REVIEW **Open access**

Explainable and Trustworthy AI for Smart Energy Systems: A Critical Review of Frameworks, Limitations, and Future Research Directions

Emma Wilson[1*], Frederik De Jong[2], James Brown[1], Lara Peters[2]

Abstract

Explainable and trustworthy artificial intelligence are becoming essential for smart energy systems that increasingly depend on complex models for forecasting, optimization, anomaly detection, and operational decision support. Although explainable AI methods and trustworthy AI principles have both expanded in the smart energy literature, they are often treated as separate concerns, leaving a gap between technical interpretability and the broader governance requirements of critical energy infrastructure. This review critically examines peer-reviewed articles published to evaluate how explainability, trustworthiness, accountability, human oversight, and regulatory readiness are addressed in smart energy AI research. The synthesis shows that current work remains dominated by method-level applications of explainability, especially feature attribution, interpretable forecasting, and post-hoc model interpretation, while system-level frameworks for stakeholder use, auditability, fairness, privacy, robustness, and institutional accountability remain underdeveloped. The review argues that explainability should not be treated as a narrow technical add-on, but as part of a broader socio-technical framework that connects AI model behavior with operator decision-making, consumer rights, regulatory oversight, and public trust. Future research should move toward energy-specific XAI frameworks that are stakeholder-validated, operationally scalable, uncertainty-aware, and embedded within lifecycle governance processes. Building trustworthy AI for smart energy systems will require closer integration among AI researchers, energy engineers, regulators, utilities, social scientists, and affected communities so that AI-supported energy decisions become not only accurate, but also transparent, contestable, accountable, and socially legitimate.

Keywords Critical review, Explainable AI, Smart energy systems, Trustworthy AI, Frameworks, Limitations

*Correspondence:
Emma Wilson
emma.wilson@ed.ac.uk
[1] Department of Intelligent Electrical Networks and AI, Faculty of Engineering, University of Edinburgh, Edinburgh, United Kingdom
[2] Department of Artificial Intelligence for Smart Microgrids, Faculty of Electrical Engineering, Delft University of Technology, Delft, Netherlands

Introduction

Artificial intelligence has become increasingly embedded in smart energy systems, including smart grids, building energy management, solar forecasting, cyber-physical power systems, and demand-side applications. Reviews of smart-grid AI show that machine learning is now used across forecasting, control, monitoring, and security functions, but this expansion also intensifies concerns about opacity, accountability, and operational trust [1]. In critical infrastructure settings, opaque decisions are not merely technical inconveniences, because forecasting errors, unjustified automated actions, or poorly understood alerts may affect reliability, safety, and public confidence [2]. Explainable AI has therefore become a central

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

response to the black-box problem, particularly as energy researchers increasingly adopt post-hoc explanation tools and interpretable modeling approaches [3].

A recurring limitation in the literature is that explainability and trustworthiness are often treated as separate design concerns rather than as mutually dependent requirements of smart energy systems. Energy-focused XAI reviews identify growing use of SHAP, feature attribution, attention mechanisms, and interpretable machine learning, but they also note that many studies remain method-driven rather than governance-driven [3, 4]. Broader XAI surveys have developed rich taxonomies of explanation methods, yet these taxonomies do not automatically resolve sector-specific questions about grid operation, regulatory accountability, consumer protection, and institutional responsibility [5, 6]. This creates a gap between technical explainability and the broader conditions under which AI can be considered trustworthy in energy infrastructure.

The objective of this review is to critically evaluate the state of explainable and trustworthy AI frameworks for smart energy systems, with attention to both technical methods and socio-technical governance. Prior work on trustworthy AI emphasizes transparency, accountability, fairness, privacy, and robustness as core dimensions of responsible AI, but these principles require careful reinterpretation when applied to energy systems that combine real-time operation, market coordination, consumer participation, and critical infrastructure risk [7, 8]. The review therefore examines whether existing energy XAI frameworks provide actionable support for operators, consumers, regulators, and system designers. It also identifies limitations in current approaches and proposes future research directions for more integrated, energy-specific frameworks.

The distinctive contribution of this review is its integrated critique of three literatures that are often discussed separately: XAI methods, trustworthy AI principles, and governance requirements for AI in energy systems. Existing surveys map technical methods and application domains, while AI ethics literature provides general principles that may not be immediately operational in electricity systems [9, 10]. Governance-focused work shows that electricity systems raise distinctive regulatory and institutional questions, especially where AI systems influence infrastructure management and public-service obligations [11]. This review connects these streams to argue that explainability should be evaluated not only by technical fidelity, but also by its contribution to accountable, usable, and trustworthy energy decision-making.

Figure 1 presents the integrated conceptual framework guiding this review, showing how smart energy AI applications, explanation methods, stakeholder interpretation, trustworthy AI principles, and governance outcomes must be connected across the full system lifecycle.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

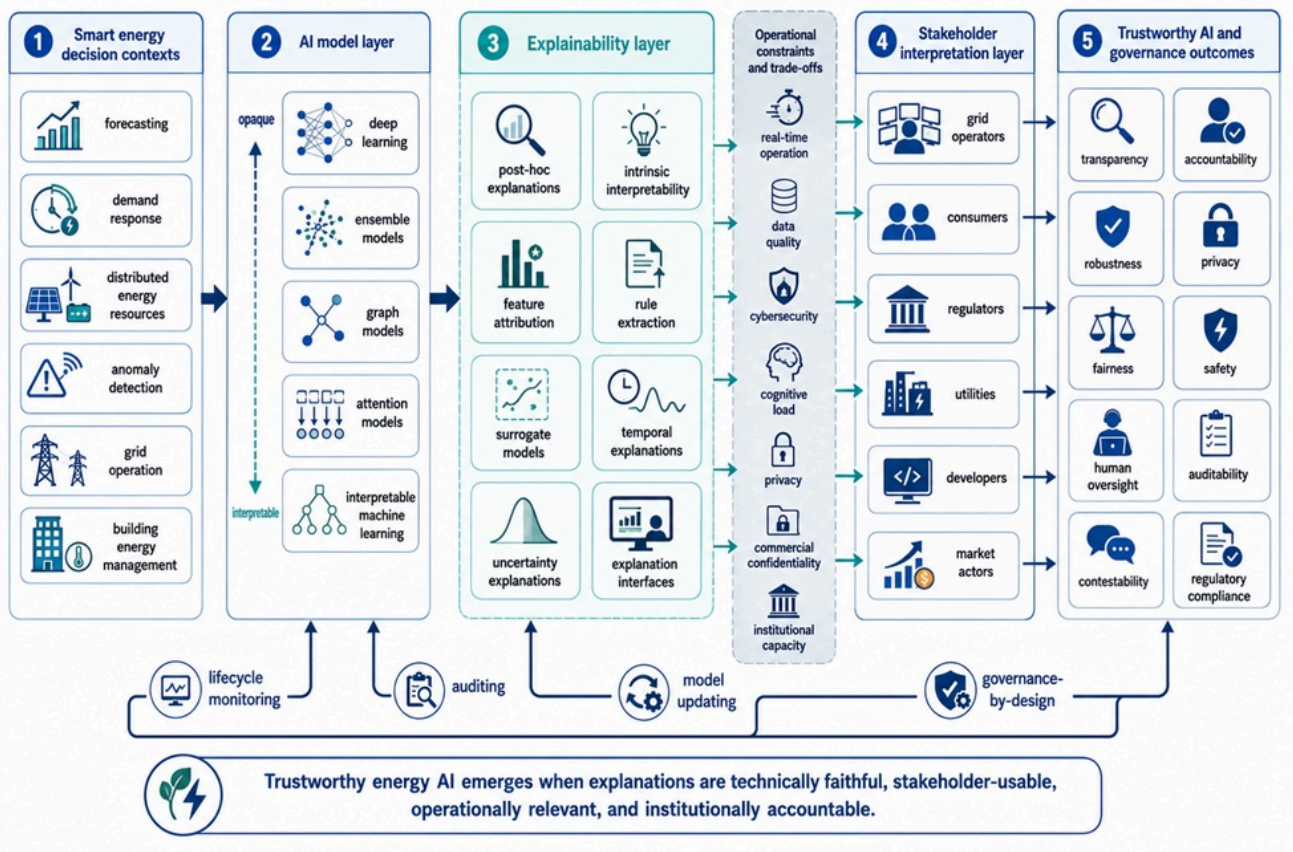

Figure 1. Integrated Explainability–Trustworthiness Framework for AI in Smart Energy Systems

Background and Conceptual Scope

Smart energy systems and the imperative for explainability and trust

Smart energy systems can be understood as digitally enabled electricity and energy infrastructures that integrate distributed generation, advanced sensing, demand response, automated control, consumer participation, and data-driven decision support. In such systems, AI models are increasingly used to forecast demand, optimize distributed resources, detect anomalies, and support grid management, as shown in reviews of machine-learning-driven smart power systems and smart-grid analytics [1, 12]. Because electricity systems are safety-critical and socially embedded, explainability and trustworthiness are not optional add-ons but preconditions for responsible deployment. The literature reveals that technical accuracy alone is insufficient where decisions must be understood by operators, justified to regulators, and accepted by consumers affected by automated or semi-automated energy decisions [11].

Defining explainable AI in technical and socio-technical terms

Explainable AI is often defined technically as a family of methods that make model predictions more interpretable through feature attribution, surrogate modeling, rule extraction, saliency mapping, attention visualization, or intrinsically interpretable model design. Major XAI surveys distinguish between transparent models and post-hoc explanations, while also emphasizing that explanation quality depends on the purpose, audience, and context of use [5, 13]. In smart energy systems, this means that an explanation useful to a machine-learning engineer may not be useful to a grid operator responding to a real-time alert or a consumer contesting a demand-response decision. Social-science-informed XAI research further shows that explanation is relational and audience-dependent, suggesting that energy XAI frameworks

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

must address understanding, contestability, responsibility, and decision context rather than merely displaying model features [14].

Principles and dimensions of trustworthy AI

Trustworthy AI frameworks commonly emphasize transparency, accountability, fairness, robustness, privacy, safety, and human oversight as normative and design requirements. AI ethics literature has consolidated these principles across many domains, but it also shows that high-level guidelines often remain abstract unless translated into concrete institutional practices [7, 10]. In energy systems, transparency may require traceable operational decisions, fairness may concern distributional impacts of demand response or energy access, robustness may involve resilience against cyber-physical disturbances, and privacy may relate to granular smart-meter or household behavior data. A recurring limitation is that energy AI studies often invoke trustworthiness rhetorically while providing limited detail on how these principles should be operationalized across technical architecture, governance processes, and stakeholder rights [8].

Existing taxonomies and reviews of XAI in energy

Existing reviews of XAI in energy provide valuable taxonomies of explanation methods, application areas, and unresolved challenges. Machlev and colleagues explicitly review XAI techniques for energy and power systems, identifying opportunities in forecasting, control, and grid management while also noting gaps in evaluation and practical adoption [3]. Smart-grid-focused reviews similarly discuss interpretable machine learning as a response to the opacity of increasingly complex energy AI models, but their primary orientation remains technical and methodological [4]. This review builds on those contributions by asking a broader framework question: whether XAI methods are being connected to trustworthy AI principles, stakeholder-specific explanation needs, and governance obligations in smart energy systems.

Governance and regulatory drivers for explainable and trustworthy AI in energy

Governance and regulatory drivers are becoming increasingly important because AI-supported energy systems operate within public-service, safety, reliability, and market-accountability regimes. Work on governing AI in electricity systems has argued that emerging AI regulation must be interpreted in relation to the distinctive risks and institutional structures of electricity infrastructure [11]. Cyber-physical power-system research also shows that digitalization increases exposure to security, resilience, and system-interdependence risks, making trustworthy AI a matter of infrastructure governance rather than only software quality [2]. These drivers suggest that explainability must support auditability, compliance, human oversight, and institutional accountability, not simply model interpretation.

Table 1 summarizes the major governance drivers shaping explainable and trustworthy AI adoption in smart energy systems.

Table 1. Governance Drivers and Institutional Requirements for Explainable AI in Smart Energy Systems

Governance driver	Energy-system relevance	Institutional requirement	Key implementation challenge
Critical infrastructure protection	AI systems increasingly influence grid stability, forecasting, and operational control	Mandatory risk assessment and operational oversight	Balancing automation efficiency with human supervision
Regulatory accountability	Energy systems operate under strict public-service and compliance obligations	Auditability, traceable decisions, and documentation standards	Lack of standardized explanation reporting practices
Consumer trust and participation	Demand-response systems and smart-meter platforms directly	Transparent consumer-facing explanations and	Limited public understanding of AI-supported energy decisions

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

	affect households and prosumers	contestability mechanisms	
Cybersecurity and resilience	AI-enabled grids are vulnerable to cyber-physical disruptions	Robustness testing, anomaly monitoring, and secure model governance	Difficulty validating AI behavior under adversarial conditions
Interoperability across platforms	Smart grids integrate distributed assets, utilities, and digital platforms	Shared explanation formats and interoperable governance standards	Fragmented technical architectures and vendor ecosystems
Ethical and fairness obligations	AI may influence pricing, access, and energy allocation decisions	Bias monitoring and fairness-oriented governance procedures	Weak integration between fairness evaluation and operational AI systems

Review of Explainable AI Frameworks in Energy

Post-Hoc and intrinsic explainability methods for energy AI

The energy XAI literature is dominated by post-hoc explanation methods applied to forecasting, anomaly detection, and operational prediction tasks. For example, SHAP-based and related explanation approaches have been used to interpret solar photovoltaic generation forecasts, showing how feature contributions can make predictive outputs more transparent to analysts and system designers [15]. Other work on interpretable electricity demand forecasting and residential load forecasting uses model design choices, attention mechanisms, or interpretable representations to expose temporal and behavioral patterns [16, 17]. However, a recurring limitation is that such studies often clarify what variables influence model output without establishing whether the resulting explanations are operationally meaningful, cognitively usable, or institutionally sufficient for accountability.

Table 2 extends the method-level discussion by mapping major XAI method families to smart energy decision requirements, stakeholder value, unresolved limitations, and governance implications.

Table 2. Analytical matrix linking XAI methods to smart energy decision requirements

XAI method family	Primary explanatory function	Smart energy use cases	Main stakeholder value	Key limitation in current literature	Governance implication
Feature attribution methods	Identify variables contributing to model outputs	Load forecasting, solar forecasting, demand prediction, anomaly detection	Helps analysts and operators see influential drivers behind predictions	Often explains correlations without proving causal or operational relevance	Requires documentation of explanation fidelity, stability, and decision relevance
Surrogate and local explanation models	Approximate complex model behavior in interpretable form	Short-term forecasting, fault classification, grid-state prediction	Supports model debugging and local decision justification	May oversimplify complex temporal or cyber-physical dependencies	Should be validated against operational scenarios before being used for accountability
Intrinsically interpretable models	Make decision logic transparent by design	Building energy modeling, demand	Improves traceability and reduces	May trade off predictive performance in	Useful for high-accountability or

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

		response, consumer-facing systems	dependence on post-hoc rationalization	highly complex system settings	consumer-impacting decisions
Attention and temporal explanation methods	Highlight relevant time steps, sequences, or system states	Load forecasting, sequence-to-sequence prediction, fault diagnosis	Helps operators understand temporal patterns and warning signals	Attention weights may be misread as causal explanations	Requires careful interface design and explanation literacy
Rule extraction and symbolic explanations	Convert model behavior into human-readable rules	Operational control, demand-response logic, compliance review	Supports auditability and regulatory inspection	Rule sets may become too complex or unstable in dynamic systems	Can support formal review, certification, and institutional accountability
Visual explanation interfaces	Present explanation outputs through dashboards, heatmaps, or structured views	Grid monitoring, control rooms, consumer energy platforms	Supports rapid interpretation and human decision-making	Interface usability is rarely tested with real energy stakeholders	Must be evaluated for cognitive load, accessibility, and decision consequences
Uncertainty-aware explanations	Communicate confidence, ambiguity, and risk around predictions	Renewable forecasting, outage prediction, operational planning	Supports calibrated trust and risk-aware action	Underused in energy XAI despite uncertainty-heavy operating conditions	Critical for preventing overtrust and supporting justified reliance
Lifecycle explanation documentation	Records explanation design, validation, use, and monitoring over time	Procurement, deployment, auditing, regulatory compliance	Enables institutional memory and accountability	Rarely integrated into technical XAI studies	Provides the bridge between explainability and enforceable trustworthy AI governance

Explanation interfaces and human-centered XAI in energy

Explanation interfaces remain less developed than explanation algorithms in the smart energy literature. General XAI evaluation research emphasizes that explanations must be designed for human users, evaluated according to task context, and assessed in relation to user understanding rather than merely technical plausibility [18, 19]. Energy applications such as sequence-to-sequence load forecasting and spectrogram-based fault classification show the potential of visual or structured explanations for operational interpretation, but they do not fully resolve how grid operators, planners, consumers, and regulators should interact with such explanations in practice [20, 21]. This creates a gap between human-centered XAI theory and the limited interface-level evidence currently available in energy applications.

Conceptual frameworks positioning XAI within energy system architecture

Framework-level accounts of XAI in energy increasingly recognize that explanation should be embedded within broader system architecture rather than added after model development. Reviews of XAI in energy and smart grids describe

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

explanation methods across forecasting, operation, control, and monitoring, but they also reveal that many frameworks stop at method classification rather than specifying governance pathways, user roles, or decision responsibilities [3, 4]. Data-driven building energy modeling illustrates the value of interpretable machine learning for generalization and practical understanding, yet it also shows the difficulty of translating interpretability into deployable decision architecture [22]. The literature therefore reveals a need for conceptual frameworks that position XAI across the full AI lifecycle, from data collection and model development to operational use, auditing, and accountability.

Review of Trustworthy AI Frameworks in Energy

Principle-based trustworthy AI frameworks applied to energy

Principle-based trustworthy AI frameworks provide an important vocabulary for evaluating smart energy AI, but their translation into energy-specific requirements remains incomplete. General trustworthy AI scholarship argues that trustworthiness requires alignment among technical performance, ethical principles, organizational controls, and social expectations [8]. In energy systems, these principles must be adapted to settings where AI decisions may affect grid reliability, consumer participation, market behavior, or infrastructure security [1, 2]. A recurring limitation is that energy-focused studies often reference transparency, robustness, and privacy without developing clear mechanisms for resolving conflicts among these principles in concrete operational settings.

Trust modeling and measurement in AI-supported energy decisions

Trust in AI-supported energy decisions is often assumed to follow from explainability, but the literature provides a more complex picture. Social-science-informed XAI research argues that explanations affect trust only when they are understandable, relevant, contrastive, and responsive to the user's decision context [14]. In energy applications, explanation methods for load forecasting, photovoltaic forecasting, and grid classification can make model behavior more visible, yet visibility does not automatically produce calibrated trust among operators or consumers [15, 20, 23]. This review therefore identifies a conceptual gap: energy AI frameworks need to distinguish between increasing trust, calibrating trust, preventing overtrust, and supporting justified reliance. **Figure 2** conceptualizes trust in AI-supported energy decisions as a calibration problem in which explanation visibility, user context, uncertainty, and oversight must align before reliance on AI outputs can be justified.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

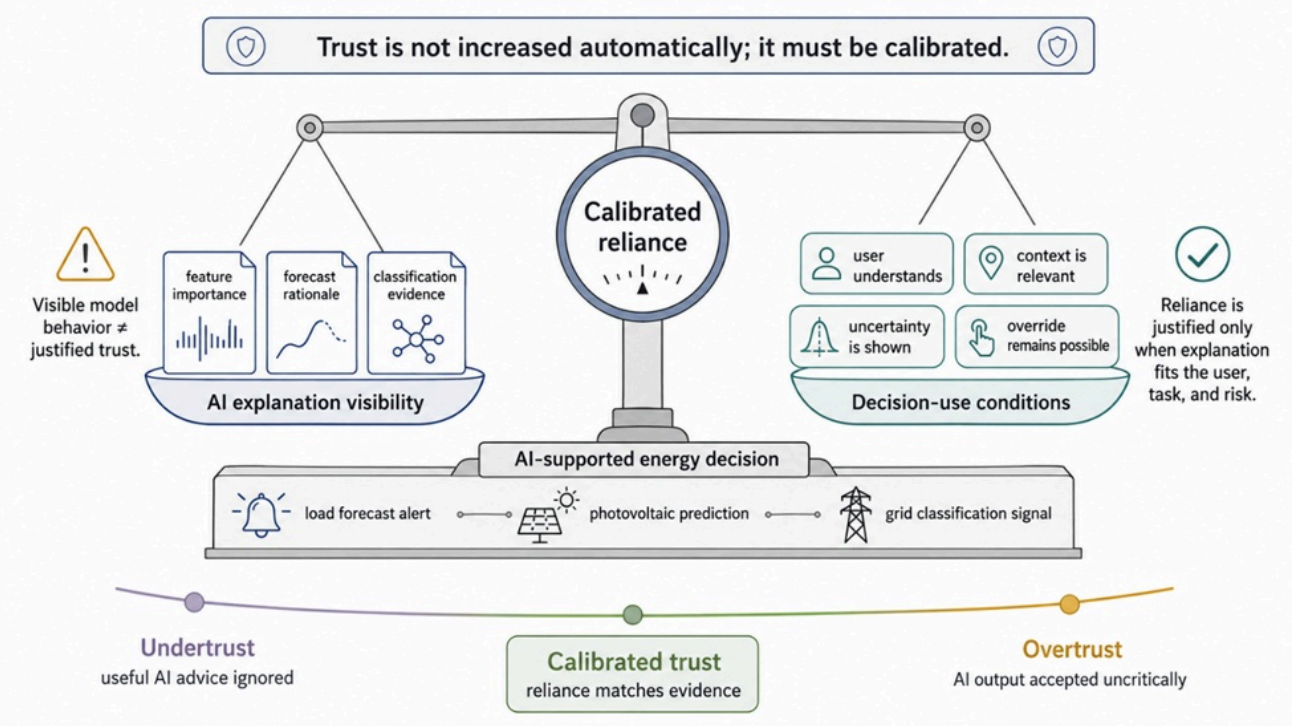

Figure 2. Trust as Calibration Rather Than Visibility in AI-Supported Energy Decisions

Certification, auditing, and compliance frameworks

Certification, auditing, and compliance frameworks remain underdeveloped in the energy XAI literature, despite their relevance to critical infrastructure. AI governance work in electricity systems highlights the challenge of aligning emerging AI regulation with electricity-sector obligations, including reliability, safety, transparency, and institutional accountability [11]. Broader AI ethics and trustworthy AI literature calls for mechanisms that move beyond principles toward assessment, oversight, and accountability structures [7, 8]. However, few energy XAI frameworks specify how explanations should be documented, audited, stress-tested, or used in regulatory review, leaving a significant gap between technical interpretability and enforceable compliance.

Integrating Explainability and Trustworthiness

The relationship between explanation and trust in energy AI

Explainability is frequently presented as a route to trust, but the relationship between explanation and trust in smart energy systems is contingent rather than automatic. Foundational XAI research shows that explanations can support understanding and accountability, but it also warns that explanation methods may be misleading if they lack fidelity, stability, or user relevance [5, 13]. In energy applications, interpretable forecasting and grid-monitoring studies demonstrate that explanations can reveal model behavior, yet they rarely test whether these explanations improve operator judgment, consumer acceptance, or regulatory confidence [16, 17, 21]. The literature therefore suggests that explainability should be treated as one condition of trustworthy AI, not as a substitute for robustness, fairness, privacy, governance, and human oversight.

Joint frameworks and their conceptual coherence

Joint frameworks linking XAI and trustworthy AI remain conceptually promising but unevenly developed in the energy literature. Energy-focused reviews and smart-grid surveys increasingly recognize that explainability supports

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

transparency and accountability, while trustworthy AI scholarship provides a broader normative structure for evaluating fairness, privacy, robustness, and responsibility [3, 4, 8]. Yet many frameworks lack coherence because they do not specify whether explainability is a technical method, a user-interface property, a governance requirement, or a stakeholder right. This ambiguity weakens operationalization, since different interpretations lead to different design obligations, evaluation criteria, and institutional responsibilities.

Table 3 consolidates trustworthy AI principles into energy-specific operational obligations, showing how abstract governance concepts can be translated into evaluable requirements for smart energy systems.

Table 3. Trustworthy energy AI consolidation framework: from abstract principles to operational obligations

Trustworthy AI dimension	Energy-specific interpretation	Operational obligation	Evidence required for evaluation	Main unresolved research gap
Transparency	AI-supported energy decisions must be understandable and traceable across data, model, and decision layers	Provide explanations appropriate to operators, consumers, regulators, and developers	Model cards, explanation records, decision logs, interface evaluations	Lack of stakeholder-specific transparency standards
Accountability	Responsibility must be assignable when AI affects grid operation, market behavior, or consumer outcomes	Define accountable actors across design, deployment, monitoring, and intervention	Governance maps, responsibility matrices, audit trails, escalation protocols	Unclear liability and responsibility allocation in multi-actor energy systems
Robustness	AI must remain reliable under uncertainty, missing data, disturbances, and changing grid conditions	Stress-test models under normal, edge-case, and adversarial scenarios	Scenario testing, resilience metrics, degradation analysis, post-deployment monitoring	Limited validation under real-time and cyber-physical operating conditions
Fairness	AI should avoid unjust distributional effects across consumers, communities, and market participants	Assess impacts of demand response, pricing, classification, and automation across groups	Bias audits, distributional impact assessments, consumer outcome analysis	Energy XAI rarely connects explanations to fairness or justice outcomes
Privacy	Smart-meter, household, and operational data must be protected while enabling useful explanation	Minimize sensitive data exposure in models, explanations, and audit systems	Data governance records, privacy-preserving explanation methods, access-control logs	Tension between explanation detail and privacy protection remains undertheorized
Safety	AI must not compromise physical grid reliability, infrastructure protection, or public-service continuity	Maintain human override, fail-safe design, and safety constraints in AI-supported decisions	Safety cases, incident records, human-in-the-loop protocols, validation reports	Few frameworks integrate XAI with safety engineering practices
Human oversight	Humans must be able to understand, challenge,	Match explanation depth to decision	Usability studies, operator trials, decision-	Limited empirical evidence on how

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

	override, or approve consequential AI decisions	criticality and user expertise	support evaluations	energy stakeholders actually use explanations
Contestability	Affected actors should be able to challenge or seek review of AI-supported decisions	Provide understandable justifications and review pathways for contested decisions	Complaint mechanisms, appeal records, explanation summaries, consumer-facing documentation	Consumer rights and procedural justice remain marginal in energy XAI
Auditability	AI behavior must be reviewable by institutions beyond the model development team	Preserve documentation across the AI lifecycle from design to deployment	Audit logs, compliance reports, model-version histories, third-party review evidence	Auditing requirements for explainable energy AI remain immature
Regulatory compliance	AI systems must align with energy-sector obligations and emerging AI governance requirements	Translate AI principles into procurement, grid operation, market, and consumer-protection rules	Compliance checklists, regulatory submissions, certification records	Weak integration between technical XAI frameworks and enforceable regulation

Stakeholder-specific requirements and trade-offs

Explainability and trustworthiness requirements differ substantially across grid operators, consumers, regulators, developers, and market actors. Operators may need concise, time-sensitive explanations that support situational awareness, while consumers may require understandable justifications for automated demand-response or energy-management decisions, and regulators may require audit trails that demonstrate accountability and compliance [11, 14]. Developers may prioritize model debugging and feature interpretation, as illustrated by technical XAI applications in solar forecasting, residential load prediction, and graph-based smart-grid modeling [15, 16, 23]. Few frameworks adequately address trade-offs among these needs, particularly where explanation depth may conflict with real-time constraints, privacy protection, cognitive load, or commercial confidentiality.

Critical Assessment of Limitations and Gaps

Technical limitations: scalability, real-time constraints, and data quality

A recurring limitation in XAI for smart energy systems is the difficulty of scaling explanation methods to high-volume, real-time, and heterogeneous operational environments. Smart grids generate data from meters, sensors, distributed energy resources, weather streams, control systems, and cyber-physical infrastructure, yet many interpretable modeling studies remain focused on bounded prediction tasks rather than full-system deployment constraints [9]. Short-term load forecasting work illustrates the growing sophistication of hybrid deep learning architectures, but it also shows that interpretability can become harder as models combine multiple architectures and temporal representations [24]. Future frameworks therefore need to address whether explanations can be generated quickly, reliably, and consistently enough for real operational use, especially under uncertainty, missing data, cyber disturbance, or rapidly changing grid conditions.

Empirical validation deficits

The literature reveals a persistent shortage of real-world validation of XAI methods in energy operations. Many studies demonstrate that explanation techniques can be attached to models for forecasting, classification, or anomaly interpretation, but fewer examine how these explanations perform when used by operators, planners, or consumers in authentic decision settings [3]. Expert-knowledge integration in solar radiation prediction suggests that domain knowledge

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

can improve interpretability, yet such work still leaves open the broader question of whether explanations support better institutional decisions across diverse energy contexts [25]. This review therefore identifies empirical validation as a central gap, because XAI claims remain incomplete when explanation usefulness is inferred from technical outputs rather than demonstrated through stakeholder-centered evaluation.

Socio-technical gaps: operator cognition, trust calibration, and usability

A major socio-technical gap concerns the limited integration of human factors, cognitive engineering, and usability research into energy XAI frameworks. Explanations are often treated as outputs of algorithms, but explanation theory emphasizes that users need contrastive, context-sensitive, and actionable accounts that fit their goals and prior knowledge [14]. Human-centered XAI evaluation frameworks similarly argue that explanation systems should be assessed according to user tasks, mental models, and decision consequences, rather than by technical properties alone [18]. In smart energy systems, this means that future work should examine how operators interpret alerts, how consumers understand automated energy decisions, and how regulators judge whether an explanation is sufficient for accountability.

Regulatory, ethical, and accountability gaps

Regulatory, ethical, and accountability gaps remain especially visible where XAI research stops at model interpretation rather than institutional responsibility. Trustworthy AI literature stresses that accountability requires identifiable responsibilities, documentation, oversight, and mechanisms for redress, yet energy AI frameworks rarely specify who is accountable when an interpretable model supports a harmful or contested operational decision [7, 8]. Work on AI governance in electricity systems shows that regulatory obligations must be adapted to sector-specific risks, including reliability, market fairness, infrastructure safety, and public-service duties [11]. A recurring limitation is that explainability is often framed as a transparency tool without being linked to contestability, liability, auditability, and enforceable accountability.

Governance, Policy, and Regulatory Challenges

Translating AI ethics principles into energy sector regulation

Translating AI ethics principles into energy-sector regulation is difficult because broad concepts such as transparency, fairness, and accountability must become concrete obligations within grid codes, market rules, procurement standards, and utility governance. Global AI ethics analyses show that principles are widely repeated across guidelines, but also that implementation mechanisms are often vague or inconsistent [10]. In electricity systems, the challenge is sharper because AI may influence dispatch, demand-response incentives, network operation, consumer classification, or infrastructure protection [11]. Future governance work should therefore focus less on restating generic AI ethics and more on defining operational duties for explainable and trustworthy AI across specific energy-sector decision points.

Standards and interoperability for explainable and trustworthy energy AI

Standards and interoperability are necessary for explainable and trustworthy energy AI, but the reviewed literature suggests that this area remains immature. Smart-grid AI surveys emphasize that energy systems depend on interoperable data flows, communication protocols, and cyber-physical coordination, yet XAI frameworks often focus on model-level explanation rather than system-level integration [12]. Cyber-physical power-system research further demonstrates that modeling, simulation, and security concerns must be addressed across interconnected infrastructures rather than isolated algorithms [2]. This creates a standards gap: energy XAI frameworks need shared documentation practices, explanation formats, audit interfaces, and lifecycle controls that can operate across vendors, regulators, utilities, and distributed energy platforms. **Figure 3** illustrates the standards gap in trustworthy energy AI: explainability must move beyond isolated model interpretation and become an interoperable governance layer connecting AI models, cyber-physical grid infrastructure, utilities, vendors, regulators, and distributed energy platforms.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

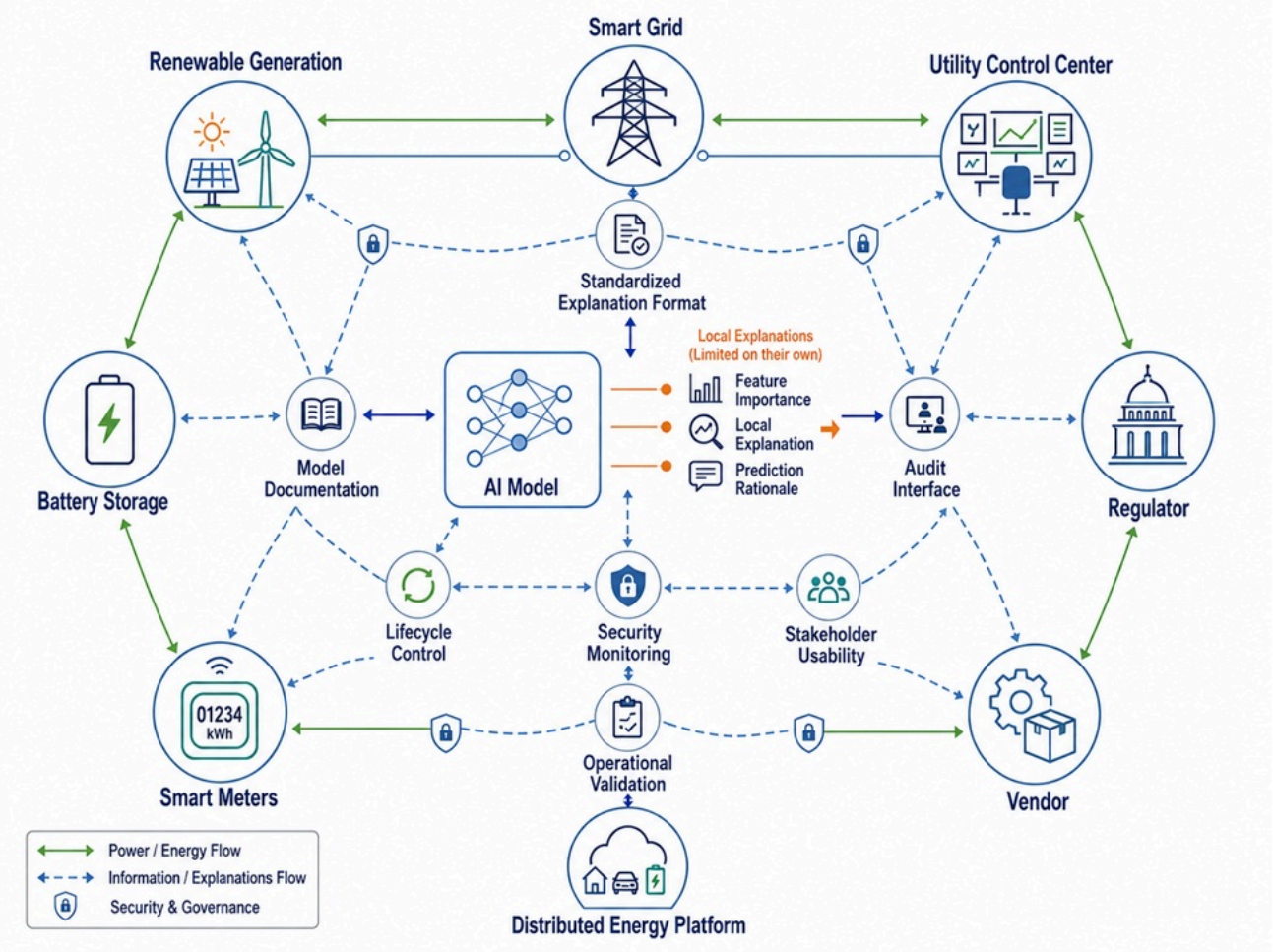

Figure 3. Interoperable Governance Layer for Explainable and Trustworthy Energy AI

Recent scholarship further emphasizes that explainability in energy systems must move beyond isolated interpretability techniques toward operationally validated and domain-specific governance frameworks. A recent systematic review of explainable AI in energy forecasting demonstrated that most studies prioritize local feature interpretation while providing limited evaluation of explanation robustness, computational feasibility, and stakeholder usability across operational energy contexts [3]. Similarly, emerging work on explainable smart-grid architectures argues that transparent AI integration can improve grid coordination and user confidence, but also highlights unresolved issues concerning scalability, interoperability, and governance alignment in real-world deployment environments [26]. These findings reinforce the argument that explainability alone is insufficient unless it is integrated with lifecycle governance, operational accountability, and stakeholder-centered evaluation mechanisms.

Institutional capacity and regulatory readiness

Institutional capacity is a critical but underdeveloped condition for trustworthy AI in smart energy systems. Regulatory agencies, utilities, system operators, and market institutions may be expected to evaluate AI systems, yet they may lack the technical expertise, audit tools, or organizational processes needed to assess complex explanations and model behavior [11]. Broader work on explainability and trustworthy AI indicates that meaningful oversight requires more than access to model outputs; it requires evaluation criteria, governance procedures, documentation, and capacity to interpret technical evidence [19]. Future work should therefore examine regulatory readiness as a research problem in its own right, rather than assuming that explainability will automatically make AI systems governable.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Public trust, energy democracy, and societal acceptance

Public trust in AI-enabled energy systems depends not only on whether algorithms can be explained, but also on whether affected communities view decisions as fair, contestable, and institutionally legitimate. AI ethics frameworks highlight the importance of justice, human autonomy, and social benefit, but these principles require domain-specific interpretation when applied to energy access, pricing, demand response, distributed resources, and consumer data [7]. Although energy XAI studies often focus on technical transparency, they rarely address broader questions of energy democracy, procedural justice, or public participation in AI governance. This review therefore identifies societal acceptance as a neglected dimension of trustworthy energy AI, especially where automated systems affect households, vulnerable consumers, or community energy arrangements.

Future Research Directions

From generic XAI to energy-specific explainability research

Future research should move from applying generic XAI tools to developing energy-specific explanation paradigms that reflect the temporal, operational, and institutional characteristics of smart energy systems. General tools such as feature attribution and local explanation methods provide useful starting points, but energy contexts require explanations that account for load dynamics, weather dependence, grid constraints, distributed assets, safety margins, and real-time decision pressures [3, 15]. Advances in local-to-global explanation methods, including tree-based explanation approaches, show how model behavior can be made more interpretable, but these methods must be adapted to the multi-actor and safety-critical context of energy systems [27]. A stronger research agenda would therefore develop explanation templates, evaluation criteria, and stakeholder-specific interfaces for forecasting, control, demand response, anomaly detection, and regulatory audit.

Table 4 presents a future research agenda linking explainability methods with governance-ready smart energy AI design.

Table 4. Future Research Priorities for Explainable and Trustworthy AI in Smart Energy Systems

Research priority	Current limitation identified in literature	Proposed future direction	Expected contribution
Stakeholder-centered explanation design	Explanations are often optimized for developers rather than operators or consumers	Develop adaptive explanation interfaces tailored to user expertise and operational context	Improved usability and calibrated trust
Real-time explainability	Many XAI methods are computationally intensive for live grid environments	Design lightweight and scalable explanation architectures for operational deployment	Enhanced applicability in real-time energy systems
Governance-by-design frameworks	Explainability is commonly added after deployment	Integrate accountability, auditability, and oversight mechanisms during system design	Stronger regulatory and institutional alignment
Explanation robustness evaluation	Existing studies rarely test explanation stability under uncertainty or adversarial conditions	Develop standardized robustness and fidelity benchmarks for energy AI explanations	More reliable and trustworthy operational explanations

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Human factors and cognitive evaluation	Limited evidence exists regarding how stakeholders interpret explanations	Conduct operator-focused and consumer-focused usability studies	Better alignment between explanation outputs and decision-making needs
Lifecycle AI governance	Documentation and monitoring remain fragmented across deployment stages	Create continuous lifecycle governance frameworks for energy AI systems	Improved accountability and long-term oversight

Integrating XAI and trustworthiness into energy AI governance by design

Future work should embed explainability and trustworthiness into energy AI governance by design, rather than treating them as post-hoc additions after model deployment. Trustworthy AI scholarship emphasizes that responsible AI requires lifecycle governance, including design choices, evaluation, monitoring, accountability, and institutional controls [8]. In smart energy systems, governance-by-design would require that explanation requirements be specified during procurement, system architecture, data governance, model validation, operational deployment, and post-deployment monitoring [11]. Such an agenda would align XAI methods with regulatory accountability, human oversight, privacy protection, robustness testing, and stakeholder rights from the beginning of energy AI development.

Methodological Reflections on the Review

Review approach and selection criteria

This review used a targeted literature approach to synthesize 29 peer-reviewed articles published between 2017 and 2023, selected for their relevance to explainable AI, trustworthy AI, smart energy systems, governance, and critical infrastructure. The selection intentionally combined energy-specific XAI studies, smart-grid AI reviews, broader explainability taxonomies, and trustworthy AI frameworks because the article's purpose was critical synthesis rather than exhaustive bibliographic coverage [2, 5, 8]. Foundational XAI and AI ethics sources were included where they provided concepts necessary to interpret energy-specific limitations, such as explanation purpose, trust calibration, transparency, and accountability [6, 7]. This targeted approach allowed the review to examine the conceptual relationship between XAI methods and trustworthy AI governance rather than merely cataloguing applications.

Quality assessment and critical appraisal

The quality assessment in this review was conceptual rather than quantitative, reflecting the article's purpose as a critical review of frameworks, limitations, and future research directions. Energy XAI studies were assessed according to whether they clarified explanation purpose, stakeholder audience, operational relevance, and governance implications, rather than according to predictive performance or statistical results [4, 22]. Broader XAI and trustworthy AI literature was appraised for conceptual transferability to energy systems, including whether its principles could support accountability, contestability, and lifecycle governance in critical infrastructure [13, 28]. This appraisal approach is appropriate for framework synthesis, but it also requires interpretive judgment because different articles define explainability, trust, and trustworthiness in different ways.

Limitations of the synthesis method

The synthesis method has limitations because it integrates diverse literatures with different assumptions, methods, and evaluation standards. Technical energy AI papers often prioritize model interpretation, while AI ethics and trustworthy AI papers emphasize governance principles, and human-centered XAI papers focus on user understanding and interaction [5, 18]. This disciplinary diversity strengthens the review's critical reach, but it also makes direct comparison difficult because the articles do not share a common framework vocabulary. As a result, the synthesis should be understood as an interpretive mapping of framework maturity and research gaps, not as a systematic measurement of field performance.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Limitations of the Review

Scope boundaries and omitted literature

The scope of this review was bounded by the 29-source requirement and the 2017–2023 literature window, which means that some relevant work may have been excluded. The review prioritized peer-reviewed journal articles on XAI, trustworthy AI, smart grids, smart energy systems, and governance, while excluding grey literature, technical standards, policy reports, and many conference papers. This boundary is defensible for an academic review article, but it may underrepresent rapidly evolving regulatory and standardization debates that often appear first outside journal literature [10, 11]. The findings should therefore be read as a critical synthesis of selected peer-reviewed scholarship rather than a complete map of all policy and technical activity.

The review as a snapshot of a rapidly evolving field

This review is also temporally limited because both XAI methods and smart energy digitalization are developing quickly. Energy AI research continues to expand across forecasting, optimization, distributed resources, cyber-physical security, and consumer-facing energy services, while explainability research continues to refine methods, evaluation approaches, and human-centered design principles [1, 12, 19]. The 2017–2023 window captures a formative period in which XAI and trustworthy AI became prominent but were not yet fully integrated into energy governance frameworks. As a result, this review should be understood as a snapshot of an emerging field whose technical, regulatory, and socio-technical foundations remain under active development.

Conclusion

The field of explainable and trustworthy AI for smart energy systems is maturing, but it remains fragmented, technically driven, and insufficiently grounded in real-world governance. The reviewed literature shows strong progress in applying explanation methods to energy forecasting, monitoring, and modeling tasks, yet weaker progress in connecting those methods to stakeholder needs, institutional responsibilities, and regulatory accountability. Trustworthy AI principles provide a valuable normative vocabulary, but they are often not translated into energy-specific design and governance requirements.

Closing these gaps is not merely a research desideratum; it is a prerequisite for safe, fair, and publicly acceptable AI in critical energy infrastructure. Smart energy systems affect grid reliability, consumer participation, market coordination, and social trust in essential services. Explainability must therefore support justified reliance, responsible oversight, and meaningful contestability rather than only technical interpretability.

The most urgent future research directions include energy-specific XAI design, systematic studies of trust calibration, and stronger integration of explainability into regulatory frameworks. Future work should also develop practical evaluation methods for explanation usability, operational relevance, auditability, and lifecycle governance. Such research would help move the field from isolated explanation tools toward mature frameworks for trustworthy energy AI.

Progress will require collaborative and interdisciplinary efforts among AI researchers, energy engineers, social scientists, regulators, utilities, and end-users. No single discipline can fully define what explanation, trust, and accountability should mean in complex smart energy systems. Co-created frameworks are therefore essential for building AI systems that are technically capable, institutionally accountable, socially legitimate, and aligned with the public importance of energy infrastructure.

Acknowledgements

None

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Sep 2022 Revised: 25 Nov 2022 Accepted: 12 Dec 2022
Published online: 15 November 2023

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

1. Ibrahim MS, Dong W, Yang Q. Machine learning driven smart electric power systems: Current trends and new perspectives. Appl Energy. 2020;272:115237.

2. Yohanandhan RV, Elavarasan RM, Manoharan P, Mihet-Popa L. Cyber-physical power system (CPPS): A review on modeling, simulation, and analysis with cyber security applications. IEEE Access. 2020;8:151019-64.

3. 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.

4. Xu C, Liao Z, Li C, Zhou X, Xie R. Review on interpretable machine learning in smart grid. Energies. 2022;15(12):4427.

5. Barredo AA, Del Ser J, Gil-Lopez S, Díaz-Rodríguez N, Bennetot A, Chatila R, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115.

6. Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-60.

7. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Mach. 2018;28(4):689-707.

8. Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. Electron Mark. 2021;31(2):447-64.

9. Hossain E, Khan I, Un-Noor F, Sikander SS, Sunny MS. Application of big data and machine learning in smart grid, and associated security concerns: A review. IEEE Access. 2019;7:13960-88.

10. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389-99.

11. Niet I, van Est R, Veraart F. Governing AI in electricity systems: Reflections on the EU artificial intelligence bill. Front Artif Intell. 2021;4:690237.

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

13. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):1-42.

14. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019;267:1-38.

15. Kuzlu M, Cali U, Sharma V, Güler Ö. Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access. 2020;8:187814-23.

16. Xu C, Li C, Zhou X. Interpretable LSTM based on mixture attention mechanism for multi-step residential load forecasting. Electronics. 2022;11(14):2189.

17. Hirose K. Interpretable modeling for short- and medium-term electricity demand forecasting. Front Energy Res. 2021;9:724780.

18. Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans Interact Intell Syst. 2021;11(3-4):1-45.

19. Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion. 2021;76:89-106.

20. Gürses-Tran G, Körner TA, Monti A. Introducing explainability in sequence-to-sequence learning for short-term load forecasting. Electr Power Syst Res. 2022;212:108366.

21. Ardito C, Deldjoo Y, Di Noia T, Di Sciascio E, Nazary F. Visual inspection of fault type and zone prediction in electrical grids using interpretable spectrogram-based CNN modeling. Expert Syst Appl. 2022;210:118368.

22. Manfren M, James PA, Tronchin L. Data-driven building energy modelling—An analysis of the potential for generalisation through interpretable machine learning. Renew Sustain Energy Rev. 2022;167:112686.

23. Luo Y, Lu C, Zhu L, Song J. Graph convolutional network-based interpretable machine learning scheme in smart grids. IEEE Trans Autom Sci Eng. 2021;20(1):47-58.

24. Xie Y, Ueda Y, Sugiyama M. A two-stage short-term load forecasting method using long short-term memory and multilayer perceptron. Energies. 2021;14(18):5873.

25. del Campo-Avila J, Takilalte A, Bifet A, Mora-Lopez L. Binding data mining and expert knowledge for one-day-ahead prediction of hourly global solar radiation. Expert Syst Appl. 2021;167:114147.

26. Bose BK. Artificial intelligence techniques in smart grid and renewable energy systems—Some example applications. Proc IEEE. 2017;105(11):2262-73.

27. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.

28. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press