

REVIEW

Open access

A Classification Review of Explainable and Physics-Informed AI for Voltage Regulation, Congestion Management, and Microgrid Control in Cyber-Physical Smart Grids

Aisha Bello^{1*}, Ibrahim Musa², Zainab Sule¹

Abstract

Artificial intelligence is increasingly used for real-time decision support and control in cyber-physical smart grids. However, the growing reliance on data-driven control creates a need for methods that are both explainable to human operators and consistent with electrical-network physics. This review proposes a two-dimensional classification scheme for AI methods in smart-grid control. The scheme combines an explainability dimension with a physics-informed dimension and applies the resulting framework to voltage regulation, congestion management, and microgrid control. A systematic literature search identified 25 peer-reviewed journal articles published between 2017 and 2025. The selected studies were classified according to their explainability mechanism, degree of physical-knowledge integration, application domain, and relevance to cyber-physical grid operation. The review shows that most studies emphasize either explainability or physical consistency, while methods combining both remain limited. Voltage regulation is dominated by reinforcement-learning and safety-constrained control studies, whereas microgrid control contains a broader variety of AI methods. The proposed classification provides a structured basis for analysing trustworthy AI in smart-grid control. It highlights methodological gaps and identifies under-developed combinations of explainability, physics-informed learning, and operational deployment.

Keywords Voltage regulation, Congestion management, Explainable AI, Physics-informed AI, Microgrid control, Classification review

*Correspondence:

Aisha Bello

aisha.bello@gmail.com

¹ Department of Smart Power Systems, Faculty of Engineering, University of Ibadan, Ibadan, Nigeria

² Department of Energy Optimization, Faculty of Technology, Ahmadu Bello University, Zaria, Nigeria

Introduction

Smart grids are evolving into cyber-physical systems in which sensors, communication networks, distributed energy resources, and automated controllers interact continuously. This evolution has increased interest in AI methods that can support fast operational decisions under uncertainty, but it has also raised concerns about transparency and physical reliability [1]. Explainable AI addresses the interpretability problem by making model behaviour more understandable to operators, engineers, and regulators. Physics-informed AI addresses the feasibility problem by embedding grid laws, constraints, and dynamic relations into learning models [2].

Voltage regulation, congestion management, and microgrid control are three central operational problems in which AI-based methods are increasingly investigated. Deep reinforcement learning has been used for Volt-VAR control,

coordinated inverter operation, and active distribution-network management [3]. Physics-informed neural networks and related grey-box methods have been applied to power-flow approximation, state estimation, and dynamic modelling, creating new possibilities for physically consistent decision support [4]. These applications show that AI for smart-grid control must be assessed not only by predictive performance but also by explainability and physical grounding.

Existing reviews provide useful foundations but do not fully integrate explainability and physics-informed learning into one classification structure. Machlev, Heistrene, Perl, Levy, Belikov, Mannor, and Levron reviewed XAI techniques for energy and power systems, emphasizing explanation methods and stakeholder needs [1]. Huang and Wang reviewed physics-informed neural networks in power systems, focusing on how physical equations improve learning reliability [2]. A classification review is therefore needed to connect these two lines of research and evaluate how they appear across specific control problems.

The contribution of this article is a two-dimensional classification scheme for AI methods used in cyber-physical smart-grid control. The first dimension classifies explainability as post-hoc or intrinsic and as model-specific or model-agnostic, while the second dimension classifies physical integration as physics-informed losses, hard constraints, or hybrid grey-box design [5]. The framework is applied to 25 peer-reviewed studies on voltage regulation, congestion-related operational support, and microgrid control. The review does not present new experiments; its primary contribution is the classification, synthesis, and gap analysis of the selected literature.

Need for Classification

Fragmented literature

The literature on trustworthy AI for smart-grid control is fragmented across explainable AI, physics-informed learning, reinforcement learning, optimal power flow, and microgrid energy management. Reviews of interpretable machine learning in smart grids often focus on transparency and explanation interfaces, whereas physics-informed studies focus on constraints, residuals, and electrical laws [5]. Emergency-control studies using explainable deep reinforcement learning illustrate this separation because the policy may be interpreted without necessarily being formulated as a physically constrained model [6]. A classification scheme is needed to identify whether a method is explainable, physics-informed, both, or neither.

Value of a two-dimensional scheme

A two-dimensional scheme is valuable because explainability and physical consistency represent different forms of trust. Model-agnostic explanations can show which measurements influenced an AI decision, while physics-informed methods can reduce violations of power-flow equations, voltage limits, or dynamic constraints [1]. Interpretable dispatching-control studies demonstrate the value of explaining AI recommendations, whereas safe voltage-control studies show the importance of preventing infeasible control actions [7]. Treating the two dimensions separately avoids the false assumption that an explainable model is physically safe or that a physics-informed model is automatically understandable.

Figure 1 presents the conceptual trust architecture used in this review, separating explainability from physical consistency while showing how both dimensions jointly support operator-relevant smart-grid control.

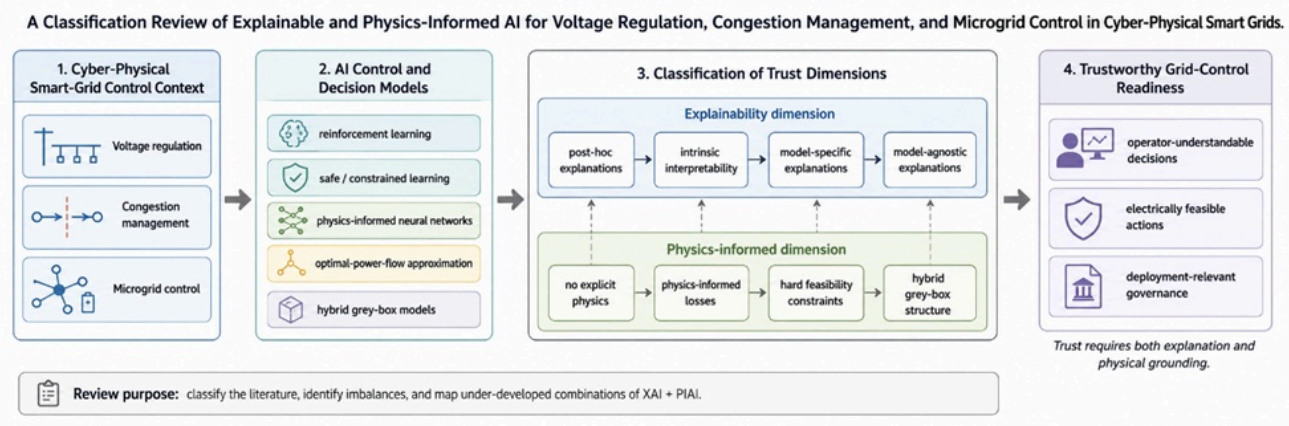


Figure 1. Two-Dimensional Trust Architecture for Explainable and Physics-Informed AI in Cyber-Physical Smart-Grid Control

Review Method

Database and search strategy

The literature search targeted peer-reviewed journal articles indexed in IEEE Xplore, ScienceDirect, Scopus, Web of Science, and publisher databases for major power and energy journals. Search terms combined “explainable AI,” “XAI,” “physics-informed,” “physics-informed neural network,” “voltage regulation,” “congestion management,” “microgrid control,” and “smart grid control,” following the thematic scope established in XAI and PIAI review studies [1]. The search prioritized IEEE Transactions on Smart Grid, IEEE Transactions on Power Systems, Applied Energy, Energy and AI, Electric Power Systems Research, International Journal of Electrical Power and Energy Systems, Renewable and Sustainable Energy Reviews, and IEEE Access. Studies were retained when their method could be classified along at least one explainability or physics-informed dimension.

Inclusion/exclusion criteria

The inclusion criteria required articles to be peer-reviewed journal publications from 2017 to 2025 and to address explainable AI, physics-informed AI, or AI-based smart-grid control. Studies on physics-informed state estimation, AC optimal power flow, transformer thermal modelling, dynamic simulation, and inverter-dominated power systems were included when they contributed transferable categories for control-oriented classification [8]. Articles were excluded if they were conference papers, preprints, non-peer-reviewed material, outside the time window, or focused only on generic prediction without a relevant control or cyber-physical smart-grid context. Review articles were included only when they contributed taxonomy, terminology, or classification logic rather than merely listing AI applications.

Screening and final selection

The screening process moved from broad keyword retrieval to classification relevance. Physics-informed studies on optimal power flow and operational support were retained because they clearly demonstrate how grid equations, gradients, and feasibility criteria can be embedded in learning methods [4]. Reinforcement-learning studies for voltage control were retained because they represent AI-based control under physical operating limits and cyber-physical constraints [9]. Table 1 summarizes the search and screening process used to identify the final set of 25 reviewed articles.

Table 1. Literature Search and Screening Summary

Database or source	Main search focus	Initial hits	Screened after title and abstract	Final articles retained
IEEE Xplore	Explainable AI, XAI, smart-grid control, and power-system decision support	42	14	4
IEEE Xplore	Deep reinforcement learning, voltage control, Volt-VAR control, and active distribution networks	87	22	8
ScienceDirect	Physics-informed neural networks, power-system control, optimal power flow, and dynamic modelling	64	18	7
ScienceDirect	Microgrid control, safe reinforcement learning, islanding, and energy management	38	11	3
Scopus and Web of Science	Classification review, interpretable machine learning, energy systems, and smart grids	55	16	2
Publisher and DOI cross-checking	Backward and forward checking from selected XAI and PIAI papers	31	9	1
Total	Combined targeted search and verification	317	90	25

Explainable AI Categories

Post-hoc vs. intrinsic explainability

Post-hoc explainability refers to methods that explain a trained model after learning, while intrinsic explainability refers to models that are interpretable by design. SHAP, LIME, attention maps, saliency methods, and feature-attribution techniques are typical post-hoc approaches because they translate complex model behaviour into more understandable evidence for users [1]. Intrinsic explainability appears when the model structure itself is transparent, such as rule-based logic, sparse models, monotonic relations, or architectures with physically meaningful variables [5]. In smart-grid control, this distinction matters because post-hoc explanations support auditing, whereas intrinsic explanations may be more suitable for real-time operator trust.

Model-specific vs. model-agnostic

Model-specific explanations depend on the internal structure of a particular model class, while model-agnostic explanations can be applied across different learning architectures. Explainable deep reinforcement learning for emergency control is model-specific when the explanation is tied to the policy architecture, state-action structure, or learned dispatch hierarchy [6]. Model-agnostic approaches can compare neural networks, tree-based models, and reinforcement-learning policies using common input-output explanation tools [1]. This classification is important because cyber-physical smart grids may contain several AI models that require consistent explanation interfaces across different control layers.

Explanation types and stakeholders

Explanation types can be classified according to the stakeholder question being answered. Operators usually need decision explanations, engineers need diagnostic explanations, and regulators need auditable explanations that show whether a model behaved consistently with safety and reliability requirements [7]. In state-estimation and topology-aware learning, explanations may also need to connect model behaviour to network structure and measurement placement [10].

Therefore, this review treats explainability as both a technical property of the AI method and a functional property of the explanation delivered to smart-grid stakeholders.

Explainability in power systems context

Explainability in power systems must be judged against electrical meaning rather than only statistical clarity. A feature-importance ranking is operationally useful only when it can be interpreted in relation to voltage magnitude, reactive power capability, line loading, network topology, or dynamic stability [5]. Physics-informed graphical neural networks show how topology-aware learning can make internal model representations more consistent with grid structure, even when the main task is state estimation rather than direct control [10]. In this review, explainability is therefore classified by both algorithmic form and operational relevance.

Physics-Informed AI Categories

Physics-informed loss functions

Physics-informed loss functions incorporate physical laws, residual equations, or operational penalties into the training objective. In power systems, this may include power-flow residuals, dynamic-equation residuals, thermal-model residuals, or gradient information derived from known physical relations [2]. Gradient-enhanced physics-informed neural networks demonstrate how derivative information can improve operational support when limited data are available [11]. This category is classified as soft physical integration because the model is encouraged to satisfy physics but is not always forced to do so exactly.

Hard physical constraints

Hard physical constraints impose feasibility directly through constrained action spaces, safety shields, projection layers, optimization layers, or rule-based limits. Safe off-policy deep reinforcement learning for Volt-VAR control illustrates this category because the controller is designed to avoid voltage violations while learning operational actions [9]. Physics-shielded multi-agent reinforcement learning further demonstrates how safety mechanisms can protect voltage control involving photovoltaic and battery energy-storage systems [12]. This category is more restrictive than physics-informed losses because operational feasibility becomes a boundary condition for control.

Hybrid grey-box models

Hybrid grey-box models combine data-driven learning with explicit physical models, simulators, graph structures, or reduced-order equations. Physics-informed transformer thermal modelling is a representative example because the neural model is guided by known thermal behaviour rather than relying only on statistical correlation [8]. Bayesian physics-informed system identification for inverter-dominated power systems also illustrates grey-box design by combining physical structure with uncertainty-aware learning [13]. These models are important for cyber-physical smart grids because they can preserve engineering meaning while adapting to data-rich and changing operating conditions.

Levels of physical integration

The physics-informed dimension can be organized into four levels: no explicit physics, soft physics through loss functions, hard physics through constraints, and hybrid physics through coupled model-learning structures. AC optimal-power-flow PINN studies show that the same application can fall into different physical-integration levels depending on whether equations are penalized, embedded, or combined with optimization structure [4]. PINNSim and trustworthy physics-informed dynamic frameworks demonstrate that dynamic simulation can also be learning-based while remaining connected to physical equations and reliability requirements [14]. **Table 2** presents the two-dimensional classification scheme used in the remainder of the review.

Table 2. Two-Dimensional Classification of XAI and PIAI Methods

Explainability dimension	No explicit physics	Physics-informed loss functions	Hard physical constraints	Hybrid grey-box models
Post-hoc, model-agnostic	SHAP, LIME, or permutation explanations applied to a black-box voltage or microgrid controller; explanation is portable but physical validity is not guaranteed.	SHAP or sensitivity analysis applied to a PINN trained with power-flow, thermal, or dynamic residual penalties; explanation supports auditing of soft physical consistency.	Attribution analysis of a controller protected by action projection, safety filters, or rule-based feasibility checks; explanation describes decisions after constraint enforcement.	Post-hoc explanation of a neural model coupled with a simulator, graph model, or reduced-order physical representation; explanation can be interpreted through physical variables.
Post-hoc, model-specific	Attention maps, saliency maps, or policy-trajectory explanations for a deep reinforcement-learning controller; explanation depends on the learned model structure.	Residual diagnostics and gradient-based interpretation for a physics-informed neural architecture; explanation is tied to the structure of the residual terms.	Policy-path or action-filter explanations for constrained reinforcement learning; explanation follows both the learned policy and the embedded operational limits.	Internal-state visualization of Bayesian, dynamic, or topology-aware grey-box models; explanation depends on learned physical states and uncertainty estimates.
Intrinsic, model-agnostic	Sparse surrogate rules or interpretable regression models approximating a controller; interpretability is high but physical validity must be separately checked.	Transparent additive or rule-based models trained with equation-residual penalties; interpretability and soft physical consistency are both present.	Rule-based or monotonic controllers with explicit voltage, current, power, or stability limits; feasibility and interpretability are built into the decision logic.	Structured graphical or reduced-order models with interpretable physical variables; learning updates parameters while the physical model form remains understandable.
Intrinsic, model-specific	Interpretable neural architectures with named latent variables but no explicit electrical equations; useful for insight but limited for feasibility assurance.	Neural architectures whose loss terms correspond directly to power-flow, thermal, or dynamic equations; explanation comes from the physical meaning of the residuals.	Neural controllers with embedded feasibility layers or differentiable optimization blocks; the architecture itself prevents selected classes of violations.	End-to-end grey-box architectures combining neural components, grid equations, topology, and uncertainty; this is the strongest but most complex classification cell.

Voltage Regulation Applications

AI methods for voltage control

AI-based voltage regulation is one of the most populated application areas in the reviewed literature because distribution networks increasingly require fast control of inverters, reactive power devices, and energy-storage resources. Two-timescale deep reinforcement learning has been used to coordinate slower mechanical devices with faster inverter-based

actions, showing how learning can support operational decisions across different control horizons [3]. Deep reinforcement learning for Volt-VAR optimization further demonstrates that AI can learn voltage-support policies from system states and control feedback rather than relying only on fixed rule-based settings [15]. Multi-agent approaches extend this idea by distributing the control task across several local controllers that must coordinate under network coupling [16].

Explainability and physics in voltage studies

Voltage-regulation studies show strong attention to operational feasibility, but their explainability level varies considerably. Safe off-policy reinforcement learning explicitly addresses voltage-limit violations, placing it in the hard-constraint region of the physics-informed dimension even when the learned policy itself remains difficult to interpret [9]. Two-stage deep reinforcement learning for inverter-based Volt-VAR control introduces a structured control process that improves operational organization, but the explanation of why a specific action is selected remains less developed than the control mechanism itself [17]. Physics-shielded multi-agent reinforcement learning strengthens the physical-integration dimension by filtering unsafe actions before they affect photovoltaic and battery-supported voltage control [12].

Gaps and opportunities

The main gap in voltage regulation is not the absence of AI control methods, but the limited integration of operator-facing explanations with physically constrained learning. Decentralized voltage control using multi-agent deep reinforcement learning demonstrates scalability for high photovoltaic penetration, yet the decision logic still requires clearer explanation for field deployment [18]. Coordinated active and reactive power optimization shows that multi-agent learning can manage complex control interactions, but the resulting policies need more transparent justification for operators responsible for voltage-security decisions [19]. Human-in-the-loop reinforcement learning is a promising direction because it connects automated Volt-VAR control to operator judgement, creating a pathway toward explanations that are useful during real-time intervention [20].

Congestion Applications

AI for congestion management

Congestion management is less directly represented than voltage regulation in the selected literature, but it is strongly connected to optimal power flow, dispatching control, and operational support. Data-driven optimal power flow provides a physics-informed machine-learning route for congestion-related decision support because line-flow and feasibility constraints are central to OPF formulations [21]. Physics-informed neural networks for AC optimal power flow similarly show how learning can approximate power-flow-constrained decisions that are relevant to congestion and security management [4]. In this review, congestion applications are therefore interpreted broadly as AI methods supporting constrained network operation, line-loading relief, and feasible dispatch.

Explainability and Physics in Congestion Studies

Congestion-oriented AI studies tend to be more physics-informed than explainable because feasibility, line limits, and power-flow consistency are built into the operational problem. Gradient-enhanced physics-informed neural networks support power-system operation by using derivative information, which can improve sensitivity-aware decision support for constrained grid operation [11]. Physics-informed graphical neural networks for state estimation contribute indirectly to congestion management by improving topology-aware situational awareness, which is necessary before congestion-relief actions can be trusted [10]. Explainable emergency dispatching and multi-hierarchical interpretable dispatching provide a complementary direction by focusing on why control actions are recommended, but these studies are not always formulated specifically as congestion-management tools [6].

Gaps and opportunities

The largest gap in congestion management is the scarcity of studies that combine explicit congestion-relief objectives, explainable decision logic, and hard physical feasibility guarantees. Physics-informed optimal-power-flow studies provide strong physical grounding, but they often emphasize solution quality and feasibility rather than operator-oriented explanation of congestion drivers [21]. Interpretable dispatching-control studies explain policy behaviour, but their connection to line-overload mitigation and congestion-cost allocation is less explicit [7]. Future congestion-management research should therefore connect counterfactual explanations, constraint-violation diagnostics, and OPF-based physical structure in a single classification cell.

Microgrid Control Applications

Primary, secondary, and tertiary control

Microgrid control spans primary stabilization, secondary restoration, and tertiary energy-management decisions, which makes it more diverse than single-objective voltage regulation. Multi-agent safe policy learning for networked microgrids demonstrates how distributed controllers can coordinate power management while respecting operational safety requirements [22]. Safe deep reinforcement learning for microgrid energy management extends this perspective by incorporating spatial-temporal perception, showing that microgrid decisions depend on both local dynamics and network-wide context [23]. These studies occupy a physically constrained but only partially explainable region of the classification because safety is emphasized more strongly than interpretability.

Islanding and resilience

Microgrids require AI methods that remain reliable during islanding, reconnection, disturbances, and uncertain renewable generation. Physics-informed dynamic simulation frameworks are relevant here because islanded operation depends on frequency, voltage, and inverter-dominated dynamics that cannot be assessed through data correlation alone [14]. Bayesian physics-informed identification of inverter-dominated power systems adds uncertainty awareness, which is important when microgrid controllers must act under incomplete measurements or changing topology [13]. A classification perspective shows that resilience-oriented microgrid AI should be evaluated by whether it explains abnormal conditions and whether it preserves physical consistency during transitions.

Explainable and physics-informed approaches

The reviewed microgrid literature includes emerging examples of physics-informed learning, but explainability remains less mature than safety and feasibility. A uniform physics-informed neural-network framework for microgrids directly addresses voltage-stability analysis, placing it within the physics-informed loss or hybrid grey-box categories depending on implementation details [24]. Trustworthy physics-informed machine learning for power-system dynamics provides a broader framework for incorporating reliability and trustworthiness into dynamic models, which is highly relevant to microgrid operation [25]. However, most microgrid studies still need clearer mechanisms for translating learned behaviour into explanations that operators can use during islanding, restoration, or energy-management decisions.

Cross-application synthesis

Across the three application domains, voltage regulation has the densest set of AI control studies, congestion management has the strongest connection to OPF-based physics, and microgrid control has the broadest cyber-physical scope. Reviews of XAI and interpretable machine learning show that explanation techniques are available, but they are not yet systematically embedded into voltage, congestion, and microgrid controllers [1]. Reviews and applications of physics-informed learning show stronger progress in physical consistency, especially through PINNs, OPF formulations, dynamic models, and grey-box identification [2]. **Table 3** maps the reviewed methods to the three operational domains and identifies the most populated classification cells.

Table 3. Mapping of XAI/PIAI Methods to Smart-Grid Control Applications

Application domain	Representative reviewed studies	Dominant AI methods	Dominant explainability category	Dominant physics-informed category	Most populated classification cells	Under-populated classification cells
Voltage regulation	Two-timescale voltage control [3], Volt-VAR optimization [15], safe off-policy Volt-VAR control [9], two-stage inverter Volt-VAR control [17], consensus multi-agent Volt-VAR control [16], decentralized PV-rich voltage control [18], coordinated active/reactive voltage control [19], physics-shielded voltage control [12], human-in-the-loop Volt-VAR control [20]	Deep reinforcement learning, multi-agent reinforcement learning, safe reinforcement learning, human-in-the-loop reinforcement learning	Mostly post-hoc or weakly intrinsic; explanations are limited unless human-in-the-loop or structured-policy elements are present	Hard constraints and safety shields are common; some studies use structured control stages rather than explicit physical residuals	Post-hoc/model-specific plus hard constraints; intrinsic/model-specific plus constrained control	Intrinsic/model-agnostic plus hybrid grey-box; post-hoc/model-agnostic plus physics-informed loss
Congestion management	Data-driven optimal power flow [21], PINN-based AC optimal power flow [4], gradient-enhanced operational support [11], interpretable dispatching control [7], explainable emergency control [6]	Physics-informed learning, optimal-power-flow approximation, interpretable dispatching, reinforcement-learning-based emergency control	Model-specific explanations appear in dispatching; model-agnostic explanations remain limited	Physics-informed losses and OPF-based constraints dominate	Post-hoc/model-specific plus physics-informed loss; intrinsic/model-specific plus hybrid OPF structure	Intrinsic/model-agnostic plus hard constraints; post-hoc/model-agnostic plus hybrid grey-box explanations
Microgrid control	Safe policy learning for networked microgrids [22], safe deep reinforcement learning for microgrid energy management [23], uniform PINN	Safe reinforcement learning, multi-agent learning, PINNs, Bayesian grey-box modelling, dynamic physics-	Explainability is mostly implicit or model-specific; stakeholder-facing explanations are rare	Hard constraints, physics-informed losses, and hybrid grey-box models all appear	Post-hoc/model-specific plus hard constraints; intrinsic/model-specific plus physics-informed loss;	Model-agnostic explanation plus hybrid grey-box; intrinsic transparent microgrid controllers with explicit physical guarantees

	framework for microgrid voltage stability [24], Bayesian physics-informed inverter identification [13], trustworthy physics-informed dynamics [25]	informed learning			hybrid grey-box cells	
Cross-domain pattern	XAI reviews [1], interpretable smart-grid review [5], PINN review [2], topology-aware state estimation [10], transformer thermal modelling [8], PINN-based simulation [14]	Review taxonomies, topology-aware neural networks, PINNs, grey-box thermal and dynamic models	General XAI categories are mature, but application-specific explanations are uneven	Physical integration is mature in modelling and support tasks, less mature in explainable closed-loop control	Physics-informed loss cells and hard-constraint cells	Fully explainable and fully physics-informed closed-loop control across all three applications

Proposed Classification

The two-dimensional framework

The proposed framework classifies each AI method by crossing an explainability dimension with a physics-informed dimension. The explainability dimension separates post-hoc from intrinsic methods and model-specific from model-agnostic methods, following categories used in XAI reviews for power and energy systems [1]. The physics-informed dimension separates no explicit physics, physics-informed loss functions, hard physical constraints, and hybrid grey-box models, following categories visible in PINN, OPF, and dynamic-modelling studies [2]. This two-dimensional structure makes it possible to classify a method without conflating transparency with feasibility.

Classification of reviewed studies

When the 25 reviewed studies are classified, voltage-regulation papers cluster around reinforcement-learning control with hard or semi-hard safety mechanisms. Studies such as safe off-policy Volt-VAR control and physics-shielded multi-agent voltage control occupy physically constrained cells, while human-in-the-loop voltage control moves closer to explainable control through operator interaction [9]. Congestion-related studies cluster around OPF and physics-informed operational support, particularly data-driven OPF and AC-OPF PINN formulations [4]. Microgrid studies are distributed more broadly across safe reinforcement learning, physics-informed voltage-stability analysis, and grey-box dynamic identification [22].

Observed patterns and imbalances

The first imbalance is that physics-informed modelling is more developed than explainable physics-informed control. Transformer thermal modelling, inverter-dominated system identification, and PINN-based simulation show strong physical integration, but they do not always provide operator-facing explanations for control decisions [8]. The second imbalance is that explainability studies often focus on interpreting black-box behaviour without enforcing hard grid constraints, as seen in the distinction between XAI reviews and safety-constrained voltage-control studies [5]. The third imbalance is application-specific: congestion management has physically grounded methods but relatively few explicit XAI mechanisms, while microgrid control has rich control diversity but uneven classification maturity.

Classification as a research roadmap

The classification can be used as a research roadmap by identifying cells that are technically important but weakly populated. For voltage regulation, the priority is to combine safe reinforcement learning with explanations that reveal why inverters, storage units, or tap changers are activated under specific grid conditions [20]. For congestion management, the priority is to combine OPF-based physical structure with counterfactual and constraint-aware explanations that identify the drivers of line overloads [21]. For microgrid control, the priority is to integrate physics-informed dynamic models with explanations suitable for islanding, resilience, and operator-supervised autonomy [24].

Figure 2 synthesizes the review's roadmap by showing how voltage regulation, congestion management, and microgrid control require different combinations of explainability, physical feasibility, and operator-supervised deployment.

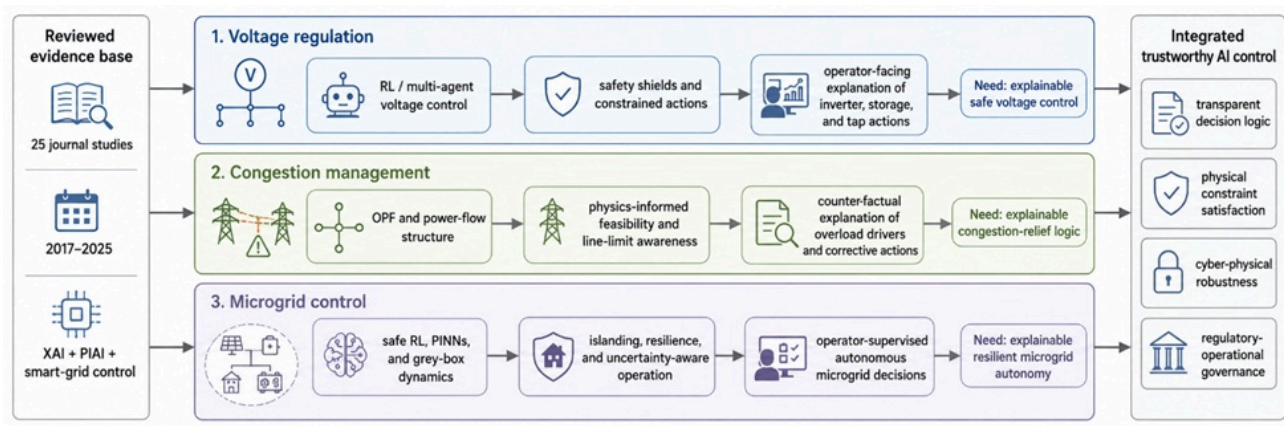


Figure 2. Cross-Domain Research Roadmap for Explainable and Physics-Informed AI in Voltage Regulation, Congestion Management, and Microgrid Control

Conclusion

This review proposed a two-dimensional classification scheme for AI methods used in cyber-physical smart-grid control. The first dimension captures how a method explains its behaviour, while the second captures how it incorporates physical knowledge. By separating these dimensions, the review avoids treating explainability and physical consistency as interchangeable forms of trust.

The classification shows that voltage regulation is the most densely studied control application among the reviewed works. Many voltage-control studies use reinforcement learning, multi-agent coordination, or safety filtering, but fewer provide strong operator-facing explanations. This makes voltage regulation a mature but still incomplete area for trustworthy AI.

Congestion management appears less populated as an explicitly labelled application, although it is strongly connected to optimal power flow, dispatching control, and operational support. The reviewed studies show that physics-informed methods are promising for congestion-related feasibility, but explainable congestion-relief logic remains underdeveloped. This gap is important because operators need to understand not only that a solution is feasible, but also why a constraint is binding and which corrective action is justified.

Microgrid control shows the broadest cyber-physical scope because it includes energy management, islanding, resilience, and dynamic stability. The main future direction is to combine physics-informed dynamic models, safe control, and

stakeholder-oriented explanations in a unified design. Overall, the proposed classification provides a roadmap for developing AI systems that are transparent, physically grounded, and suitable for critical smart-grid operation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 05 Oct 2024

Revised: 11 Nov 2024

Accepted: 07 Mar 2025

Published online: 20 March 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

- Machlev R, Heistrene L, Perl M, Levy KY, Belikov J, Mannor S, et al. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy AI*. 2022;9:100169.
- Huang B, Wang J. Applications of physics-informed neural networks in power systems-a review. *IEEE Trans Power Syst*. 2022;38(1):572-88.
- Yang Q, Wang G, Sadeghi A, Giannakis GB, Sun J. Two-timescale voltage control in distribution grids using deep reinforcement learning. *IEEE Trans Smart Grid*. 2019;11(3):2313-23.
- Nellikath R, Chatzivasileiadis S. Physics-informed neural networks for ac optimal power flow. *Electr Power Syst Res*. 2022;212:108412.
- Xu C, Liao Z, Li C, Zhou X, Xie R. Review on interpretable machine learning in smart grid. *Energies*. 2022;15(12):4427.

Zhang K, Zhang J, Xu PD, Gao T, Gao DW. Explainable AI in deep reinforcement learning models for power system emergency control. *IEEE Trans Comput Soc Syst*. 2021;9(2):419-27.

Zhang K, Zhang J, Xu P, Gao T, Gao W. A multi-hierarchical interpretable method for DRL-based dispatching control in power systems. *Int J Electr Power Energy Syst*. 2023;152:109240.

Bragone F, Morozovska K, Hilber P, Laneryd T, Luvisotto M. Physics-informed neural networks for modelling power transformer's dynamic thermal behaviour. *Electr Power Syst Res*. 2022;211:108447.

Wang W, Yu N, Gao Y, Shi J. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. *IEEE Trans Smart Grid*. 2019;11(4):3008-18.

Ngo QH, Nguyen BL, Vu TV, Zhang J, Ngo T. Physics-informed graphical neural network for power system state estimation. *Appl Energy*. 2024;358:122602.

Mohammadian M, Baker K, Fioretto F. Gradient-enhanced physics-informed neural networks for power systems operational support. *Electr Power Syst Res*. 2023;223:109551.

Chen P, Liu S, Wang X, Kamwa I. Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems. *IEEE Trans Smart Grid*. 2022;14(4):2656-67.

Stock S, Babazadeh D, Becker C, Chatzivasileiadis S. Bayesian physics-informed neural networks for system identification of inverter-dominated power systems. *Electr Power Syst Res*. 2024;235:110860.

Stiasny J, Zhang B, Chatzivasileiadis S. Pinnsim: A simulator for power system dynamics based on physics-informed neural networks. *Electr Power Syst Res*. 2024;235:110796.

Zhang Y, Wang X, Wang J, Zhang Y. Deep reinforcement learning based volt-var optimization in smart distribution systems. *IEEE Trans Smart Grid*. 2020;12(1):361-71.

Gao Y, Wang W, Yu N. Consensus multi-agent reinforcement learning for volt-var control in power distribution networks. *IEEE Trans Smart Grid*. 2021;12(4):3594-604.

Liu H, Wu W. Two-stage deep reinforcement learning for inverter-based volt-var control in active distribution networks. *IEEE Trans Smart Grid*. 2020;12(3):2037-47.

Cao D, Zhao J, Hu W, Ding F, Huang Q, Chen Z, et al. Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of PVs. *IEEE Trans Smart Grid*. 2021;12(5):4137-50.

Hu D, Ye Z, Gao Y, Ye Z, Peng Y, Yu N. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. *IEEE Trans Smart Grid*. 2022;13(6):4873-86.

Sun X, Xu Z, Qiu J, Liu H, Wu H, Tao Y. Optimal volt/var control for unbalanced distribution networks with human-in-the-loop deep reinforcement learning. *IEEE Trans Smart Grid*. 2023;15(3):2639-51.

Lei X, Yang Z, Yu J, Zhao J, Gao Q, Yu H. Data-driven optimal power flow: A physics-informed machine learning approach. *IEEE Trans Power Syst*. 2020;36(1):346-54.

Zhang Q, Dehghanpour K, Wang Z, Qiu F, Zhao D. Multi-agent safe policy learning for power management of networked microgrids. *IEEE Trans Smart Grid*. 2020;12(2):1048-62.

Ye Y, Wang H, Chen P, Tang Y, Strbac G. Safe deep reinforcement learning for microgrid energy management in distribution networks with leveraged spatial-temporal perception. *IEEE Trans Smart Grid*. 2023;14(5):3759-75.

Feng R, Wajid K, Faheem M, Wang J, Subhan FE, Bhutta MS. Uniform physics informed neural network framework for microgrid and its application in voltage stability analysis. *IEEE Access*. 2025;13:8110-26.

Ellinas P, Karampinis I, Nadal IV, Nellikkath R, Vorwerk J, Chatzivasileiadis S. Physics-informed machine learning for power system dynamics: A framework incorporating trustworthiness. *Sustain Energy Grids Netw*. 2025;43:101818.
<https://doi.org/10.1016/j.segan.2025.101818>.