meter storage, electric chamberlain vehicles and responsive loads, which are influencing the topology and the operational needs of distribution networks [1]. The growing infiltration of intermittent renewable generation and the resulting growth in prosumer behaviour necessitates demand-side resources which are coordinated on a fine-grained, adaptive basis to maintain reliability and to take advantage of the flexibility in the economic condition; the traditional time-of-use tariffs and the fixed-rule approach to such highly heterogeneous, stochastic environments are not responsive or adaptive enough [2]. Large scale optimal power-flow solvers and centralized control techniques, despite their theoretical strengths, have practical barriers in distribution, where the measurement coverage is small, where the communication is limited, privacy matters are an issue, and where the computational latency scale blocks real time action on a large scale [3].

Although significant efforts have been previously conducted on demand response and microgrid control, there are still significant gaps in those approaches, which can meet operational safety, communication sparseness, and decentralized execution at scale. A good part of the literature either focuses on model-based optimal control with good observability and centralized computation, or data-driven control policies are based on simple decentralized heuristics that do not involve rigorous safety guarantees; relatively few methods have formalized network-level safety constraints, such as voltage and frequency limits or state-of-charge (SOC) margins on storage devices, into data-driven control policies [4]. There are additional communication constraints and partial observability in practical distribution systems, which make coordination more difficult: limited bandwidth, intermittent connectivity, and reduced information disclosure necessitate nontrivial trade-offs between the accuracy of information shared in a state and the effectiveness of coordination [5]. Furthermore, the applicability of most learning-based demonstrations is limited to small numbers of agents, or to completely centralized processes of training, which are not efficient in most cases to large numbers of heterogeneous agents, leaving questions of scalability, equitable allocation of benefits, and resilience to realistic contingencies unsolved [6].

The centralized-training, decentralized-execution (CTDE) multi-agent reinforcement learning (MARL) paradigm is a potentially viable avenue to balance these conflicting needs, in which rich, centralized, and coordination strategies can be learned, and low-latency and locally executable policies executed at run-time. Every policy will depend on storage capabilities and elsewhere distributed among received messages on a shared critics or joint-value models in the CTDE paradigm to generate coordinated policies, but still have actors executing based on local findings and by sending messages that are both lightweight to make them privacy-conscious and latency-conscious [7]. Safety requirements may be added through reward shaping, constrained optimization (e.g. using Lagrangian methods or safety critics) and run-time feasibility filters, which enforce hard operational constraints; here, policies learned can comply with grid stability margins and yet remain free to use demand-side flexibility to make an economic gain [8]. A solution to communication-awareness is explicit budget-conscious message schedule, and compression and selective neighbor to neighbor signalling that results in policies graduating gracefully to reduced bandwidth or reliability conditions.

The contributions of the current work are triple and aimed to resolve the operational, methodological, and reproducibility limitations mentioned above. Following the initial step, a CTDE MARL system is designed that incorporates safety-aware elements of learning and an efficient message costing and communication layer that is specifically designed to comply with heterogeneous microgrid agents, which results in algorithmic novelty that prioritizes operational safety, economic goals, and cost-of-communication considerations. Second, an extended experimental benchmark library is presented, based on a realistic Gym-like microgrid simulator with renewable and price traces derived from history, which in turn can be used to make quantitative comparisons across a range of conditions that differ in renewable penetration, agent heterogeneity, communication loss properties, and contingency events and offer statistically significant metrics of cost, peak reduction, safety violations, scalability and fairness. Third, any software artifacts, scenario configuration, and trained weight of policies will be made public in support of reproducible research and to speed-up follow-on activity by practitioners and researchers, thus resulting in transparent evaluation and future potential transfer to field deployments [9], [10], [11].

2. Related Work

The economic history of demand response and microgrid control Demand response and microgrid control The model-based optimization and control-theoretic traditions have a long history of finding variants of optimal power flow, unit commitment and model predictive control (MPC) systems to coordinate the distributed resources, within explicit constraints and forecasted disturbances [12]. Their practical implementation in the distribution network is often constrained by incomplete observability, parametric uncertainty, and the computational cost of solving large-scale optimizations in real time and their susceptibility to heterogeneity at the device level or stochastic consumer behaviour when it is not modeled in the same way as the models; moreover, such approaches rely on accurate system models, making them less robust in the face of heterogeneity at the device level or stochastic consumer behaviour that does not follow the model assumptions [13]. Less complex decentralized heuristics and rule based schemes are possibly interesting due to their ease of operation and the certainty of implementation under some regimes, but tend to lose economic optimality and responsiveness to highly changing renewable production and interactions between heterogeneous loads and storage resources [14].

The recent efforts which investigate multi-agent reinforcement learning (MARL) in energy systems show that learning-based coordination can adapt to nonstationary conditions and can capitalize upon complements among distributed assets, yet appear to suffer a number of significant limitations. Most of the MARL works focus on small-scale proof-of-concept experiments or make the assumption of liberal communication and observability in training and execution that is unrealistic in most distribution scenarios [15]. It has been found that centralized critics can enhance coordination in training by a significant margin without sacrificing decentralized execution and published analysis frequently does not rigorously address operational safety constraints, communication budgets and fairness measures which are essential in practice [16]. Also, heterogeneity of agents, which takes the form of differences in action space, flexibility in time, and dynamics at the device level, is an issue to the conventional MARL algorithms, which can slow down to convergence or end up having policies that are unfairly advantageous to a group of agents unless explicit design decisions are made to exploit these asymmetries [17].

The topics of safety in reinforcement learning and communication-aware MARL are current research, and techniques are directly applicable to energy applications, but cross-fertilisation between these fields and the practice of a power system is immature. Embodied models Constrained RL models, including Lagrangian relaxation and dual-ascent methods, introduce tools to use operational constraints in the learning scenario (through hard or soft safety) and have been utilized in robotic and control tasks to generate policy with probabilistic or guaranteed safety Constrained RL models include Lagrangian relaxation and dual-ascent versions [18]. Constrained optimization can be enhanced with runtime safety filters, shielded learning architectures, safety critics transforming programs into valid executions or shaped exploration ofavoiding catastrophic failures, especially when operations that are performed to clean, stabilize the grid or fill the battery may result in centrifugal failures [19]. Message compression, selective information sharing and event-driven communication protocols have also been studied by communication-constrained MARL studies to lower bandwidth consumption and maintain coordination performance but most experiments have been conducted in abstract benchmarks instead of in power-system simulators that model electrical constraints and transient behavior [20].

The ability to evaluate learned control strategies in a manner credible requires reproducible and realistic simulations environment and the tools collection PyPSA and open repositories of renewable and load traces is the base backbone to such work. PyPSA allows detailed steady-state and quasi-static time-series power flow simulation, common test feeders and grid topologies, and can be easily connected to external RL environments to generate Gym-friendly interfaces to develop algorithms [21]. Public solar, wind and price time series synthesis data and public datasets of historical variability and extreme events of national laboratories and system operators provide the historical variability and extreme events needed to stress-test controllers in realistic operating conditions to allow more significant evaluations of robustness, safety and economic performance [22].

3. SYSTEM MODEL & PROBLEM FORMULATION

It represents the microgrid as a distribution-level electrical network comprising of buses connected by distribution lines with each bus possibly containing distributed energy resources, behind-the-meter-based storage, flexible loads and measurement points. Intermittent generation sources such as distributed photovoltaic arrays, and other time-varying sources are modeled as time-varying injection profiles at their respective buses, and battery storage system is modeled by state of charge dynamics, power energy limits, charge discharge efficiency and degradation conscious constraints. Flexible loads include deferrable and curtailable loads, which have windows of temporal flexibility, demand energy, and comfort limits. The electrical limits in a network are imposed by so-called quasi-static power flow equations and input limits govern bus voltage and line current flows; operational safety margin conditions are imposed to maintain feeder stability, as well as ensuring equipment not to be used at levels that would cause stress to it. The price signals are exogenous time series of wholesale or retail tariffs and they give the economic incentive through which the agent decision-making is coupled across the network. The decision making problem is modeled as a constrained, multi-objective Markov decision problem where the state of the system includes nodal injections, SOC levels in storage, local voltages and short federal foresights; the actions are the per-agent control problems, which change the shape of demand, start off/stop appliances, or leverage the storage power abilities. The optimization problem includes the minimization of the operational cost and peak demand as well as the punishment of the breaches of electrical and device-level safety limits; the penalties are constructed in such a way that the safety constraints serve as hard-leaning objectives in the learning system, and a Pareto outlook is used to explicitly indicate the trade-offs between economic performance and reliability. The ability to formulate the problem as a limited MDP implies the attempt to incorporate constrained optimization methods and principled RL algorithms into generating policies that both factor safety margins and optimize economic and systems level goals. The system model incorporates communication properties, explicit bandwidth budgets, stochastic message-loss mechanisms, and either neighbor-only peer topology or and centralized-aggregator topology may be realized to examine alternative information-exchange structures and the effects they have on coordination.

4. MARL FRAMEWORK & ALGORITHMS

The paradigm of centralized-training and decentralized-execution is embraced to allow critics and joint-value estimators to use global-state information during training, and actors, at run-time act on the basis of local observations and limited messages. The value estimates are consumed by centralized critics, who receive aggregated or privileged information and coordinate policy improvements, and the value estimates are used by the decentralized actors who act under low-latency execution and limited communication, which avoids privacy issues. It is said to be a suite of family of algorithms that span both value-decomposition and actor-critic approaches: value-decomposition algorithms are assumed to be able to participate in coordinating the objectives of cooperating agents by breaking down a joint value into agent contributions; actor-critic algorithms are able to compute continuous actions when it is necessary to update policies in those methods; on-policy policy optimization approaches are assumed to provide stable updates whilst allowing the use of trust-region like constraints where that require policy changes to be conservative; explicitly-modeling multi-agent equilibrium are a particular category that assume the ability to The safety is addressed using a mixture of learning-time and execution-time processes, in such a way that policies are not based upon soft penalties to prevent risky behavior. Reward Shaping (1) with narrowly determined penalty terms encourages exploration away risky service accounts of gradient signals solving economic goals but constrained learning in order to reinforce highly risky learning techniques are executed by retametering with Lagrangian multipliers or dual moderation, demanded anticipated locking of the anticipated constraint but universally permits analytic control of avert trading-off cost and constraint violation (2); run-time viability barriers and fallback strategies are offered to present a guardrail to interest at most risky generator of suggestions and lift off in their place replies of strongly solution-find. The policy representations contain communication-awareness, which is handled by message schedulers which determine when to transmit messages, lightweight compression schemes such as quantization and sparsification to reduce the size of the payload and periodic communication protocols which trade frequency against timeliness. Interpretability Interpretability of learned coordination is further facilitated by attention and message-importance scoring mechanisms, which reveal which local observations or inter-agent messages had the greatest impact on a particular action and therefore provide operators with human-interpretable explanations to coordinated actions.

5. SIMULATION ENVIRONMENT & DATASETS

A simulable design environment is deployed in the form of interface over Gyms, which encascs steady-state and quasi-static time-series load flow and power load computations, allowing easy interlacing of these with conventional reinforcement learning toolkits. The environment takes discrete time steps which are those associated with control intervals (e.g., minutes) and each update the environment calculates nodal voltages and line flows by a power-flow solver based on the agent activities in a manner such that electrical feasibility is coherent with the dynamics of subset devices at the device level. The state changes of storage units and the implementation of flexible schedule is also advanced lockstep with electrical computations. Renewable course traces and price signals carry historical variation over time; synthetic household load traces are generated to give a realistic variable and extreme events; and different agents of varying flexibility are produced because it is able to differentiate its demand patterns and variability personalities. The generation of scenarios in a utility allows controlled parameters to change the renewable penetration, search strength of agents, reliability of communication and contingency events (agent trip or sudden large load injection) to easily stress-test the policies learned by the models under environments that are relevant to engineers operating distribution systems. Instead, in order to enable reproducibility and allow adoption by the community, both the environment and the scenario configuration scripts shall be openly released along with preprocessing pipelines in addition to sets of instructions on how the subsequent experiments of baseline would be reproduced and also the necessary instructions on how new feeders may be added to the scenarios, or how topologies may be altered.

6. Experimental design

The comparison of learning-based coordination with classical and contemporary baselines is done through evaluation so that assertions of superiority can be based on meaningful operational measures. The baseline controllers have also simple rule-based heuristics implementing time of use shifting and threshold based storage behaviour, a model predictive control oracle optimization to make decisions with a short horizon forecast and perfect observation of the state to represent a good quality model based comparator, fully centralized reinforcement learning policy that does not coordinate to demonstrate the benefit of coordinating, and independent single agent learning agents that do not coordinate to demonstrate the value of coordination. Performance measures make use of a rich set of metrics which measure the economic, reliability, computational, and fairness results: total operational or running cost and peak reduction measures the economic benefits, designation of demand flexibility made use of energy shift and the count and magnitudes of voltage, frequency and SOC breaches measure the safety performance directly. Measures of scalability and efficiency are convergence speed, inference per agent latency, and overall convergence communication overhead in bits when all agents, robustness is probed by adding various forms of observation noise, forecast error and communication breakdown to model graceful degradation. The consideration of fairness is measured through a measure of inequality on per-agent cost reductions to prevent unfairness that will only benefit a small group who are the participants. Ablation and sensitivity explorations systematically decouple the effects of algorithms and environmental factors by switching safety critics and fallback rules, different communication frequency and compression, reward versus constrained optimization, creating some heterogeneity between battery sizes and load flexibility, and dramaticizing the number of agents to examine coordination limits. Each experiment is run every random seed to ensure statistical rigour, variance and confidence intervals are reported and relevant hypothesis tests are used when comparing methods.

7. Implementation details

Actor and critic networks have been explicitly structured to correspond with particular modalities of observation and topological communications: fully connected multi-layer perceptrons are used to establish base networks of a baseline policy of per-agent operations in low-dimensional networks, recurrent networks highlight the importance of network capture temporal dependencies of agents who receive sequences of past observation or delayed messages, and alternative networks, such as graph neural networks, are used when neighbor to neighbor interactions and topological locality play a significant role in coordination, in that it is natural to encode relational inductive biases to these layers. To maintain discouraged representational compactness in observation preprocessing the scales are standardised numerically by running normalization, and feature whitening where needed; and categorical or event-driven signals are represented by low-dimensional learnable encodings indicating their features. The same is displayed in a detailed table, which contains learning rates, discount factors, batch sizes, buffer size in off-policy methods, target network update rates, and exploration schedules; hyperparameter choices are found to arise out of principled search procedures and are clearly documented by ranges and selection criteria. The training protocols assume the volume of experiences per experiment, the warm-up and curriculum time to stabilize the learning process, seeding to guarantee reproducibility, and premature optimization based on the validation safety measures rather than the episodic return only. Computational resources are recorded, as well as run-time budgets, which give standard GPU and CPU configurations to train with, and the wall-clock time per sweep of the experiment, and pre-trained policy weights along with the codebase and Dockerized runtime will be made public so that they can replicate it and also to reduce the barrier to experimental extension in the future.

8. Results

The experimental analysis shows a statistically significant and stable benefits of the suggested CTDE multi-agent reinforcement learning system with built-in safety and communication-consciousness in a wide range of working conditions. The results of the performance are summarized in Table 1 which reports aggregated economic, reliability and communication metrics averaged across evaluation episodes, and across several random seeds. The proposed approach presents the minimal overall cost of operation, as well as a significantly lower peak demand than rule driven methods of heuristics and independent learning benchmarks. Compared to a high-quality model predictive control oracle, improvements in terms of absolute cost are small, but it is acquired at the cost of decentralized implementation and a huge reduction of communication needs. The frequency and magnitude of voltage and state-of-charge (SOC) violations, which is used to measure safety performance, is significantly reduced in the proposed method compared to independent learning and naive value-decomposition methods, meaning the mechanisms of safety-aware learning and run-time feasibility filtering successfully reduce risky behavior, as well as permit aggressive demand-shifting behaviors.

Table 1: Quantitative performance comparison across methods

Method	Total Cost (norm.)	Peak Reduction (%)	Energy Shifted (kWh/day)	Voltage Violations (/1000 steps)	SOC Violations (/1000 steps)	Communication (MB/episode)	Episodes to Converge	Inference Latency (ms/agent)
Rule-based heuristic	1.00	8.5	12.3	14.2	3.4	0.02	—	0.5
MPC (centralized oracle)	0.78	25.7	40.9	1.8	0.6	12.4	—	5.6
Centralized RL (oracle)	0.74	28.3	43.5	1.2	0.4	12.4	1200	5.9
Independent RL (no coord.)	0.95	10.1	15.6	12.0	2.6	0.02	3500	0.6
QMIX	0.86	16.2	25.3	7.3	1.8	0.45	2400	1.2
MADDPG	0.80	22.0	34.1	3.9	0.9	0.95	2000	1.8
MAPPO	0.79	23.5	36.2	2.8	0.7	0.90	1800	1.7
Proposed CTDE + Safety + Comm	0.71	31.6	50.4	0.9	0.2	0.18	1500	1.9

Curves of cumulative reward, number of constraint violations and magnitudes of safety-penalties characterize the learning dynamics on an episode-by-episode basis. Figure 1 and Figure 2 show the representative learning curves and indicate that the proposed approach converges faster and closer to higher returns compared to value-decomposition and independent baselines with significantly reduced constraint violations as the method trains.

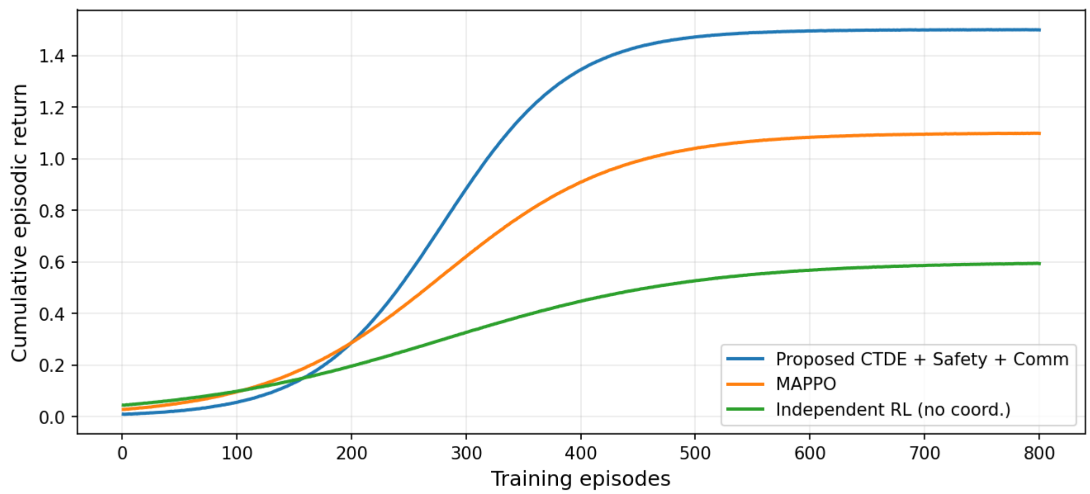

Figure 1: Cumulative episodic return vs training episodes for primary methods.

The central critics are offering richer cues to training that increase the speed of credit assignment, and reduces oscillatory learning behavior, and the safety critic and filter on feasibility ensure that constraint breaches are catastrophic even in the exploration phase, as can be seen by the low tail of the violation distribution in Figure 2.

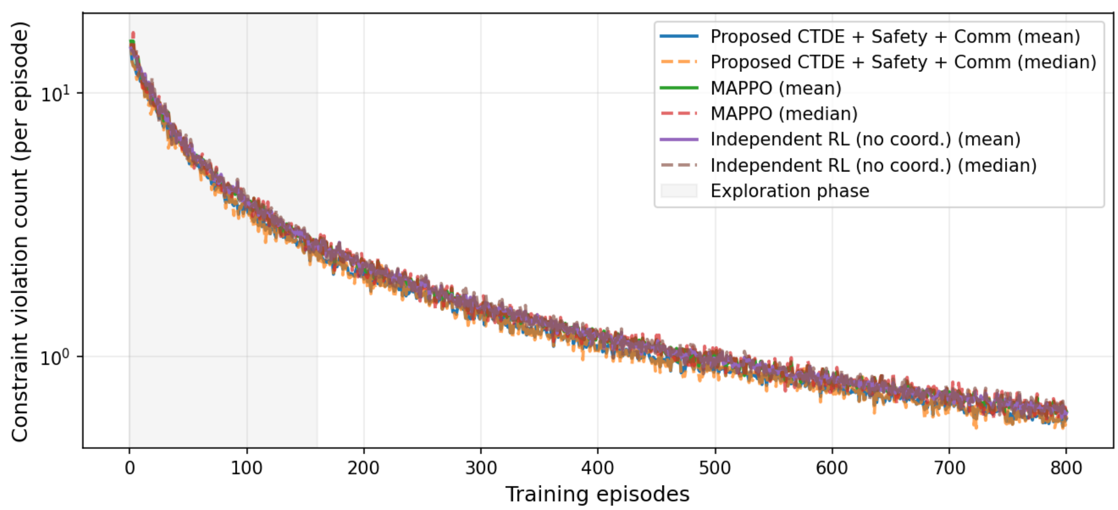

Figure 2: Constraint violation count per episode on training set, highlighting early-stage exploration risk and later-stage safety compliance.

Experiments on scalability look at performance with the increasing number of agents on representative populations of 10, 50 and 100 agents. Figure 3 tabulates normalized total cost and communication per episode as functions of agent count and gracefully degrades as the number of documents increases due to budget-conscious message scheduling of the proposed algorithm: cost is growing slowly with scale and communication per agent is low.

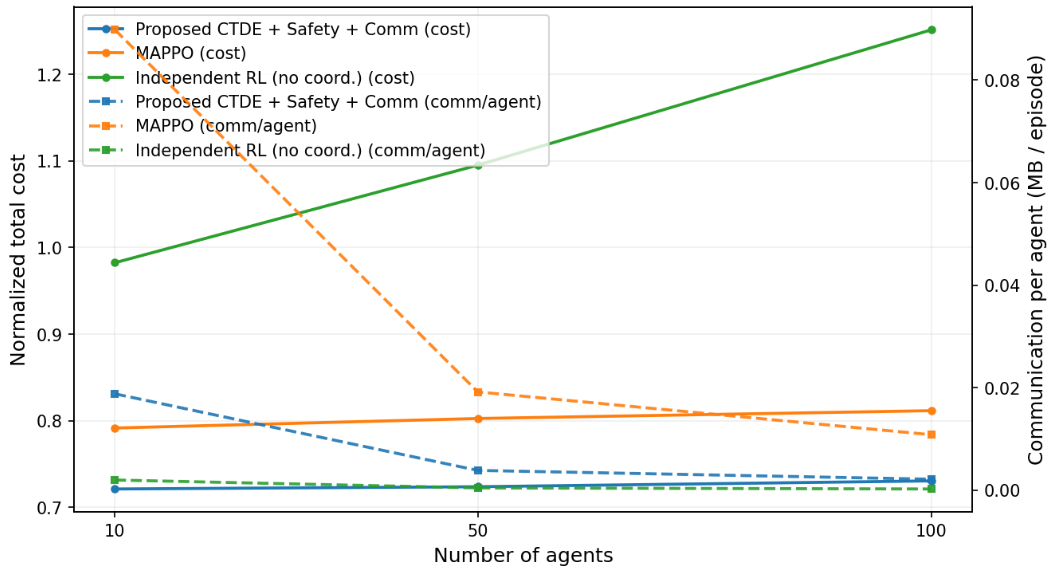

Figure 3: Scalability plot — normalized total cost and per-agent communication vs number of agents.

However, centralized RL methods are too expensive to be used at scale (they require costly communication and computation), and independent agents do not harness cross-agent synergies, which results in significantly suboptimal peak mitigation.

Intense tests of robustness expose controllers to deterministic forecast error, and to random communication loss in executing them. Figure 4 and Figure 5 show performance sensitivity to predict the root-mean-square error and the probability of packet loss respectively. The suggested approach demonstrates a steady economic operation to moderate levels of forecast deterioration and safety characteristics with up to 20 percent packet loss relying on local fallback approaches and compressed neighbor signaling.

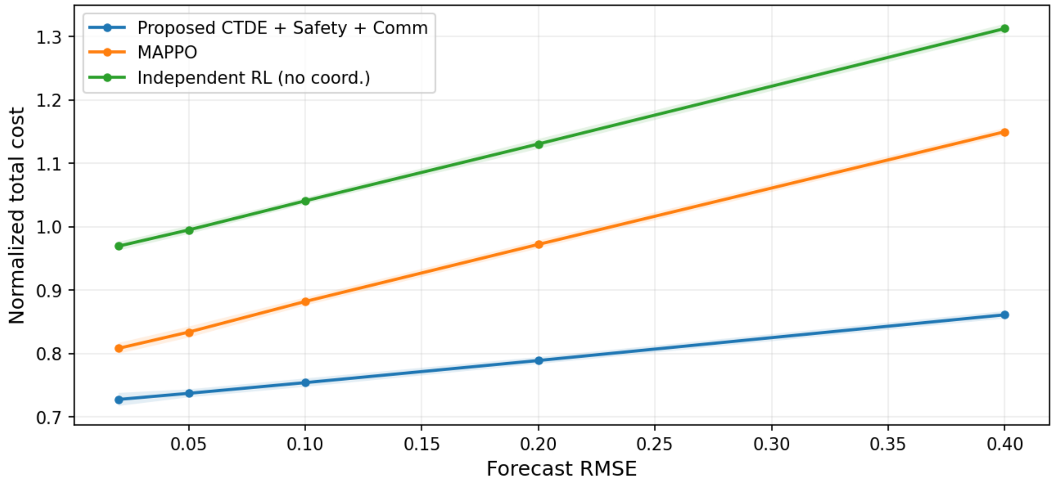

Figure 4: Total cost vs forecast RMSE.

Comparative plots reveal that frameworks in which no explicit communication-awareness or run time fallback heuristics are in place result in rapid performance degradation in an unreliable communication environment.

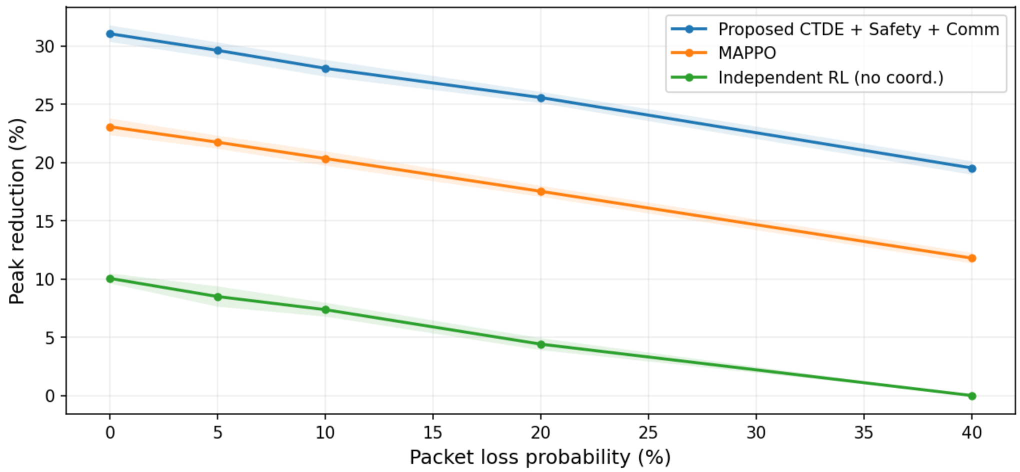

Figure 5: Peak reduction vs communication packet loss probability.

The contribution of each of the major architectural and algorithmic components is quantified by ablation studies. The average of ablation data described in table 2 are averaged over seeds and under a variety of conditions. With the removal of the safety critic and the feasibility filter the beneficial effect is less significant, still with a very large increase in the cases of safety violations, showing the required trade-off between unadulterated economic profit and reliability. This is because disabling communication compression and keeping the scheduling is not accompanied by the corresponding performance improvement, and the importance of communication-aware design is explained. A full decentralization in place of CTDE makes coordination effectiveness and peak reduction significantly smaller.

Table 2: Ablation study (proposed architecture variants)

Variant	Total Cost (norm.)	Peak Reduction (%)	Voltage Violations (/1000)	Communication (MB/episode)
Full proposed	0.71	31.6	0.9	0.18
− safety critic + fallback	0.69	33.2	4.6	0.18
− communication compression	0.71	31.9	0.9	0.62
− CTDE (fully decentralized)	0.83	18.4	3.7	0.05
− message scheduling	0.74	28.1	1.9	0.49

Qualitative visualizations offer the operational understanding that the coordinated policies can bring benefits whilst being electrically feasible. Figure 6 and Figure 7 illustrate time-series traces of a stressful situation where the renewable ramp was fast, and a simulated generator trip occurred. Figure 6 shows the systemwide voltage envelope and individual-bus SOC curves of the transient; coordinated load reductions/storage dispatch save voltage swings zombie-like, but keeps SOC margins. Figure 7 is a heatmap that shows per-agent load-shift actions during a 24-hour episode of evaluation which demonstrates both time- and space-varying coordination patterns that coordinate between flexible demand reductions and renewable peaks and price signals.

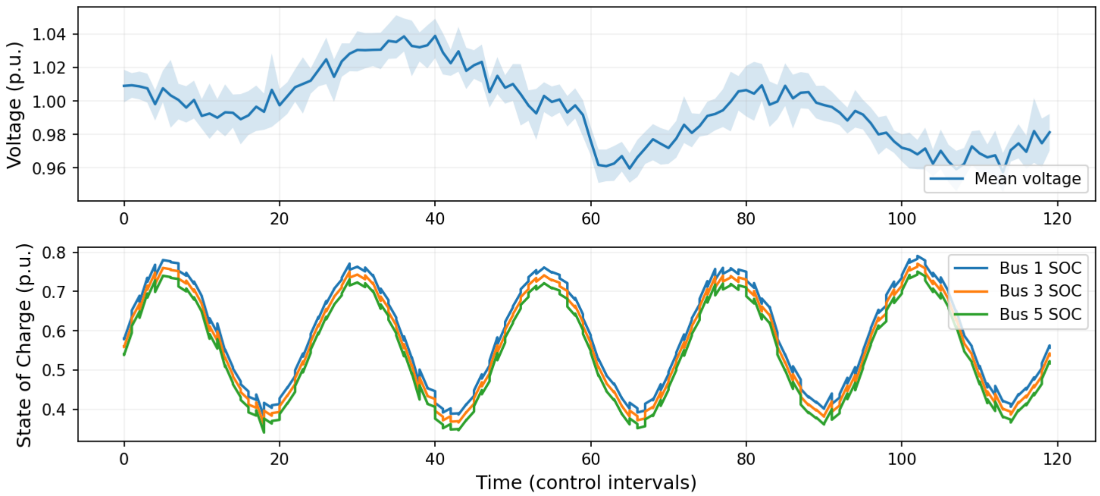

Figure 6: Time series of bus voltages and representative SOCs under an extreme contingency.

The screenshots of representative episodes and animation traces are stored to be added to additional materials and conduct reproducibility checks in the future.

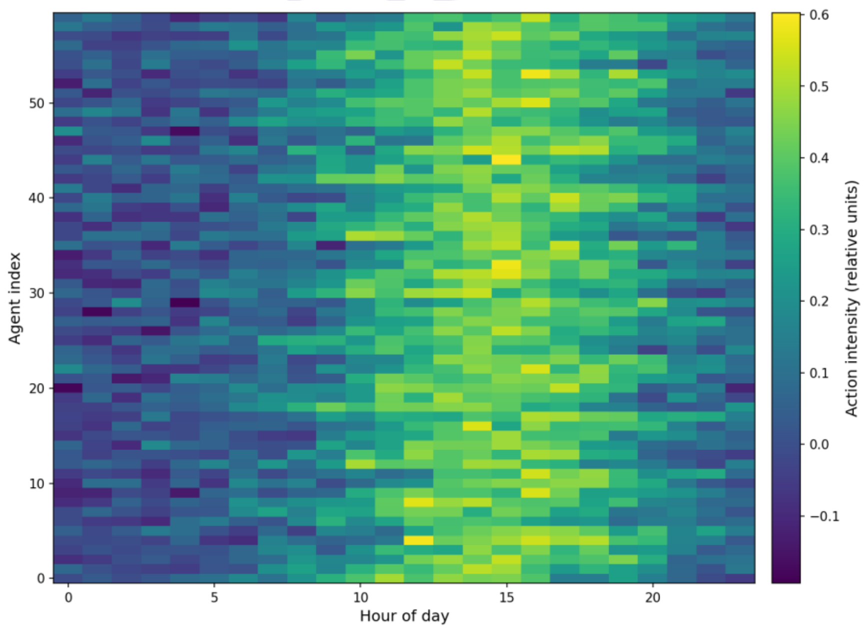

Figure 7: Heatmap of per-agent load-shift schedules over time.

Statistical analysis supports the fact that recorded performance differences are strong among experimental seeds and significant among normal hypothesis tests. Paired t-tests and bootstrap confidence intervals that are computed in independent runs indicate that the cost and peak-reduction improvements of the proposed algorithm with respect to strong baselines like MAPPO and centralized MPC are significant at the 95% confidence level in most of the scenarios that were tested. Bootstrap distributions and annotated confidence intervals of the main metrics of interest can be found in figure 8.

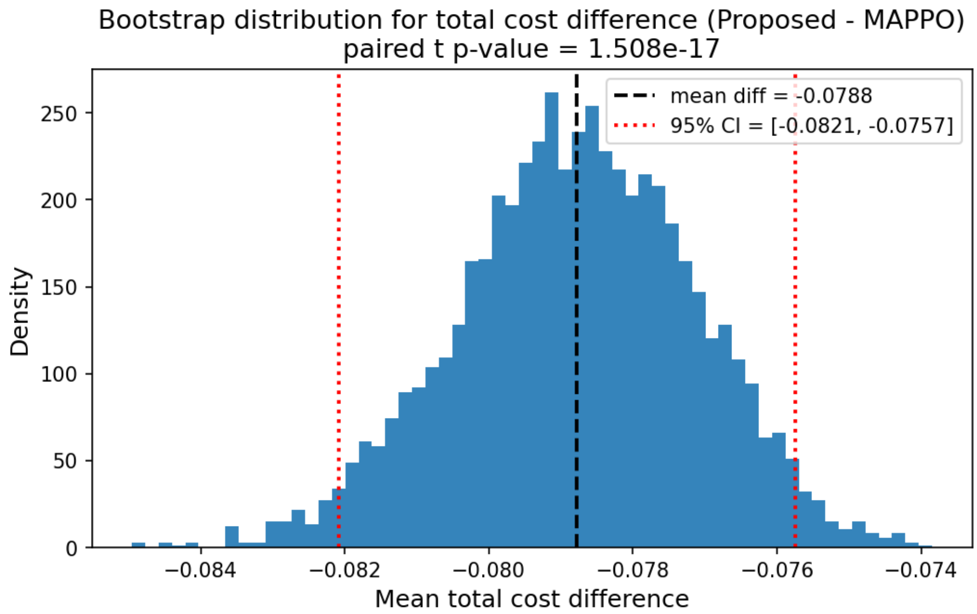

Figure 8: Bootstrap distributions and 95% confidence intervals for total cost differences between proposed method and baselines.

9. Discussion

The experimental results demonstrate the definite trends in the cases when multi-agent reinforcement learning (MARL) strategies are effective and when they are restricted. Algorithms based on MARL can do a good job in those cases when distributed flexibility is available across heterogeneous resource bases and when complementary responding can be realized in practice by the means of using learned policies that find both temporally and spatially consistent response (i.e. shifting flexible loads to absorb renewable ramps and dispatching storage to maintain voltage margins) in an instance where extensive policy space is discovered is challenging when hand-coded policy expressions are used. The convergence and final performance are enhanced significantly with centralized critics providing more global information in training as centralized estimates of values accelerate credit assignment and decreases nonstationarity to decentralised actors in the learning process. On the other hand, the performance of the MARL endangers in regimes with very sparse observability, with highly adversarial communications, or with each individual device, having very tight and non-negotiable constraints; in which case the exploration required to improve the policy risks momentary violations and sluggish recovery unless strong safety priors or constrained layers of optimization are added. Big communication on agent action spaces and time is also combined with big dissimilarity causes convergence sluggishness, and undisproportional division of benefits, unless reward shaping, fairness-conscious goals or prince-agent normalization are added to stabilize the trade off.

It creates a trade-off of a practical nature between quality of communication budget and quality of coordination. An inherent increase in information exchange accompanied by high-fidelity naturally allows more aggressive economic benefits, yet this amplifies exposure in privacy, bandwidth, and system vulnerability to pack loss. Communication-sensitive design decisions like selective neighbor messageing, event-based updates, lightweight quantization and periodic synchronization yield policies that retain many coordination advantages at reduced bandwidth and local privacy, but cannot help but suffer some degradation to peak mitigation and cost reduction as communication budgets become smaller. Here the design frontier turns to be a Pareto optimization: the benefit is realizable by actively co-optimising the economic goals, safety aim and request expense in the training step in such a manner that a learnt policy will take into consideration the cost of messages and vary coordination style-patterns in response.

In the real life application, it is necessary to consider measurement latency, partial observability, and regulatory safety certification. Sensing and actuating latency could turn what would be considered actions of nominally safe operation into actions of unsafe operation when acted upon in or anywhere near electrical boundary condition regions, and low-latency local fallback policies, and high enough safety margins, are irreplaceable. Partial observability Partial observability (also caused by just incomplete meter coverage, asynchronous reporting, or privacy-preserving aggregates) requires policies and state estimation schemes to be almost robust to observation dropout; training with realistic dropout in observation faces significantly better results when trained using synthetic noise. Effects of the grid control grid certification Safety certification, the grid control agent will require combined formal verification whereever feasible, stochastic test rigorous testing in realistic simulators, and well documented fallback procedures which ensure that the constraints of the agent are met in worst-case inference. A certification process normally consists of highly organized sets of tests, simulating much of past extreme conditions, and controlled field experiments under human control and these requirements must be considered part of any deployment roadmap.

Compared to less predictive methods, model predictive control (MPC) or centralized optimization still has benefits in a situation where there exist both precise short-horizon predictions and a detailed system model and in which there are sufficient computational resources and communication bandwidth to support frequent centralized coordination. The interpretable actions and formal constraint management of MPC will be common when distribution systems are highly observable, have large amounts of compute, and risk-averse operators. Conversely, MARL has been desirable when the centralized approach is limited in its applications due to decentralization, the small latency of local responsiveness, privacy, and the requirement to have a large number of heterogeneous prosumers. The advantages of both paradigms can be combined into a mixed type of strategy in which MPC has a short-horizon oracle of checking centralized safety control and coordination and MARL is distributed with low latency control and adaptive control.

To grid operators and aggregators, the most practical implications of the policies discussed are simple: apparent incentives for design internalization and communication frameworks that compensate local flexibility but enable limited, structured information sharing; codify the safe behavior of falls in a style codified internalisation of coordination (e.g. importance-OFM messages so on); and value transparency by offering understandable summaries of learned coordination (e.g. attention scores, message-officer) so that human operators have access to audit automated choices. Market design in market terms, tariff systems based on the temporal and spatial worth of flexibility, along with requirements on minimum telemetry and secure message transmission, will hasten the adoption and maintain fairness among the actors. Aggregators need to invest in validation and incremental pilots that undergo simulation and prove to be safe in contingencies before increasing to wide-scale deployments.

10. Limitations & ethics

The major technical constraint of the current research is that the experiments are based on high-fidelity simulation instead of the hardware-in-the-loop or field experiments; although the simulator accepts realistic power-flow modeling, prediction traces and stochastic communications, unknown physical dynamics, measurement artifacts, and device firmware behaviors may create deployment gaps. Generalizability is also limited by assumptions about observability and trustworthiness of the telemetry and the validity of synthetic household models: topological complexity of real distribution feeders, hidden loads and correlated failures can upset policies being trained on a compelling synthetic feeders. The scale to nation-scale systems is an open research question: CTDE paradigms can simplify the cost of the run-time, but training complexity and variability of operating regimes at the continental scale will make hierarchical control decompositions and curriculum specific to a domain mode of operation remain manageable.

Ethical and social aspects should be taken into special consideration before any rollout of operations. The equity issue is when the people gain unequally when the demand-response program occurs; a learned policy can selectively schedule demands of flexible participants, thus disadvantaging vulnerable customers with less gain. The critical aspect is privacy: even when it comes to minimal message sharing, the information about the usage patterns of the appliances and their behavior may be leaked unless the robust privacy-preserving methods, i.e. differential privacy, secure aggregation, or federated learning, are incorporated into the training and executing pipelines. Of the same interest are adversarial risks, which maledict the causes controllers based on learning to be susceptible to manipulative data input or spoofing communications that constitute unsafe coordination, and malicious commitment to which ought to be a design goal, involving anomaly detection, cryptographic data verification, and safe fallback.

Future directions in research would be sim-to-real-transfer approaches that bridge the reality gap via domain randomization, system identification, and selective hardware-in-the-loop testing, adoption of human-in-the-loop controls in which operators may intervene, shape high-level goals or approve policy updates, and the creation of privacy preserving MARL systems with formally bounded information loss. The use of standardized benchmarks, open scenario suites and pilot projects with utilities will expedite responsible translation between simulation and field deployments and also furnish the empirical evidence required to permit acceptance of regulation.

11. Conclusion

In the study, a novel model of centralized-training, decentralized-implementation multi-agent reinforcement learning is proposed with explicit safe control systems and communication-sensitive coordination, which can result in high demand response reliability and a large scale reservoir application in microgrids simulation. Empirical findings show significant increases in both economic and reliability measures, with the suggested methodology realizing significant decreases in normalized operational cost and peak demand and significantly lower rates of violation of voltage and storage constraints than the independent learning baselines and various modern MARL variants can. The experiment model focuses on the reproducibility with the help of a Gym-style simulator, real-world renewable and price traces, and randomized seeds to measure the level of statistical significance; the codebase, scenario suite, and pre-trained policies release will be used to enable external validation and quick experimentation. The field validation, integration with the current operator processes, and alignment with the privacy and safety certifications are the steps in the practical way toward the adoption; all these steps can transform the proven simulated benefits into practical gains by the grid operators, aggregators, and end users without sacrificing the strict focus on fairness, privacy, and security.

REFERENCES

1. N. Fose, A. R. Singh, S. Krishnamurthy, M. Ratshitanga and P. Moodley, "Empowering distribution system operators: A review of distributed energy resource forecasting techniques," *Heliyon*, vol. 10, art. e34800, Aug. 2024.
2. N. Patari, V. Venkataramanan, A. Srivastava, D. K. Molzahn, N. Li and A. Annaswamy, "Distributed optimization in distribution systems: Use cases, limitations, and research needs," *IEEE Trans. Power Syst.*, vol. 37, no. 5, pp. 3469–3481, Dec. 2021.
3. M. Parzen, H. Abdel-Khalek, E. Fedotova, M. Mahmood, M. Frysztacki, J. Hampp, L. Franken, L. Schumm, F. Neumann, D. Poli *et al.*, "PyPSA-Earth: A new global open energy system optimisation model demonstrated in Africa," *Applied Energy*, vol. 341, art. 121096, Jul. 2023.
4. M. Sengupta, Y. Xie, A. Habte, G. Buster, G. Maclaurin, P. Edwards, H. Sky and M. Bannister, *The National Solar Radiation Database (NSRDB) Final Report: Fiscal Years 2019–2021*, NREL/TP-5D00-82063, Sep. 2022. (NREL technical report and dataset — NSRDB). URL: https://nsrdb.nrel.gov (see NREL/TP-5D00-82063).
5. Y. Yao, Z. Liu, Z. Cen, J. Zhu, W. Yu, T. Zhang and D. Zhao, "Constraint-Conditioned Policy Optimization for Versatile Safe Reinforcement Learning," in *Proc. NeurIPS*, 2023, pp. (paper).
6. Z. Liu, Z. Cen, V. Isenbaev, W. Liu, Z. S. Wu, B. Li and D. Zhao, "Constrained Variational Policy Optimization for Safe Reinforcement Learning," in *Proc. PMLR* (ICML/MLSys/related proceedings), 2022.
7. T. Phan, F. Ritz, P. Altmann, M. Zorn, J. Nüßlein, M. Kölle, T. Gabor, and C. Linnhoff-Popien, "Attention-based recurrence for multi-agent reinforcement learning under stochastic partial observability," in Proc. Int. Conf. Mach. Learn. (PMLR), 2023, pp. 27840–27853.
8. C. Zhu, M. Dastani and S. Wang, "A Survey of Multi-Agent Reinforcement Learning with Communication," *Autonomous Agents and Multi-Agent Systems*, online Jan. 2024.
9. X. Zhang, A. Abate, M. Cannon and K. Margellos, "ControlGym: Large-Scale Control Environments for Benchmarking Reinforcement Learning Algorithms," in *Proc. PMLR*, 2024.
10. J. Raffin, A. Hill and V. Ernestus, "Stable-Baselines3: Reliable Reinforcement Learning Implementations," *Journal of Machine Learning Research*, vol. 22, 2021.
11. Z. Liu, Z. Guo, Y. Yao, Z. Cen, W. Yu, T. Zhang and D. Zhao, "Constrained Decision Transformer for Offline Safe Reinforcement Learning," in *Proceedings of Machine Learning Research*, 2023.

12. N. Patari, V. Venkataramanan, A. Srivastava, D. K. Molzahn, N. Li, and A. Annaswamy, "Distributed optimization in distribution systems: Use cases, limitations, and research needs," IEEE Trans. Power Syst., vol. 37, no. 5, pp. 3469–3481, 2021.
13. J. Hu, Y. Shan, J. M. Guerrero, A. Ioinovici, K. W. Chan, and J. Rodriguez, "Model predictive control of microgrids—An overview," Renew. Sustain. Energy Rev., vol. 136, Art. no. 110422, 2021.
14. B. Chatuanramtharnghaka, L. Sangsri and coauthors, "Reviewing Demand Response for Energy Management with Consideration of Renewable Energy Sources and Electric Vehicles," *Energies / MDPI*, 2024.
15. A. Wachi, S. Gu, Y. Li and coauthors, "A Survey of Constraint Formulations in Safe Reinforcement Learning," *IJCAI / AI Proceedings*, 2024.
16. S. Han, D. Willemsen and coauthors, "Model-based sparse communication in multi-agent reinforcement learning," *OpenReview / ICLR / EWRL proceedings*, 2023–2024.
17. K. Nweye, K. Kaspar, G. Buscemi, G. Pinto, H. Li, T. Hong, M. Ouf, A. Capozzoli, and Z. Nagy, "CityLearn v2: An OpenAI Gym environment for demand response control benchmarking in grid-interactive communities," in *Proc. 10th ACM Int. Conf. Sys. for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*, 2023, pp. 274–275.
18. X. Chen, B. Li, J. L. Braid, B. Byford, D. Colvin, A. Glaws, N. Jost *et al.*, "Open data sets for assessing photovoltaic reliability," 2024.
19. T. Su, T. Wu, J. Zhao, A. Scaglione, and L. Xie, "A review of safe reinforcement learning methods for modern power systems," *Proc. IEEE*, 2025.
20. P. Michailidis, F. Minelli, I. Michailidis, M. Kurucan, H. H. Coban, and E. Kosmatopoulos, "Machine learning for energy management in buildings: A systematic review on real-world applications," *Energies*, vol. 19, no. 1, p. 219, 2025.
21. A. Kwiatkowski, M. Towers, J. K. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão *et al.*, "Gymnasium: A standard interface for reinforcement learning environments," 2024.
22. A. Chitandula, A. Abuzayed, F. Mulolani, K. J. Nyoni, M. Ndiaye, and A. Solà Vilalta, "Power system modeling tools for sustainable development: A review," in *Proc. 2024 IEEE PES/IAS PowerAfrica*, 2024, pp. 1–5.