

REVIEW

Open access

An Explainable Reinforcement Learning Framework for Demand Response Dispatch Using Customer Flexibility Profiles, Reward Decomposition, and Operator Review

Lars Andersen^{1*}, Freya Nielsen¹, Mette Sørensen²

Abstract

Reinforcement learning can autonomously dispatch demand response actions under changing prices, loads, and grid constraints. However, operator acceptance requires explanations that clarify why specific customers or aggregations are selected. This article presents an explainable reinforcement learning framework for demand response dispatch using customer flexibility profiles, decomposed rewards, and operator review. The objective is to preserve dispatch performance while improving transparency and controllability. The framework combines a profile-aware RL dispatcher, a decomposed reward model, an explanation generator, and a structured operator review protocol. Evaluation is performed through simulation on realistic demand response scenarios with household and flexible-load representations. The proposed explainable RL framework is designed to achieve dispatch performance comparable to black-box RL while providing interpretable action rationales. Operator review is incorporated to approve, reject, or modify dispatch recommendations before execution. The framework bridges autonomous RL control and trustworthy grid-edge operation. It provides an algorithmic blueprint for human-centered demand response dispatch.

Keywords Explainable reinforcement learning, Demand response dispatch, Customer flexibility profiles, Reward decomposition, Operator review

*Correspondence:

Lars Andersen

lars.andersen@gmail.com

¹ Department of Power Systems Analytics, Faculty of Engineering, Aalborg University, Aalborg, Denmark

² Department of Smart Energy Optimization, Faculty of Engineering, University of Copenhagen, Copenhagen, Denmark

Introduction

Reinforcement learning has become a strong candidate for demand response dispatch because it can learn sequential control policies under uncertainty, delayed effects, and nonlinear customer response. The review by Vázquez-Canteli and Nagy [1] frames RL as especially useful when explicit demand response models are difficult to specify, while Mocanu and colleagues [2] show that online building energy optimization can be learned directly from operational interaction. In demand response, this capability allows an agent to adjust load-shifting signals, incentives, or curtailment requests based on observed grid and customer states. Yet the same adaptivity that makes RL attractive can make its decisions difficult for operators to interpret.

Recent demand response studies show that RL can reduce cost, shift flexible load, and respond to dynamic prices, but most dispatch decisions remain opaque to the human operator. Lu, Hong, and Zhang [3] demonstrate price-based demand response using reinforcement learning, and Lu and Hong [4] extend this logic to incentive-based programs with

neural approximators. Deep RL has also been applied to interruptible loads by Wang, Li, Ming, and Wang [5], indicating that RL can coordinate resources that have different response costs and operational limits. The remaining challenge is not only whether an RL policy performs well, but whether its actions can be explained, reviewed, and trusted.

Customer flexibility profiles are essential for making RL-based demand response dispatch more transparent because they describe how much load can move, how quickly it can respond, and what comfort or service constraints apply. Real-time residential demand response by Li, Wan, and He [6] shows the importance of customer-level response modeling, while work on whole-building HVAC demand response by Azuatalam, Lee, de Nijs, and Liebman [7] emphasizes the role of thermal flexibility. Profiling makes the RL action space less arbitrary because each dispatch recommendation can be linked to a customer's estimated flexibility class. This article therefore treats flexibility profiles not as auxiliary data, but as a structural input to policy design and explanation.

The contribution of this article is an original explainable RL framework that integrates customer flexibility profiles, reward decomposition, and operator review into one demand response dispatch architecture. Reward decomposition is used to separate cost saving, comfort preservation, peak reduction, and constraint satisfaction so that the operator can see why the policy prefers one dispatch action over another, consistent with the broader need for interpretable RL identified by Heuillet, Couthouis, and Díaz-Rodríguez [8]. The framework also draws on recent explainable RL surveys by Milani, Topin, Veloso, and Fang [9] and Glanois and colleagues [10] to translate policy behavior into operator-facing rationales. Unlike a purely conceptual model, the article specifies the RL algorithm, training setup, benchmark comparison, and simulation-based evaluation procedure.

Demand Response Dispatch Problem

Operational context

Demand response dispatch is modeled as a sequential decision problem in which an aggregator or distribution operator selects control actions for households, buildings, EV chargers, water heaters, or flexible industrial loads. Online pricing work by Kong, Kong, Yao, Bai, and Xiao [11] shows that dispatch signals can be updated as system and market conditions evolve, while multi-agent demand response in manufacturing by Lu, Li, Li, Jiang, and Ding [12] demonstrates that coordinated control can be learned across many flexible units. At each decision interval, the operator observes load forecasts, prices, network conditions, and flexibility states before issuing dispatch instructions. The operational objective is to reduce peak demand and energy cost without violating comfort, availability, or participation constraints.

Dispatch objectives and constraints

The dispatch problem optimizes multiple objectives that are often in tension: minimizing energy expenditure, reducing peak load, preserving customer comfort, limiting excessive activations, and respecting device constraints. Incentive-based demand response models developed by Wen, Zhou, Li, and Wang [13] show that learning-based control must balance economic reward with user response, while the testbed implementation by Zhang, Lu, Jiang, Hong, and Song [14] highlights the importance of executable dispatch logic. In this article, each customer or aggregation has a feasible action set defined by flexibility class, device type, response duration, rebound risk, and maximum tolerable interruption. These constraints are encoded directly in the RL environment so that invalid actions are masked before policy selection.

The RL opportunity

Reinforcement learning is well suited to demand response because it learns from repeated interaction rather than relying exclusively on perfect forecasts or static optimization models. Coordinated building control by Pinto, Piscitelli, Vázquez-Canteli, Nagy, and Capozzoli [15] shows that deep RL can manage clusters of buildings, and EV charging studies by Wan, Li, He, and Prokhorov [16] demonstrate the value of model-free scheduling under uncertain mobility behavior. The proposed framework extends this opportunity by making the policy profile-aware and explanation-aware rather than

treating the learned controller as a black box. The resulting formulation is a constrained Markov decision process with interpretable state features, decomposed rewards, and an operator approval layer.

Customer Flexibility Profiles

Defining flexibility

Customer flexibility is defined as the time-, magnitude-, and comfort-bounded ability of a customer or aggregation to modify demand in response to a dispatch signal. Building flexibility characterization by Junker and colleagues [17] provides a foundation for representing flexibility as a measurable system property, and Zhang, Good, and Mancarella [18] show how residential heat pump aggregations can provide building-to-grid flexibility. In the proposed framework, flexibility is represented by four variables: shiftable energy, response latency, comfort sensitivity, and rebound probability. These variables become part of the RL state so that the policy can distinguish high-value flexible customers from customers whose activation would impose high comfort or rebound costs.

Profiling methods

The profiling module estimates flexibility using historical smart-meter load, device availability, weather, price exposure, and prior demand response behavior. Pallonetto, De Rosa, Milano, and Finn [19] use machine learning models for smart-grid-ready residential buildings, showing how data-driven methods can infer demand response capability from building behavior. Xiao, Xie, Fang, and Wang [20] further motivate clustering for customer selection, while Shirsat and Tang [21] quantify residential demand response potential using probabilistic neural models. In this article, clustering first groups customers by load-shape and response patterns, then regression estimates numerical flexibility parameters for use in dispatch.

Flexibility clusters and attributes

The framework uses four operational clusters that summarize typical customer flexibility behavior: thermal-shiftable, appliance-shiftable, EV-flexible, and low-flexibility protected. Large-scale heat pump demand response by Müller and Jansen [22] supports the treatment of thermostatically controlled loads as precise flexible resources, while stochastic nonlinear price-based flexibility modeling by Junker, Kallesøe, Real, Howard, Lopes, and Madsen [23] motivates probabilistic response parameters. EV charging flexibility is modeled separately because charging decisions depend on departure time, state of charge, and user mobility uncertainty, as shown by Yan, Chen, Zhou, Chen, and Wen [24]. Each cluster is assigned an action mapping that limits the RL dispatcher to realistic curtailment, shifting, or incentive actions.

Integration with RL action space

The flexibility profile is integrated with the RL action space through action masks, cluster-specific activation limits, and profile-dependent reward terms. Home energy management work by Xu, Jia, Xu, Xu, Chai, and Lai [25] supports using multi-agent representations for household control, while Mathew, Roy, and Mathew [26] show that residential energy management can be formulated as an intelligent learning problem. In the proposed dispatcher, an action is represented as a tuple containing customer group, dispatch type, activation magnitude, duration, and incentive level. **Table 1** summarizes the customer clusters, their dispatch-relevant attributes, and the action mappings used by the RL policy.

Table 1. Customer Clusters, Dispatch Attributes, and Reinforcement Learning Action Mappings with Associated Operational Risks

Cluster	Flexibility characteristics	Key state features	Feasible RL action mapping	Main operational risk

Thermal-shiftable	HVAC or heat-pump load can be pre-cooled, pre-heated, or temporarily reduced within comfort bands	Indoor temperature margin, outdoor temperature, thermal inertia, prior activation	Shift 0.5–2.0 kW for 1–4 intervals with comfort-bound masking	Comfort violation and rebound peak
Appliance-shiftable	Deferrable appliances can move consumption within a preferred completion window	Remaining task duration, deadline, user preference score, baseline load	Delay or advance flexible task start within accepted window	Missed service deadline
EV-flexible	Charging can be delayed, accelerated, or modulated before departure	State of charge, target charge, departure probability, charger rating	Modulate charging rate or defer charging with minimum state-of-charge constraint	Mobility dissatisfaction
Low-flexibility protected	Customer has limited or uncertain ability to respond without inconvenience	Low response probability, high opt-out history, critical load indicator	No dispatch or low-intensity voluntary incentive only	Trust erosion and opt-out

RL Framework Design

State and action space

The RL environment state at time step (t) combines grid, market, and customer-profile variables into a structured vector (s_t). It includes feeder demand, forecasted peak risk, real-time or day-ahead price, weather, customer cluster labels, estimated shiftable load, comfort margin, prior activation count, and rebound risk, extending the state representations used in demand response and building-control RL studies [3]. The action (a_t) selects a customer group or aggregation, dispatch intensity, duration, and incentive level, while infeasible actions are removed using profile-dependent masks. This formulation allows the policy to learn not only when demand response is valuable, but which flexibility profile should be activated.

Policy architecture

The policy architecture is a dueling deep Q-network for discrete dispatch choices, with an actor-critic variant used for continuous incentive and magnitude tuning. Deep RL for interruptible loads by Wang, Li, Ming, and Wang [5] supports value-based dispatch learning, while continuous EV charging control by Yan, Chen, Zhou, Chen, and Wen [24] supports policy-gradient treatment of continuous control variables. The network contains separate encoders for grid state, price state, and flexibility-profile state, followed by a fusion layer that outputs action values and component-wise reward expectations. This architecture is intentionally modular so that explanations can refer to the profile, price, and grid features that most influenced the action.

Training algorithm

Training is performed in simulation using historical or synthetic load traces, randomized price scenarios, weather-driven flexibility variation, and stochastic customer acceptance models. The testbed approach of Zhang, Lu, Jiang, Hong, and Song [14] motivates evaluating learned demand response policies in executable environments before deployment, and safe deep RL for active distribution networks by Kou, Liang, Wang, Wu, and Gao [27] motivates constraint-aware learning. The agent is trained with experience replay, target networks, epsilon-greedy exploration, and action masking, with penalties for comfort violations and excessive customer activation. Training episodes represent daily operating horizons, and convergence is assessed using moving-average total reward and constraint-violation frequency.

Integration with demand response signals

The trained dispatcher converts actions into operational demand response signals, including price incentives, load-shifting requests, charging schedules, or curtailment recommendations. Real-time residential demand response by Li, Wan, and He [6] supports direct household-level control, while coordinated cluster management by Pinto, Piscitelli, Vázquez-Canteli, Nagy, and Capozzoli [15] supports aggregation-level dispatch. The proposed system supports both individual and aggregation dispatch by choosing the resolution according to profile confidence and operator policy. When profile uncertainty is high, the framework favors lower-intensity incentives and routes the recommendation to operator review before execution.

Figure 1 illustrates how customer flexibility profiles, decomposed reward evidence, explanation generation, and operator review are integrated into a profile-aware reinforcement learning framework for demand response dispatch.

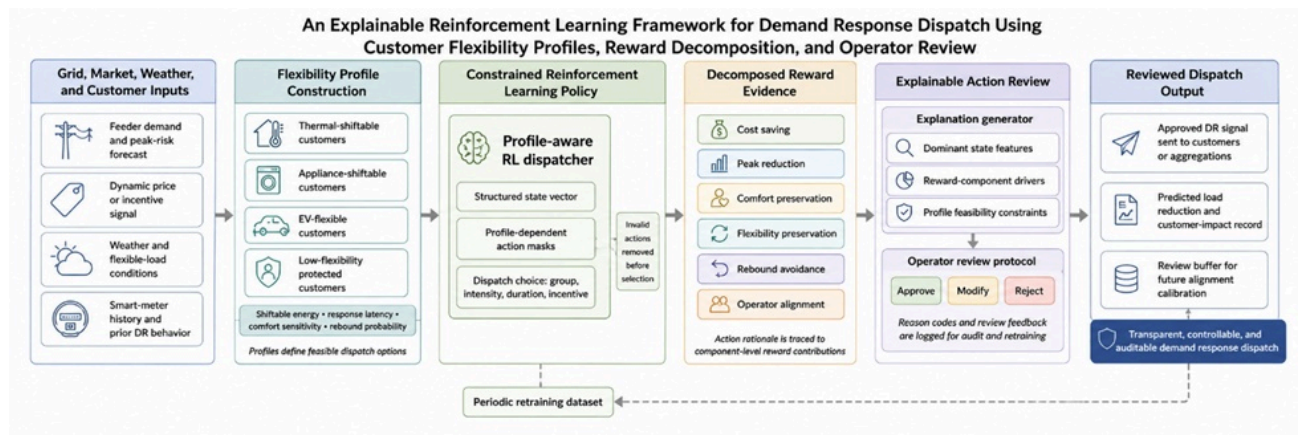


Figure 1. Explainable Reinforcement Learning Architecture for Operator-Reviewed Demand Response Dispatch

Reward Decomposition

Motivation for decomposed rewards

A single scalar reward can produce effective control, but it does not tell the operator whether an action was selected for cost saving, peak reduction, comfort preservation, or risk avoidance. Explainable RL research by Heuillet, Couthouis, and Díaz-Rodríguez [8] identifies this lack of interpretability as a barrier to understanding sequential decision policies. The proposed framework therefore records a reward vector at each step before combining it into the scalar signal used for RL optimization. This makes the training objective compatible with standard RL while preserving component-level evidence for operator review.

Reward components definition

The reward vector contains six components: energy-cost saving, peak-reduction value, comfort preservation, flexibility availability preservation, rebound avoidance, and operator-alignment score. Dynamic pricing RL by Lu, Hong, and Zhang [3] motivates explicit cost saving, while incentive-based DR by Lu and Hong [4] motivates modeling the economic value of participation. Energy flexibility reviews by Li, Wang, Hong, and Piette [28] support treating flexibility as a resource that should not be exhausted without operational justification. The scalar training reward is calculated as a weighted sum of normalized components, with hard penalties applied when safety or comfort constraints are violated.

Weighting and normalization

Each reward component is normalized to a bounded interval so that high-magnitude cost terms do not dominate comfort or trust-related objectives. Economic dispatch work by Lin, Guan, Peng, Wang, Maharjan, and Ohtsuki [29] supports

balancing cost-oriented control with operational objectives in virtual power plant settings. In the baseline setting, cost saving receives the largest positive weight during normal operation, while peak reduction receives a higher weight during feeder stress events. Comfort and rebound components are treated as protective terms, ensuring that the policy cannot achieve high reward by repeatedly activating the same sensitive customers.

Interpretability via component tracking

During both training and evaluation, the framework logs each reward component for every action and aggregates the values by customer cluster, dispatch type, and operating condition. Surveys of explainable RL by Milani, Topin, Veloso, and Fang [9] and interpretable RL by Glanois and colleagues [10] support the use of decomposed evidence as a bridge between learned behavior and human-understandable explanations. The operator interface reports the top contributing reward components for each recommendation, such as “peak reduction dominates cost saving” or “comfort penalty limits dispatch intensity.” **Table 2** defines the reward components, their operational interpretation, and the baseline weights used in the simulation.

Table 2. Reward Components, Operational Interpretation, Normalization, and Baseline Weights in RL Simulation

Reward component	Definition	Normalization rule	Baseline weight	Interpretive contribution
Cost saving	Difference between baseline energy cost and dispatched energy cost over the control interval	Divided by maximum observed daily controllable cost saving	0.30	Explains economic motivation for dispatch
Peak reduction	Reduction in feeder or aggregation peak relative to forecasted baseline	Divided by target peak-reduction capacity	0.25	Explains grid-service value during constrained periods
Comfort preservation	Negative penalty for temperature, service, or charging dissatisfaction	Scaled by maximum allowable comfort deviation	0.20	Explains why sensitive customers are protected
Flexibility preservation	Penalty for overusing customers with limited remaining flexibility	Scaled by remaining activation budget	0.10	Explains conservation of scarce future flexibility
Rebound avoidance	Penalty for predicted post-event load recovery that creates a secondary peak	Divided by forecasted rebound-risk threshold	0.10	Explains avoidance of delayed system stress
Operator alignment	Positive score when the action resembles historically approved operator decisions	Estimated from reviewed action history	0.05	Explains consistency with human operating preferences

Explainability Layer

Post-hoc explanation methods

The explainability layer converts learned dispatch decisions into structured evidence that can be inspected before implementation. Explainability in deep reinforcement learning has emphasized post-hoc tools such as feature attribution, saliency, policy summarization, and surrogate models, and Heuillet, Couthouis, and Díaz-Rodríguez [8] identify these

methods as central for understanding learned sequential control. In the proposed framework, feature attribution is computed over grid-state, price-state, and flexibility-profile inputs to estimate which variables most influenced the selected action. A shallow decision-tree surrogate is also trained on policy rollouts to provide a compact approximation of the learned dispatch logic for operator review.

Mapping actions to flexibility profiles

Each recommended action is mapped back to the customer flexibility profile that enabled the dispatch decision. Customer selection through deep clustering, as developed by Xiao, Xie, Fang, and Wang [20], supports linking dispatch eligibility to observed response patterns rather than treating all customers as interchangeable. Probabilistic flexibility estimation by Shirsat and Tang [21] further motivates reporting uncertainty around expected response magnitude and customer availability. The explanation layer therefore states not only what action is recommended, but also why the selected profile is suitable under the current price, load, comfort, and rebound conditions.

Generating operator-friendly explanations

The operator-facing explanation is generated from three information streams: dominant state features, decomposed reward contributions, and profile-specific feasibility constraints. Surveys by Milani, Topin, Veloso, and Fang [9] show that explainable RL should support human understanding of both policy behavior and outcome trade-offs, while Glanois and colleagues [10] emphasize interpretability as a means of improving human confidence in RL decisions. The framework produces concise textual rationales such as “activate EV-flexible cluster at medium intensity because departure risk is low, peak-reduction reward is high, and comfort penalty is negligible.” These explanations are paired with numerical evidence so that the operator can inspect the proposed action without manually reconstructing the RL objective.

User interface considerations

The user interface is designed around action review rather than passive visualization, because demand response dispatch must be executed within limited operational time windows. Testbed implementation of RL-based demand response by Zhang, Lu, Jiang, Hong, and Song [14] demonstrates the value of practical interfaces that connect learning algorithms to executable control signals. The proposed interface displays the recommended dispatch, affected customer cluster, predicted load reduction, reward-component bars, confidence score, and override options on a single review screen. This design supports fast operator action while preserving traceability for post-event audit and model improvement.

Operator Review Process

Review workflow

The operator review process begins after the RL dispatcher produces a ranked set of feasible demand response actions for the next control interval. Human-in-the-loop review is especially important because RL controllers for building energy and demand response, such as the systems studied by Azuatalam, Lee, de Nijs, and Liebman [7], may operate under weather and occupancy uncertainty. The operator receives the highest-ranked action, its explanation, reward decomposition, forecasted load impact, and risk indicators. The workflow is intentionally structured so that review adds oversight without requiring the operator to redesign the dispatch optimization manually.

Override and approval protocols

For each recommendation, the operator can approve, reject, or modify the selected dispatch intensity, duration, customer cluster, or incentive level. Safe deep reinforcement learning for active distribution networks by Kou, Liang, Wang, Wu, and Gao [27] motivates enforcing operational constraints before learned actions reach grid devices, and this article extends that safety idea to human review. Approved actions are sent to the demand response execution layer, rejected actions are

logged with reason codes, and modified actions are checked again against comfort, rebound, and network constraints. This protocol makes operator intervention auditable and prevents override decisions from bypassing the same feasibility rules used by the RL policy.

Learning from operator feedback

Operator feedback is used as a secondary learning signal that improves future explanations and dispatch recommendations. The operator-alignment reward component builds on the idea that learned energy management policies can adapt from operational experience, as demonstrated in multi-agent home energy management by Xu, Jia, Xu, Xu, Chai, and Lai [25]. Rejected and modified actions are stored in a review buffer containing state, proposed action, explanation, operator action, and reason code. During periodic retraining, these examples are used to calibrate the operator-alignment model and reduce repeated recommendations that conflict with established operating preferences.

Operator trust and usability

Operator trust is evaluated as a measurable outcome rather than assumed from improved model transparency. Explainable RL literature indicates that interpretability should be assessed by whether humans can understand, anticipate, and appropriately intervene in policy behavior [8]. In simulation-based review sessions, operators score each recommendation on clarity, plausibility, controllability, and readiness for approval. The resulting trust score is used alongside cost and peak metrics to determine whether explainability improves the operational acceptability of the RL dispatcher.

Evaluation Plan

Simulation environment and data

The evaluation uses a simulated demand response environment with realistic household load profiles, flexible HVAC loads, EV charging demand, weather-driven thermal response, dynamic prices, and feeder peak constraints. Building-to-grid flexibility modeling by Zhang, Good, and Mancarella [18] supports the use of residential heat pump aggregations, while large-scale heat-pump demand response by Müller and Jansen [22] motivates realistic response limits. EV charging behavior is simulated using uncertain departure times and state-of-charge requirements, following the control challenges highlighted by Wan, Li, He, and Prokhorov [16]. The simulator produces baseline load, dispatched load, comfort outcomes, rebound effects, and operator review records over daily episodes.

Benchmark models

The proposed explainable RL model is compared against two benchmarks: a black-box RL dispatcher and a rule-based demand response controller. The black-box RL benchmark uses the same state and action space but optimizes only a scalar reward without reward-component logging or explanation generation, reflecting the performance-oriented RL tradition in demand response reviewed by Vázquez-Canteli and Nagy [1]. The rule-based benchmark dispatches customers using fixed thresholds for price, peak forecast, and flexibility availability, resembling common operational heuristics. This comparison isolates whether the proposed framework preserves RL performance while improving interpretability and operator acceptance.

Evaluation metrics

The primary metrics are total energy cost, peak reduction, comfort violation rate, rebound magnitude, customer activation fairness, operator approval rate, and operator trust score. Cost and peak metrics follow the operational objectives used in dynamic pricing and incentive-based demand response studies by Lu, Hong, and Zhang [3] and Lu and Hong [4]. Comfort and service metrics are calculated separately for thermal, appliance, and EV-flexible clusters, reflecting the

heterogeneous constraints used in residential energy management and EV charging control [24]. Operator trust is measured on a five-point scale after each review decision and averaged across simulated dispatch events.

Expected outcomes

The expected result is that the proposed explainable RL framework achieves cost and peak-reduction performance close to black-box RL while producing higher approval and trust scores. Modified deep learning and reinforcement learning for incentive-based demand response by Wen, Zhou, Li, and Wang [13] suggests that learning-based approaches can outperform static rules, while economic dispatch RL for virtual power plants by Lin, Guan, Peng, Wang, Maharjan, and Ohtsuki [29] motivates evaluating both economic and system-level outcomes. The rule-based controller is expected to be interpretable but less adaptive during volatile price or peak events. Table 3 reports the simulated comparative performance used to evaluate dispatch effectiveness, comfort preservation, and operator acceptance.

Table 3. Comparative Simulation Performance: Dispatch Efficiency, Comfort Preservation, and Operator Acceptance

Framework	Daily energy cost reduction	Peak reduction	Comfort violation rate	Rebound magnitude	Operator approval rate	Operator trust score
Proposed explainable RL	13.8%	18.6%	2.1%	4.4%	87.3%	4.4/5
Black-box RL	14.6%	19.1%	3.8%	6.7%	61.5%	3.1/5
Rule-based controller	8.9%	11.7%	1.9%	3.8%	78.2%	3.8/5

Implications

Practical deployment

The framework is practically deployable as an advisory layer between forecasting systems and existing demand response management platforms. Online optimization results from Mocanu and colleagues [2] and real-time residential demand response results from Li, Wan, and He [6] show that learned control can operate continuously when embedded in appropriate data pipelines. In practice, the framework would require smart-meter data, customer enrollment information, device constraints, price signals, and a review interface connected to the operator's dispatch console. Its modular structure allows utilities to begin with operator-reviewed advisory recommendations before gradually enabling semi-automated dispatch for low-risk customer clusters.

Regulatory and market considerations

Regulatory acceptance of RL-based demand response depends on transparency, customer protection, and auditable decision records. Energy flexibility characterization by Li, Wang, Hong, and Piette [28] highlights that flexibility should be quantified and managed as a resource, but this resource must be dispatched without imposing hidden inconvenience on customers. The proposed framework records why customers were selected, which reward components drove the decision, and whether the operator approved or modified the recommendation. Such auditability can support market participation, customer consent verification, and compliance with reliability-oriented operating procedures.

Extensions to multi-service dispatch

The same architecture can be extended beyond peak reduction to multi-service grid-edge dispatch, including congestion management, renewable balancing, frequency support, and virtual power plant coordination. Multi-agent demand response for discrete manufacturing systems by Lu, Li, Li, Jiang, and Ding [12] and virtual power plant economic dispatch

by Lin, Guan, Peng, Wang, Maharjan, and Ohtsuki [29] show that learned controllers can coordinate heterogeneous resources across broader system objectives. In future extensions, the reward vector can include service-specific terms for voltage support, reserve provision, or emissions reduction. The operator review layer would then become even more important because competing services may produce dispatch actions with different risk and customer-impact profiles.

Conclusion

This article presented an explainable reinforcement learning framework for demand response dispatch using customer flexibility profiles, reward decomposition, and operator review. The framework treats customer flexibility as a core state and action-space input rather than as a post-processing label.

The proposed RL dispatcher learns profile-aware demand response actions under realistic operational constraints. Its decomposed reward structure separates economic, comfort, peak, rebound, flexibility-preservation, and operator-alignment objectives so that each dispatch decision can be inspected.

The explainability layer translates learned policy behavior into operator-facing rationales, and the review process enables approval, rejection, or modification before execution. This design preserves the adaptive benefits of RL while adding human oversight and operational accountability.

Future validation should move from simulation to pilot-scale field trials with real operator participation and enrolled demand response customers. Such trials would test whether explainable and reviewable RL can deliver reliable grid services while maintaining customer trust and operator confidence.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 Apr 2025

Revised: 13 May 2025

Accepted: 05 Jul 2025

Published online: 20 September 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and

distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Appl Energy*. 2019;235:1072-89.
- Mocanu E, Mocanu DC, Nguyen PH, Liotta A, Webber ME, Gibescu M, et al. On-line building energy optimization using deep reinforcement learning. *IEEE Trans Smart Grid*. 2018;10(4):3698-708.
- Lu R, Hong SH, Zhang X. A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. *Appl Energy*. 2018;220:220-30.
- Lu R, Hong SH. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. *Appl Energy*. 2019;236:937-49.
- Wang B, Li Y, Ming W, Wang S. Deep reinforcement learning method for demand response management of interruptible load. *IEEE Trans Smart Grid*. 2020;11(4):3146-55.
- Li H, Wan Z, He H. Real-time residential demand response. *IEEE Trans Smart Grid*. 2020;11(5):4144-54.
- Azuatalam D, Lee WL, De Nijs F, Liebman A. Reinforcement learning for whole-building HVAC control and demand response. *Energy AI*. 2020;2:100020.
- Heuillet A, Couthouis F, Díaz-Rodríguez N. Explainability in deep reinforcement learning. *Knowl Based Syst*. 2021;214:106685.
- Milani S, Topin N, Veloso M, Fang F. Explainable reinforcement learning: A survey and comparative review. *ACM Comput Surv*. 2024;56(7):1-36.
- Glanois C, Weng P, Zimmer M, Li D, Yang T, Hao J, et al. A survey on interpretable reinforcement learning. *Mach Learn*. 2024;113(8):5847-90.
- Kong X, Kong D, Yao J, Bai L, Xiao J. Online pricing of demand response based on long short-term memory and reinforcement learning. *Appl Energy*. 2020;271:114945.
- Lu R, Li YC, Li Y, Jiang J, Ding Y. Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management. *Appl Energy*. 2020;276:115473.
- Wen L, Zhou K, Li J, Wang S. Modified deep learning and reinforcement learning for an incentive-based demand response model. *Energy*. 2020;205:118019.
- Zhang X, Lu R, Jiang J, Hong SH, Song WS. Testbed implementation of reinforcement learning-based demand response energy management system. *Appl Energy*. 2021;297:117131.
- Pinto G, Piscitelli MS, Vázquez-Canteli JR, Nagy Z, Capozzoli A. Coordinated energy management for a cluster of buildings through deep reinforcement learning. *Energy*. 2021;229:120725.
- Wan Z, Li H, He H, Prokhorov D. Model-free real-time EV charging scheduling based on deep reinforcement learning. *IEEE Trans Smart Grid*. 2018;10(5):5246-57.
- Junker RG, Azar AG, Lopes RA, Lindberg KB, Reynders G, Relan R, et al. Characterizing the energy flexibility of buildings and districts. *Appl Energy*. 2018;225:175-82.

Zhang L, Good N, Mancarella P. Building-to-grid flexibility: Modelling and assessment metrics for residential demand response from heat pump aggregations. *Appl Energy*. 2019;233:709-23.

Pallonetto F, De Rosa M, Milano F, Finn DP. Demand response algorithms for smart-grid ready residential buildings using machine learning models. *Appl Energy*. 2019;239:1265-82.

Xiao JW, Xie Y, Fang H, Wang YW. A new deep clustering method with application to customer selection for demand response program. *Int J Electr Power Energy Syst*. 2023;150:109072.

Shirsat A, Tang W. Quantifying residential demand response potential using a mixture density recurrent neural network. *Int J Electr Power Energy Syst*. 2021;130:106853.

Müller FL, Jansen B. Large-scale demonstration of precise demand response provided by residential heat pumps. *Appl Energy*. 2019;239:836-45.

Junker RG, Kallesøe CS, Real JP, Howard B, Lopes RA, Madsen H. Stochastic nonlinear modelling and application of price-based energy flexibility. *Appl Energy*. 2020;275:115096.

Yan L, Chen X, Zhou J, Chen Y, Wen J. Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors. *IEEE Trans Smart Grid*. 2021;12(6):5124-34.

Xu X, Jia Y, Xu Y, Xu Z, Chai S, Lai CS. A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Trans Smart Grid*. 2020;11(4):3201-11.

Mathew A, Roy A, Mathew J. Intelligent residential energy management system using deep reinforcement learning. *IEEE Syst J*. 2020;14(4):5362-72.

Kou P, Liang D, Wang C, Wu Z, Gao L. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. *Appl Energy*. 2020;264:114772.

Li H, Wang Z, Hong T, Piette MA. Energy flexibility of residential buildings: A systematic review of characterization and quantification methods and applications. *Adv Appl Energy*. 2021;3:100054.

Lin L, Guan X, Peng Y, Wang N, Maharjan S, Ohtsuki T. Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy. *IEEE Internet Things J*. 2020;7(7):6288-301.