

ORIGINAL RESEARCH

Open access

A Layered Governance Architecture for Autonomous Distribution Grids Integrating Explainable Control, Audit Logs, Human Override, and Safety Constraints

Ji-eun Choi^{1*}, Hyejin Park¹, Min-seo Kang², Seung-ho Lee¹

Abstract

Autonomous control is transforming distribution grids, but governance mechanisms have not kept pace. Without built-in explainability, traceability, human oversight, and safety limits, autonomous systems risk becoming opaque and unaccountable. This article proposes a layered governance architecture that integrates explainable control, audit logs, human override, and safety constraints into a unified framework for autonomous distribution grids. The objective is to clarify how trustworthy autonomy can be designed before deployment rather than repaired after failure. The architecture is designed conceptually, with four interacting layers that address distinct governance needs. Implementation scenarios and a validation checklist are used to illustrate how the framework can guide planning, procurement, and regulatory review. The architecture defines clear roles and interfaces for each layer, ensuring that autonomous decisions remain transparent, auditable, overridable, and safe. Two tables summarize the layer responsibilities and provide a governance assessment checklist for future use. The proposed architecture offers a blueprint for building trustworthy autonomous distribution grids that meet emerging regulatory and societal expectations. It emphasizes governance as a design requirement rather than an external compliance burden.

Keywords Autonomous distribution grid, Layered governance, Explainable control, Audit logs, Human override, Safety constraints

*Correspondence:

Ji-eun Choi

jieun.choi@gmail.com

¹ Department of Intelligent Power Systems, College of Engineering, Yonsei University, Seoul, South Korea

² Department of Energy Analytics and Control, Faculty of Engineering, Pusan National University, Busan, South Korea

Introduction

Autonomous distribution grids are emerging as a response to increasingly complex operating conditions created by distributed energy resources, digital substations, active consumers, and rapid fault dynamics. Recent reviews of agentic artificial intelligence and smart-grid autonomy show that grid controllers are moving from advisory analytics toward systems capable of initiating actions with limited human involvement [1]. This shift can improve responsiveness, but it also changes the moral and institutional status of control decisions because actions are no longer simply executed by operators after deliberation. The governance challenge is therefore not only how to make autonomous controllers technically competent, but how to make their decisions understandable, reviewable, contestable, and bounded.

Distribution-grid autonomy is particularly consequential because decisions are made close to customers, critical loads, and distributed assets that may respond unpredictably to local conditions. Deep reinforcement learning and related

control methods have been proposed for smart-grid operation because they can adapt to changing states and learn complex operational policies [2]. Yet such methods often produce decisions whose internal rationale is difficult for operators and regulators to inspect. A governance architecture must therefore surround the autonomous controller with mechanisms that translate machine action into institutional accountability.

Existing work on artificial intelligence in smart grids has emphasized optimization, stability, forecasting, and control, while governance requirements have often been treated as secondary or external to system design [3]. This separation is risky because autonomous control decisions can trigger switching, curtailment, voltage regulation, and restoration actions before human actors have time to reconstruct the underlying reasoning. In a safety-critical infrastructure, accountability cannot depend on after-the-fact interpretation alone. It must be embedded into the architecture through explainability, auditability, override authority, and safety constraints.

This article proposes a layered governance architecture in which four governance layers are wrapped around autonomous distribution-grid controllers. The architecture builds on research showing that reinforcement-learning and predictive-control methods are becoming operationally relevant in modern power and energy systems [4]. It extends that technical trajectory by asking what governance functions must accompany such controllers if they are to operate legitimately in public infrastructure. The resulting design is conceptual and architectural rather than experimental, with emphasis on responsibilities, interfaces, and decision hierarchy.

The article contributes a framework that links technical control design with institutional oversight. Predictive control of distributed energy resources illustrates how autonomous coordination can operate across many devices, while digital-twin approaches show how virtual representations may support monitoring and decision support [5, 6]. These advances create the conditions for more capable autonomy, but they also heighten the need for transparent explanations, tamper-evident records, meaningful operator intervention, and non-negotiable safety boundaries. The remainder of the article develops these requirements and organizes them into a layered governance architecture for autonomous distribution grids.

Autonomous Distribution Grid Problem

What makes a grid autonomous

A distribution grid becomes autonomous when sensing, interpretation, decision-making, and actuation are combined so that the system can initiate operational changes without waiting for direct human command. Surveys of artificial intelligence in smart grids describe this transition as a movement from passive automation toward adaptive control capable of learning from system states and responding to disturbances [7]. In this article, autonomy does not mean independence from institutions, rules, or operators. It means that the controller has delegated authority to select and execute actions within a defined operational envelope.

Operational benefits and risks

Autonomous control can improve the speed of fault isolation, service restoration, voltage regulation, and distributed resource coordination because decisions can be made faster than conventional manual workflows. Reviews of secure smart-grid automation show that artificial intelligence and blockchain are increasingly considered together because automated action raises simultaneous concerns about performance, cyber integrity, and accountability [8]. The same autonomy that enables rapid correction can also propagate mistaken actions if the controller misreads the state, applies an unsuitable policy, or operates outside the assumptions under which it was designed. Governance must therefore treat operational benefit and operational risk as inseparable.

The governance vacuum

The governance vacuum arises when autonomous controllers are deployed with technical control objectives but without equally explicit rules for explanation, logging, intervention, and safety precedence. Research on explainable artificial intelligence in energy and power systems shows that transparency is now recognized as a condition for trust, especially where operators must rely on AI-generated recommendations or actions [9]. However, explainability alone cannot establish accountability if records are incomplete, override rights are unclear, or safety limits can be bypassed under pressure. A complete architecture must define how governance functions interact before, during, and after each autonomous decision.

Governance Requirements

Transparency and explainability

Transparency requires that the autonomous controller provide reasons for its actions in forms that are intelligible to human stakeholders. General research on explainable artificial intelligence emphasizes that trustworthy AI depends not only on technical interpretability but also on explanations that answer the practical questions of affected users [10]. In distribution-grid control, this means explaining why a feeder was reconfigured, why a voltage-control action was selected, or why a resource was curtailed. Explanations must connect the observed grid condition, the selected action, the rejected alternatives, and the expected safety implications.

Traceability and auditability

Traceability requires that every relevant input, recommendation, decision, command, override, and safety intervention be recorded in a durable sequence. Blockchain-enabled smart-grid research demonstrates the value of secure data-sharing and auditable records where multiple actors exchange operationally sensitive information [11]. For autonomous grids, auditability is not merely a cybersecurity feature; it is a governance function that allows post-event reconstruction and regulatory review. The architecture must therefore make logging an integral companion to control, rather than a passive database added after deployment.

Human authority and override

Human authority requires that operators retain meaningful capacity to veto, suspend, modify, or escalate autonomous actions under defined conditions. The growth of human-in-the-loop and human-on-the-loop approaches in automated systems indicates that oversight must be designed around real decision contexts rather than symbolic human presence [12]. In distribution-grid operations, a human override mechanism must specify who may intervene, what information must be displayed, which actions are reversible, and how conflicts between operator judgment and machine recommendation are resolved. Without such provisions, autonomy may undermine rather than support responsible operation.

Inviolable safety limits

Inviolable safety limits define the physical and operational boundaries that neither an autonomous controller nor an operator may intentionally exceed. Safety-critical control research using barrier-function approaches shows that some constraints must be enforced as structural safeguards rather than advisory preferences [13]. In distribution grids, such limits may concern equipment ratings, protection coordination, voltage boundaries, switching permissions, and emergency operating envelopes. The governance architecture therefore treats safety constraints as the highest-priority layer in the decision hierarchy.

Figure 1 presents the proposed layered governance architecture, showing how explainable control, audit logging, human override, and non-bypassable safety constraints operate around autonomous distribution-grid controllers to make machine-initiated actions transparent, auditable, contestable, and physically bounded.

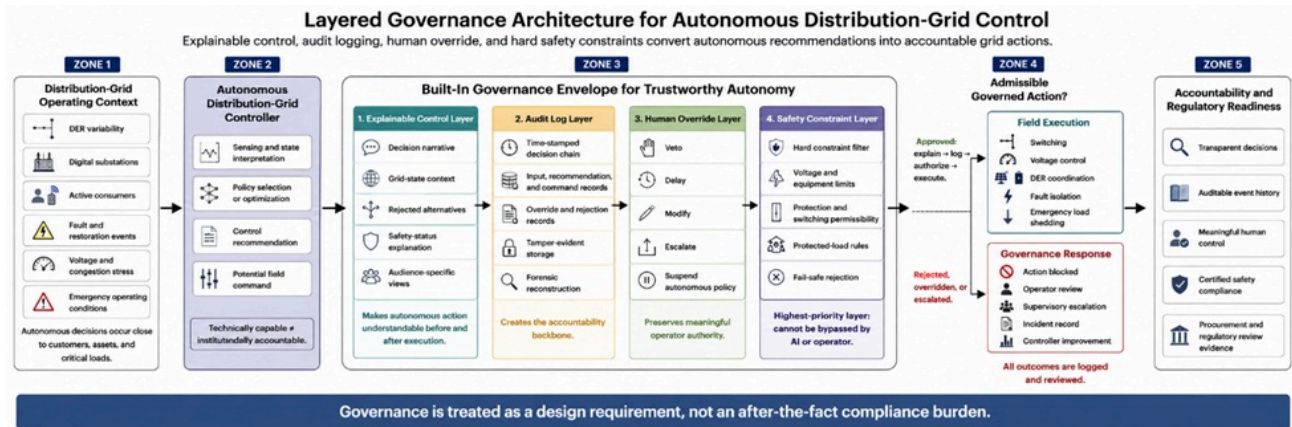


Figure 1. Layered Governance Architecture for Trustworthy Autonomous Distribution-Grid Control

Explainable Control Layer

Function and scope

The Explainable Control Layer converts autonomous control decisions into structured explanations that support operator understanding, regulatory review, and public accountability. Explainable AI research in energy and power systems emphasizes that explanations should be embedded in operational workflows rather than generated only after failures [9]. In the proposed architecture, each autonomous action is accompanied by a reason statement, a context statement, a confidence or uncertainty narrative expressed in qualitative terms, and an indication of safety checks performed. The layer does not replace the controller; it governs how the controller's reasoning is communicated.

Explanation generation methods

Explanation generation may use model-based summaries, feature-attribution narratives, counterfactual descriptions, policy rationales, and rule-based translations of control logic. Broader explainable artificial intelligence literature argues that explanations should be tailored to the type of model, the decision context, and the user's need for recourse or understanding [10]. For autonomous distribution grids, the most useful explanation is not a technical trace of every internal computation but a decision account that links grid state, operational objective, constraint status, and chosen action. This layer therefore prioritizes actionable explanation over abstract interpretability.

Target audiences: operator, regulator, consumer

The Explainable Control Layer must serve different audiences without assuming that all stakeholders need the same level of detail. Operators need timely operational explanations; regulators need evidence that decisions followed approved governance rules; consumers may need simplified accounts when autonomous actions affect service quality or distributed resources. Research on trusted autonomy and human oversight highlights that explanations must support human judgment at the point where intervention remains possible [14]. The architecture therefore provides tiered explanations that share a common decision record while varying in technical depth and institutional purpose.

Link to other layers

The Explainable Control Layer is linked to the Audit Log Layer because every explanation must be stored with the decision it describes. It is linked to the Human Override Layer because operators need explanations before deciding whether to permit, veto, or modify an autonomous action. It is linked to the Safety Constraint Layer because explanations must disclose whether safety constraints shaped, blocked, or modified the controller's preferred action. Reviews of

explainability in deep reinforcement learning show that explanation and control are increasingly intertwined when autonomous policies operate in complex environments [15].

Audit Log Layer

Purpose and legal basis

The Audit Log Layer creates a durable institutional memory of autonomous grid operation. Blockchain-based demand-response research illustrates how secure records can support accountability among distributed actors who may not fully trust one another [16]. In the proposed architecture, audit logs record not only the final action but also the decision context, explanation, authority path, safety checks, and any human intervention. This makes the log a governance instrument for compliance, learning, dispute resolution, and forensic reconstruction.

Data to be logged

The audit record should include the triggering event, relevant grid state, data sources, controller recommendation, explanation output, safety-layer response, operator action, and final executed command. Surveys of blockchain for future smart grids emphasize that decentralized energy systems require trustworthy information handling across operational and organizational boundaries [17]. For autonomous distribution grids, logging must also capture the absence of action when the controller chooses to wait, defer, or escalate. A complete log therefore records both action and restraint, because both can have safety and accountability consequences.

Integrity and non-repudiation

Integrity means that records cannot be silently altered after the fact, while non-repudiation means that responsible actors and systems cannot plausibly deny their recorded actions. Blockchain-enabled smart-grid surveys identify tamper resistance, distributed verification, and secure data provenance as recurring motivations for ledger-based approaches [18]. The proposed architecture does not require a single technical implementation, but it requires that the audit mechanism provide equivalent governance properties. Whether implemented through distributed ledgers, cryptographic sealing, secure event stores, or institutional repositories, the log must be trustworthy enough to support review.

Integration with explainability and safety

The Audit Log Layer binds the other governance layers into a coherent record by preserving explanations, override decisions, and safety interventions together. Reviews of blockchain for smart-grid management show that logging and control become more valuable when they are connected to operational processes rather than isolated administrative systems [19]. In this architecture, an explanation without a log is ephemeral, an override without a log is institutionally fragile, and a safety intervention without a log is difficult to certify or learn from. The audit layer therefore acts as the accountability backbone of autonomous distribution-grid governance.

Human Override Layer

When override is permitted

The Human Override Layer defines the conditions under which an operator may interrupt, delay, veto, or modify an autonomous grid action. Human-in-the-loop machine-learning research emphasizes that human involvement must be connected to concrete intervention points, not treated as a symbolic assurance mechanism [20]. In this architecture, override is permitted when the controller's explanation is incomplete, when field conditions conflict with the digital state estimate, when social consequences require judgment, or when the operator identifies a credible safety or reliability concern. Override is therefore framed as a governed authority, not as an informal escape route from automation.

Override mechanisms and authority levels

Override mechanisms must distinguish between routine modification, emergency veto, supervisory escalation, and system suspension. Research on human-in-the-loop approaches shows that users' roles must be designed explicitly because different forms of interaction create different responsibilities and risks [21]. A control-room operator may be authorized to delay a switching command, while a senior reliability coordinator may be required to suspend an autonomous policy across a broader operating area. The architecture assigns each override type to a defined authority level so that intervention is fast, legitimate, and reviewable.

Training and decision support for operators

Operators must be trained not only to understand grid physics but also to interpret autonomous explanations, audit cues, and safety-layer warnings. Work on trusted autonomy argues that human-on-the-loop systems require interfaces that help people remain situationally aware even when automation performs most operational tasks [14]. The Human Override Layer therefore includes decision-support displays that show the controller's proposed action, the explanation, affected assets, safety status, and available alternatives. Training should cultivate disciplined intervention, avoiding both blind trust in autonomy and excessive interruption of safe automated behavior.

Override logging and review

Every override must be logged with the reason for intervention, the authority invoked, the information available to the operator, and the final system response. Philosophical work on meaningful human control stresses that human authority must be reason-responsive, meaning that interventions should be connected to understandable reasons and institutional responsibilities [22]. In the proposed architecture, override review is used to improve procedures, detect recurring controller weaknesses, and assess whether operators are being asked to intervene under unrealistic time pressure. This review function connects human authority to organizational learning.

Safety Constraint Layer

Defining safety constraints

The Safety Constraint Layer defines the non-negotiable physical and operational boundaries within which autonomous control must remain. Research on meaningful human control in dual-mode autonomous systems shows that responsibility depends on systems responding appropriately to reasons and constraints, especially when authority is shared between humans and automation [23]. In distribution grids, safety constraints may include equipment protection, voltage quality, switching permissibility, islanding rules, fault-current limits, and emergency operating restrictions. These constraints are defined before operation and are treated as architectural commitments rather than adjustable preferences.

Implementation as hard limits

Safety constraints must be implemented as hard limits that can block, reshape, or downgrade an autonomous action before execution. Control barrier function research demonstrates how safety-critical systems can be designed so that unsafe actions are prevented at the control level rather than merely discouraged [13]. In the proposed architecture, the safety layer evaluates controller outputs before commands reach field devices, ensuring that operational rules remain enforceable even when the AI policy is uncertain. The controller may recommend an action, but the safety layer determines whether that action is admissible.

Interaction with AI and operator

The Safety Constraint Layer has priority over both autonomous optimization and human override. Predictive safety-filter research for learning-based control shows that safety can be inserted as a supervisory mechanism around controllers that

learn or adapt under uncertainty [24]. In this architecture, neither the AI nor the operator may authorize a command that violates a hard safety boundary; instead, the system must reject the action, explain the rejection, and log the event. This hierarchy prevents governance from collapsing during emergencies, when both automated and human actors may face pressure to act quickly.

Safety certification and compliance

Safety certification requires evidence that constraints are complete, current, enforceable, and auditable across operating conditions. Model-free voltage-control research in active distribution systems illustrates the operational promise of learning-based control, but such approaches require governance mechanisms that demonstrate safe behavior before broad deployment [25]. The proposed architecture supports certification by linking each safety constraint to its technical rationale, responsible owner, implementation point, exception policy, and audit record. **Table 1** summarizes how the four governance layers distribute accountability across explanation, traceability, authority, and constraint enforcement.

Table 1. Governance Architecture for Accountability, Oversight, and Safety in Autonomous Control Systems

Governance layer	Primary governance function	Key mechanisms	Accountability implications
Explainable Control Layer	Makes autonomous decisions understandable before and after action	Decision narratives, counterfactual explanations, audience-specific explanation views, safety-status summaries	Establishes whether operators and reviewers can understand why a control action was proposed or executed
Audit Log Layer	Creates a trustworthy record of system behavior and institutional response	Tamper-evident event records, cryptographic sealing, time-stamped decision chains, forensic reconstruction tools	Enables compliance review, dispute resolution, organizational learning, and responsibility assignment
Human Override Layer	Preserves meaningful human authority over autonomous action	Veto rights, escalation protocols, role-based permissions, operator decision-support interfaces	Clarifies who may intervene, when intervention is legitimate, and how intervention is reviewed
Safety Constraint Layer	Enforces non-negotiable physical and operational limits	Hard constraint filters, protection-aware command screening, fail-safe rejection, certification evidence	Ensures that neither AI optimization nor human discretion can authorize unsafe commands

Implementation Scenarios

Fault isolation and service restoration

In a post-fault reconfiguration scenario, the autonomous controller identifies a faulted section, proposes isolation, and recommends a restoration path for unaffected customers. Multi-task reinforcement-learning work on distribution-system voltage control shows that autonomous policies may need to remain effective even when network topology changes [26]. The Explainable Control Layer would state why a feeder segment is isolated, the Audit Log Layer would preserve the sequence of sensor inputs and switching recommendations, the Human Override Layer would allow an operator to delay a command if field crews report conflicting conditions, and the Safety Constraint Layer would block any action that violates switching or protection limits. This scenario shows how the layers operate as a coordinated governance envelope around fast restoration.

Voltage control with high DER penetration

In a high distributed-energy-resource scenario, the autonomous controller may coordinate inverters, tap changers, and flexible loads to maintain acceptable voltage conditions. Reviews of blockchain-based distributed energy emphasize that distributed coordination introduces institutional questions about participation, verification, and responsibility among many actors [27]. The Explainable Control Layer would identify the operating condition and affected resources, while the Audit Log Layer would record the control sequence for later review by utilities, aggregators, or regulators. The Human Override Layer would permit operator intervention during abnormal feeder conditions, and the Safety Constraint Layer would prevent control actions that compromise equipment limits or customer protection expectations.

Emergency load shedding

In an emergency load-shedding scenario, autonomy may be asked to act quickly to prevent cascading deterioration or protect critical equipment. Systematic reviews of blockchain in the energy sector show that distributed digital infrastructures create opportunities for coordination but also require clear governance of trust, access, and accountability [28]. The proposed architecture requires the controller to explain the emergency rationale, the audit system to preserve the decision chain, the operator interface to support escalation where time permits, and the safety layer to enforce protected-load and equipment constraints. This scenario is especially important because emergency conditions test whether governance remains active under stress rather than disappearing when speed becomes urgent.

Validation/Governance Checklist

Checklist development principles

The validation checklist translates the conceptual architecture into a qualitative assessment tool for developers, utilities, regulators, and auditors. Governance research on artificial intelligence argues that accountability must be distributed across institutions, design processes, and operational practices rather than assigned only after harm occurs [29]. The checklist therefore evaluates whether each layer has a defined function, a responsible owner, an interface with other layers, and evidence suitable for review. It is not a numerical scoring instrument, but a structured maturity guide for determining whether autonomous grid governance is credible.

Sample checklist items per layer

Checklist items should ask whether explanations are understandable, whether logs are tamper-evident, whether override authority is meaningful, and whether safety limits are enforceable. Research on accountability in artificial intelligence highlights that accountability depends on answerability, responsibility, and mechanisms for correction when systems fail or cause harm [30]. For the Explainable Control Layer, evaluators should examine whether explanations are available at the moment of decision; for the Audit Log Layer, they should examine whether records can reconstruct events without ambiguity. For the Human Override and Safety Constraint Layers, they should examine whether operators have usable authority while safety boundaries remain inviolable.

Using the checklist for regulatory readiness

Regulatory readiness requires showing that autonomy has been designed with governance functions that can be inspected before deployment and reviewed after operation. Research on secure private data sharing in smart grids indicates that governance mechanisms must protect both operational integrity and accountable access to information [11]. The checklist can support procurement, pre-deployment review, incident investigation, and periodic reassessment as controllers, assets, and institutional expectations evolve. **Table 2** presents a governance assessment checklist that can be adapted for utility planning, regulatory hearings, or independent assurance reviews.

Table 2. Governance Assessment Checklist for Explainability, Accountability, Human Oversight, and Safety Assurance in Autonomous Grid Control Systems

Governance domain	Sample criterion	Evaluation question	Evidence expected
Explainable Control Layer	Explanations are generated for every autonomous recommendation and action	Can an operator understand the reason for the proposed action before deciding whether to intervene?	Explanation records, interface screenshots, explanation templates, operator review notes
Explainable Control Layer	Explanations are tailored to different audiences	Can regulators and affected stakeholders receive appropriate explanations without exposing unnecessary operational detail?	Audience-specific explanation formats, access rules, communication procedures
Audit Log Layer	Logs are complete and tamper-evident	Can a reviewer reconstruct the decision chain from input data to final command?	Time-stamped logs, integrity proofs, access records, forensic review procedures
Audit Log Layer	Logs include autonomous restraint and rejected actions	Are blocked, deferred, or overridden actions recorded as governance-relevant events?	Rejection records, safety-filter outputs, override entries, non-action explanations
Human Override Layer	Override authority is role-based and operationally usable	Does the operator know when and how to veto, modify, or escalate an autonomous action?	Role matrices, control-room procedures, training materials, interface permissions
Human Override Layer	Overrides are reviewed for learning and accountability	Are interventions analyzed to improve system design and operational procedures?	Review minutes, incident reports, retraining decisions, controller update records
Safety Constraint Layer	Hard limits cannot be bypassed by AI or operator command	Does the system reject unsafe actions even when they are recommended or manually requested?	Safety-filter design documents, rejection logs, certification evidence, exception policies
Safety Constraint Layer	Constraints are maintained as grid conditions evolve	Are safety rules updated when assets, protection settings, or operating policies change?	Constraint registers, change-control records, approval workflows, periodic reassessments

Conclusion

This article has proposed a layered governance architecture for autonomous distribution grids, centered on explainable control, audit logs, human override, and safety constraints. The architecture treats governance as part of the control system's design rather than as a documentary obligation imposed after deployment. Its purpose is to ensure that autonomous decisions can be understood, reconstructed, challenged, and bounded.

The Explainable Control Layer supports transparency by turning autonomous decisions into accounts that operators, regulators, and affected stakeholders can interpret. The Audit Log Layer creates a trustworthy record of decisions, inputs, explanations, safety interventions, and human actions. The Human Override Layer preserves meaningful authority for

trained operators, while the Safety Constraint Layer establishes limits that neither machine nor human discretion may violate.

The architecture also clarifies the hierarchy among governance functions. Explanations support understanding, logs support accountability, overrides support human authority, and constraints support safety supremacy. Together, these layers create a governance envelope that can accompany autonomous controllers in restoration, voltage control, emergency response, and future grid-management functions.

Future work should translate this conceptual architecture into pilot deployments, procurement specifications, certification procedures, and regulatory guidance. Utilities and regulators can use the proposed checklist to examine whether autonomous grid systems are mature enough for operational deployment. Trustworthy autonomy in distribution grids will depend not only on better algorithms, but on architectures that make responsibility, safety, and institutional legitimacy visible by design.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 Oct 2025

Revised: 16 Nov 2025

Accepted: 16 Jan 2026

Published online: 20 March 2026

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

Kiasari M, Aly H. Agentic Artificial Intelligence for Smart Grids: A comprehensive review of autonomous, safe, and explainable control frameworks. *Energies*. 2026;19(3):617.

Li Y, Yu C, Shahidehpour M, Yang T, Zeng Z, Chai T. Deep reinforcement learning for smart grid operations: Algorithms, applications, and prospects. *Proc IEEE*. 2023;111(9):1055-96.

Shi Z, Yao W, Li Z, Zeng L, Zhao Y, Zhang R, et al. Artificial intelligence techniques for stability analysis and control in smart grids: Methodologies, applications, challenges and future directions. *Appl Energy*. 2020;278:115733.

Cao D, Hu W, Zhao J, Zhang G, Zhang B, Liu Z, et al. Reinforcement learning and its applications in modern power and energy systems: A review. *J Mod Power Syst Clean Energy*. 2020;8(6):1029-42.

Babayomi O, Zhang Z, Dragicevic T, Hu J, Rodriguez J. Smart grid evolution: Predictive control of distributed energy resources—A review. *Int J Electr Power Energy Syst*. 2023;147:108812.

Sifat MM, Choudhury SM, Das SK, Ahamed MH, Muyeen SM, Hasan MM, et al. Towards electric digital twin grid: Technology and framework review. *Energy AI*. 2023;11:100213.

Omitaomu OA, Niu H. Artificial intelligence techniques in smart grid: A survey. *Smart Cities*. 2021;4(2):548-68.

Khan AA, Laghari AA, Rashid M, Li H, Javed AR, Gadekallu TR. Artificial intelligence and blockchain technology for secure smart grid and power distribution Automation: A State-of-the-Art Review. *Sustain Energy Technol Assess*. 2023;57:103282.

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.

Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf Fusion*. 2023;99:101805.

Wang Y, Su Z, Zhang N, Chen J, Sun X, Ye Z, et al. SPDS: A secure and auditable private data sharing scheme for smart grid based on blockchain. *IEEE Trans Ind Inform*. 2020;17(11):7688-99.

Middleton SE, Letouzé E, Hossaini A, Chapman A. Trust, regulation, and human-in-the-loop AI: Within the European region. *Commun ACM*. 2022;65(4):64-8.

Wu S, Liu T, Egerstedt M, Jiang ZP. Quadratic programming for continuous control of safety-critical multiagent systems under uncertainty. *IEEE Trans Autom Control*. 2023;68(11):6664-79.

Nahavandi S. Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. *IEEE Syst Man Cybern Mag*. 2017;3(1):10-7.

Hickling T, Zenati A, Aouf N, Spencer P. Explainability in deep reinforcement learning: A review into current methods and applications. *ACM Comput Surv*. 2023;56(5):1-35.

Jindal A, Aujla GS, Kumar N, Villari M. GUARDIAN: Blockchain-based secure demand response management in smart grid system. *IEEE Trans Serv Comput*. 2019;13(4):613-24.

Mollah MB, Zhao J, Niyato D, Lam KY, Zhang X, Ghias AM, et al. Blockchain for future smart grid: A comprehensive survey. *IEEE Internet Things J*. 2020;8(1):18-43.

Liu C, Zhang X, Chai KK, Loo J, Chen Y. A survey on blockchain-enabled smart grids: Advances, applications and challenges. *IET Smart Cities*. 2021;3(2):56-78.

Aklilu YT, Ding J. Survey on blockchain for smart grid management, control, and operation. *Energies*. 2022;15(1):193.

Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev*. 2023;56(4):3005-54.

Gómez-Carmona O, Casado-Mansilla D, Lopez-de-Ipina D, García-Zubia J. Human-in-the-loop machine learning: Reconceptualizing the role of the user in interactive approaches. *Internet Things*. 2024;25:101048.

Santoni de Sio F, Van den Hoven J. Meaningful human control over autonomous systems: A philosophical account. *Front Robot AI*. 2018;5:323836.

Mecacci G, Santoni de Sio F. Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics Inf Technol*. 2020;22(2):103-15.

Wabersich KP, Zeilinger MN. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*. 2021;129:109597.

Cao D, Zhao J, Hu W, Ding F, Yu N, Huang Q, et al. Model-free voltage control of active distribution system with PVs using surrogate model-based deep reinforcement learning. *Appl Energy*. 2022;306:117982.

Pei Y, Zhao J, Yao Y, Ding F. Multi-task reinforcement learning for distribution system voltage control with topology changes. *IEEE Trans Smart Grid*. 2023;14(3):2481-4.

Ahl A, Yarime M, Tanaka K, Sagawa D. Review of blockchain-based distributed energy: Implications for institutional development. *Renew Sustain Energy Rev*. 2019;107:200-11.

Andoni M, Robu V, Flynn D, Abram S, Geach D, Jenkins D, et al. Blockchain technology in the energy sector: A systematic review of challenges and opportunities. *Renew Sustain Energy Rev*. 2019;100:143-74.

Taeihagh A. Governance of artificial intelligence. *Policy Soc*. 2021;40(2):137-57.

Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: What it is and how it works. *Ai Soc*. 2024;39(4):1871-82.