REVIEW **Open access**

Trustworthy AI in Renewable and Smart Energy Systems: A Review of Explainability, Fairness, Accountability, and Governance Challenges

Wei Zhang[1*], Chen Hui[2], Michael Tan[1]

Abstract

The integration of artificial intelligence into renewable and smart energy systems offers significant potential to improve forecasting accuracy, demand-side flexibility, building energy control, grid optimisation, and operational decision support. However, as AI becomes increasingly embedded in critical energy infrastructures, its technical performance alone is insufficient to establish trust. Explainability, fairness, accountability, and governance are now essential conditions for ensuring that AI-enabled energy systems are transparent, equitable, responsible, and institutionally reliable. Despite the rapid expansion of AI research in energy forecasting, demand response, building control, and smart-grid applications, the literature remains fragmented, with many studies addressing isolated technical or ethical issues without linking them to a comprehensive trustworthiness framework. This systematic review synthesises 29 peer-reviewed journal articles published between 2017 and 2024 to examine how trustworthy AI has been conceptualised and operationalised in renewable and smart energy systems. Using a PRISMA-guided study selection process, the included studies were thematically classified across four core dimensions: explainability, fairness, accountability, and governance. The findings show that explainability is the most developed area, particularly through post-hoc interpretation techniques and interpretable forecasting approaches. By contrast, fairness and accountability remain comparatively underdeveloped, often receiving limited empirical attention or being treated as secondary concerns. Governance research is also frequently generic, with insufficient adaptation to the specific institutional, regulatory, and operational characteristics of energy systems. The review further identifies cross-dimensional gaps, including weak integration between technical transparency and social justice, limited accountability mechanisms for AI-supported energy decisions, and a lack of empirically validated governance frameworks. Strengthening trustworthy AI in renewable and smart energy systems requires moving beyond isolated model-level solutions toward integrated approaches that connect transparency, equity, responsibility, and oversight. Future research should develop energy-specific trustworthiness frameworks that are empirically testable, operationally implementable, and responsive to evolving regulatory expectations.

Keywords Explainability, Fairness, Smart energy systems, Trustworthy AI, Accountability, Governance

*Correspondence:
Wei Zhang
wei.zhang@unimelb.edu.au
[1] Department of Artificial Intelligence for Sustainable Energy Systems, Faculty of Engineering, University of Melbourne, Melbourne, Australia
[2] Department of Intelligent Grid Optimization and Machine Learning, College of Design and Engineering, National University of Singapore, Singapore, Singapore

Introduction

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.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

AI is increasingly embedded in renewable and smart energy systems through forecasting, demand-side response, grid management, building control, and distributed energy resource coordination. In these settings, AI is not merely a computational tool but part of critical infrastructure where decisions can affect reliability, affordability, and public welfare. The broader literature on explainable and responsible AI has therefore become directly relevant to energy systems, particularly where opaque models are used in operational decisions [1]. Energy-focused reviews similarly show that machine learning is now widely applied to demand response and energy resource management, intensifying the need for trustworthiness [2].

The literature on trustworthy AI in energy remains fragmented because studies often focus on one dimension of trust while neglecting its links with others. XAI research in energy systems has expanded rapidly, especially in forecasting and power-system applications, yet it often treats explanation as a technical layer rather than a governance requirement [3]. Fairness research has developed more slowly, despite evidence that data-driven energy models can reproduce unequal outcomes across buildings, users, and communities [4]. Accountability and governance studies frequently draw on general AI ethics frameworks, but these frameworks are not always adapted to energy-sector institutions, regulatory obligations, or infrastructure risks [5].

A holistic review is needed because explainability, fairness, accountability, and governance are mutually dependent in renewable and smart energy systems. An explanation that is technically accurate may still be insufficient if it does not reveal whether a model distributes benefits and burdens unfairly [6]. A fairness intervention may also be incomplete if affected users cannot contest decisions or if utilities lack audit mechanisms for responsibility attribution [7]. This review therefore treats trustworthy AI as an integrated socio-technical problem rather than a set of isolated model properties.

The objective of this article is to systematically review peer-reviewed literature from 2017 to 2024 on trustworthy AI in renewable and smart energy systems. The review classifies studies according to four pillars: explainability, fairness, accountability, and governance, while also examining their interdependencies. The study selection process follows a PRISMA-like logic and results in 29 included studies, enabling a transparent synthesis rather than an informal narrative review [8]. The article proceeds from background and methodology to thematic synthesis, cross-dimensional gap analysis, review limitations, and future research priorities.

Background and Review Methodology

Trustworthy AI in critical energy infrastructure

Trustworthy AI refers to AI systems that can be understood, evaluated, governed, and held accountable in ways that support legitimate and socially acceptable use. In critical energy infrastructure, this requirement is intensified because AI-enabled decisions may influence grid reliability, renewable integration, demand flexibility, consumer costs, and access to essential services. General AI ethics scholarship identifies transparency, justice, non-maleficence, responsibility, and accountability as recurring principles, but these principles require sector-specific interpretation when applied to electricity systems [5]. Energy AI reviews further show that the technical complexity of renewable forecasting and demand response creates operational conditions where trustworthiness cannot be treated as optional [2].

Dimensions of trustworthy AI: explainability, fairness, accountability, governance

Explainability concerns whether AI decisions can be interpreted by relevant stakeholders, while fairness concerns whether outcomes and procedures avoid unjustified bias or exclusion. Accountability concerns the capacity to assign responsibility, audit decisions, and provide redress, whereas governance concerns the institutional rules, standards, and participatory processes that shape AI deployment. The global AI ethics literature shows that these dimensions are often listed as principles, but principles alone do not ensure practice without implementation mechanisms [9]. In energy systems, this implementation problem is especially important because model outputs are embedded in operational workflows, market arrangements, and consumer-facing decisions [10].

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Systematic review methodology and study selection

The review identified records through targeted searches covering trustworthy AI, explainable AI, fairness, accountability, governance, smart grids, renewable energy, energy forecasting, building energy management, and demand response. Peer-reviewed journal articles published between 2017 and 2024 were eligible when they addressed at least one of the four trustworthy AI dimensions in energy systems or provided a directly relevant conceptual foundation for AI trustworthiness. Records were excluded when they were conference papers, reports, standards, policy documents, non-peer-reviewed items, outside the time window, or unrelated to renewable and smart energy contexts. The review identified 214 database records, removed 48 duplicates, screened 166 titles and abstracts, assessed 64 full-text articles, and included 29 studies, as shown in **Table 1**; this process follows the structured review logic used in prior energy AI and demand-response reviews [2].

Table 1. Study Selection and Screening Process for the Structured Review of Trustworthy AI in Renewable and Smart Energy Systems

Stage	Number of records/articles	Screening decision	Main reason or interpretation
Records identified from database searches	214	Retained for deduplication	Searches combined terms for trustworthy AI, XAI, fairness, accountability, governance, renewable energy, smart grids, and demand response
Duplicate records removed	48	Excluded before screening	Duplicate indexing across multidisciplinary databases
Records screened by title and abstract	166	102 excluded	Exclusions due to non-energy focus, purely technical optimisation without trustworthiness relevance, or non-AI energy governance focus
Full-text articles assessed for eligibility	64	35 excluded	Excluded for conference format, grey literature status, missing DOI, weak relevance to the four dimensions, or publication outside 2017–2024
Studies included in qualitative synthesis	29	Included	Final peer-reviewed journal articles used for thematic synthesis
Final review corpus	29	Classified across four dimensions	Explainability, fairness, accountability, and governance categories were applied to all included studies

Overview of included studies

The included studies were published between 2018 and 2024, with a concentration after 2020, reflecting the rapid expansion of responsible and explainable AI research. Energy-specific studies focused on power systems, renewable forecasting, demand response, building energy management, local energy systems, and energy justice, while broader AI ethics studies provided conceptual grounding for accountability and governance. Explainability was represented by reviews and applied studies in energy forecasting, smart grids, and building energy models [3]. Fairness appeared most explicitly in building energy prediction, thermal control, local energy systems, and energy justice research [4].

Analytical framework for the review

The analytical framework classified each study according to its primary trustworthy AI dimension, energy domain, stakeholder relevance, methodological orientation, and governance implications. Explainability studies were evaluated by

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

whether they used post-hoc, intrinsic, visual, or comparative interpretation methods, while fairness studies were assessed through distributive, procedural, and recognition-based justice categories. Accountability studies were examined through auditability, redress, responsibility attribution, and meaningful human oversight, drawing on broader critiques of AI ethics implementation [11]. Governance studies were assessed according to whether they translated principles into operational mechanisms, institutional responsibilities, or sector-specific requirements [12].

The thematic review framework guiding the synthesis is illustrated in **Figure 1.**

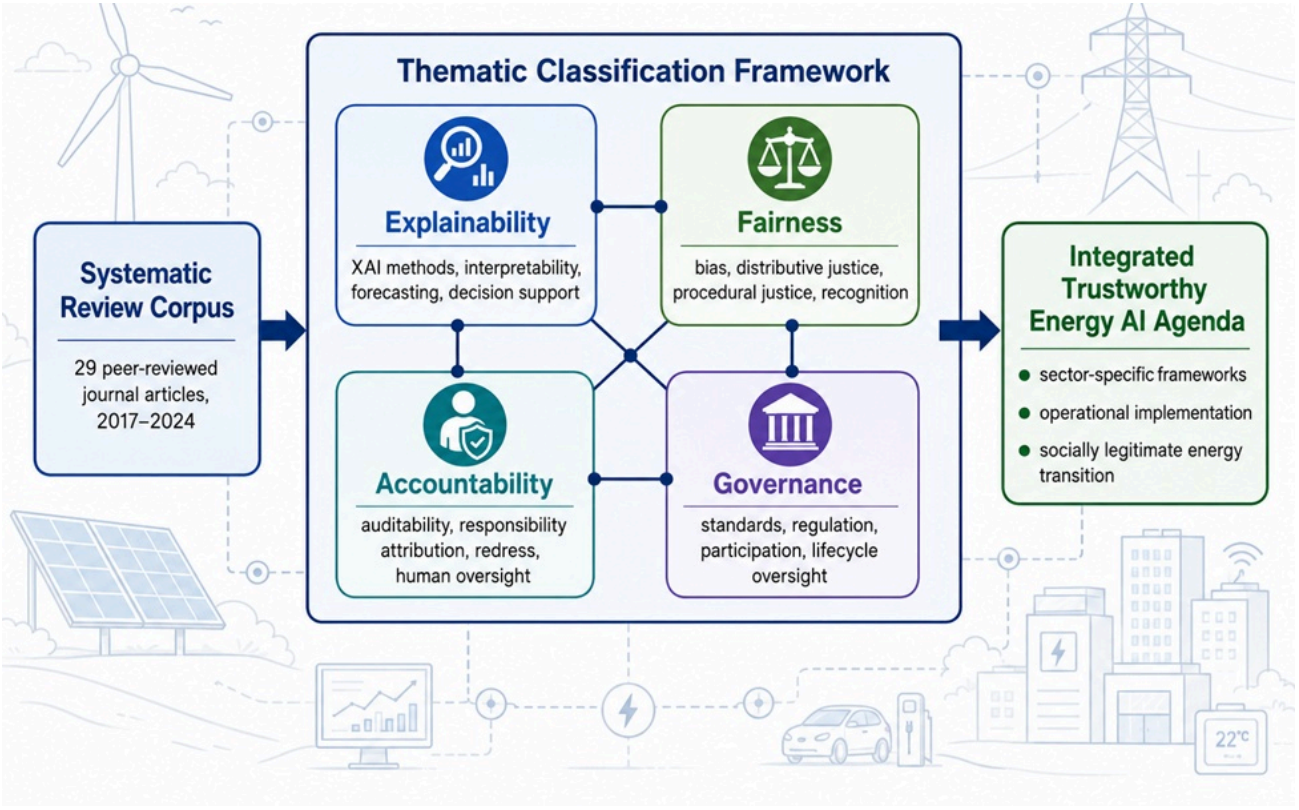

Figure 1. Thematic Review Framework for Trustworthy AI in Renewable and Smart Energy Systems

Trustworthy AI: Dimensions and Energy System Context

The convergence of AI ethics and energy governance

AI ethics and energy governance converge because energy systems increasingly depend on algorithmic decisions while remaining subject to public-interest obligations. General AI ethics frameworks emphasise beneficence, justice, autonomy, explicability, and accountability, yet energy systems add concerns of reliability, affordability, decarbonisation, and infrastructure resilience [5]. Studies on AI ethics guidelines show wide convergence around high-level principles but also considerable variation in how these principles are operationalised [9]. In renewable and smart energy systems, this creates a governance challenge because ethical commitments must be translated into procurement rules, model documentation, operator training, and consumer protection.

Energy system requirements for trustworthy AI

Energy systems impose distinctive requirements on trustworthy AI because model outputs may affect real-time operations, consumer participation, and long-term investment decisions. Renewable forecasting models, for example, must support operational decisions under uncertainty, while demand-response algorithms may influence how flexibility

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

burdens are distributed across households or communities [13]. Smart-grid and building energy applications also involve heterogeneous users, changing weather conditions, legacy infrastructure, and uneven data quality [14]. These features mean that trustworthy energy AI cannot be assessed only by predictive utility; it must also be assessed through interpretability, fairness, auditability, and institutional legitimacy.

The Interplay between dimensions: A conceptual map

The four dimensions of trustworthy AI are best understood as interdependent rather than sequential. Explainability can support accountability by making model reasoning inspectable, but explanations may be inadequate if they do not reveal fairness-relevant patterns or decision consequences [1]. Fairness requires governance because technical bias mitigation alone cannot determine which inequalities are socially acceptable or how trade-offs should be justified [15]. Accountability, in turn, depends on governance arrangements that define who must document, audit, contest, and correct AI-supported energy decisions [7].

Explainability in Energy AI

Prevalence and types of explainability methods

Explainability is the most visible trustworthy AI dimension in the reviewed energy literature. Energy XAI research frequently uses post-hoc methods such as SHAP, LIME, feature importance, sensitivity analysis, and visual explanation tools to interpret forecasting, load prediction, and operational decision models [3]. Reviews of interpretable machine learning in smart grids show that these techniques are attractive because they can be added to high-performing models without replacing them [16]. However, the wider XAI literature warns that post-hoc explanations can create an appearance of transparency without necessarily making the underlying model genuinely interpretable [17].

Explaining energy decisions: use cases and stakeholders

The reviewed studies apply explainability to several energy decision contexts, including renewable energy forecasting, electricity load prediction, building energy management, demand response, and smart-grid monitoring. In forecasting studies, explanations often help analysts understand which weather, temporal, or operational variables shape predictions [18]. In building energy management, interpretable methods are used to clarify how occupancy, indoor conditions, weather, and system settings influence energy outcomes [14]. These explanations may be useful for operators, engineers, facility managers, regulators, and consumers, but the literature often gives more attention to technical users than to affected publics.

Limitations and gaps in energy XAI research

The energy XAI literature remains limited by insufficient user testing, weak integration with governance, and limited evidence of real-world decision impact. Some studies compare explanation techniques in energy forecasting, but they rarely examine whether different stakeholders interpret or act upon explanations appropriately [18]. Building energy studies have begun to develop intrinsically interpretable models, suggesting an alternative to black-box explanation, yet such approaches are not yet dominant across energy AI research [19]. As shown in **Table 2**, the reviewed literature contains a range of explanation methods, but each method faces limitations when transferred from model interpretation to operational trustworthiness.

Table 2. Summary of Explainability Methods in Energy AI Systems, Applications, Stakeholders, and Key Limitations

Explainability method	Energy application	Explanation type	Stakeholder	Key limitation

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

SHAP-based feature attribution	Load forecasting, building energy prediction, renewable forecasting	Local and global contribution of input variables	Data scientists, energy analysts, facility managers	Often explains correlations without proving causal mechanisms or operational relevance
LIME-style local approximation	Smart-grid prediction and demand-side response models	Local surrogate explanation for individual predictions	Operators and model reviewers	Sensitive to perturbation design and may be unstable across similar cases
Feature importance ranking	Renewable energy forecasting and energy resource management	Global ranking of influential predictors	Engineers and planners	May obscure subgroup-specific effects and fairness-relevant differences
Grad-CAM and visual explanation	Time-series energy forecasting	Visual saliency over temporal or structured inputs	Technical analysts	Interpretation may be difficult for non-expert stakeholders
Intrinsically interpretable models	Building energy load prediction	Transparent model structure and readable decision logic	Facility managers and operational decision-makers	May require trade-offs in model flexibility or domain-specific design effort
Sensitivity and scenario explanation	Building control and energy management	How outputs change under altered inputs or operating conditions	Operators, regulators, and auditors	Can support reasoning but may not identify responsibility for harmful decisions
Physics-informed explanation	Renewable energy forecasting	Alignment between model behaviour and physical system knowledge	Energy engineers and system operators	Requires strong domain assumptions and may not address fairness or accountability
Taxonomy-based XAI review synthesis	Power and energy systems	Classification of explanation families and use cases	Researchers and governance designers	Provides conceptual structure but limited direct implementation guidance

Fairness and Bias Challenges

Mapping fairness issues in energy AI

Fairness is less developed than explainability in the reviewed literature, but it is increasingly recognised as essential for trustworthy energy AI. Building energy studies show that data-driven models may produce uneven prediction accuracy across subgroups, building types, or environmental conditions, raising concerns about unequal model reliability [4]. Local energy systems and demand-response research also suggest that algorithmic coordination can distribute benefits and burdens unevenly among participants [20]. Energy justice scholarship extends this concern by warning that machine learning bias may reinforce existing inequities in access, affordability, and recognition [10].

Types of fairness: distributive, procedural, and recognition

The reviewed literature can be classified into distributive, procedural, and recognition-based fairness concerns. Distributive fairness concerns who receives energy benefits, cost savings, flexibility payments, or prediction errors, and it is visible in studies on fairness-aware building models and local energy systems [6]. Procedural fairness concerns whether affected users can understand, challenge, or influence AI-supported decisions, connecting fairness to

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

explainability and accountability [12]. Recognition fairness concerns whether vulnerable or underrepresented groups are properly represented in data, model assumptions, and governance processes, an issue highlighted by energy justice research on machine learning bias [10].

The fairness–energy justice nexus

The fairness–energy justice nexus shows that technical fairness metrics alone cannot resolve the normative questions raised by AI in renewable and smart energy systems. Fairness-aware machine learning studies can reduce uneven model performance across defined groups, but they do not automatically determine which groups should be protected or how trade-offs should be governed [21]. Research on fairness-aware model predictive control further illustrates that fairness can be incorporated into energy optimisation, yet such approaches require institutional decisions about acceptable balances between efficiency and equity [22]. **Table 3** therefore categorises fairness issues not only by model type but also by the justice implications they create in energy contexts.

Table 3. Typology of Fairness Dimensions and Associated Energy Justice Implications in AI-Enabled Energy Systems

Fairness type	Energy context	AI system involved	Justice implication	Representative studies
Distributive fairness	Building energy prediction	Regression-based data-driven building models	Unequal prediction errors may cause some buildings or users to receive poorer control or planning decisions	Sun, Haghighat and Fung [4]; Sun, Fung and Haghighat [6]
Distributive fairness	Thermal energy storage and residential control	Fairness-aware model predictive controller	Optimisation may prioritise efficiency while unevenly distributing comfort or energy flexibility burdens	Sun, Haghighat and Fung [22]
Procedural fairness	Demand response and local energy systems	AI-enabled flexibility allocation and coordination algorithms	Users may be affected by automated flexibility decisions without sufficient explanation, consent, or contestability	Antonopoulos and colleagues [2]; Soares and colleagues [20]
Recognition fairness	Energy justice and smart-energy data systems	Machine learning models trained on unevenly representative data	Vulnerable communities may be misclassified, ignored, or underserved when data fail to capture social context	Chen and colleagues [10]
Representational fairness	Smart-grid and building energy analytics	Forecasting and classification models using historical consumption data	Historical patterns may encode structural inequalities and reproduce them through automated decision support	Mehrabi and colleagues [15]; Ntoutsi and colleagues [23]
Governance-linked fairness	Renewable and smart energy planning	AI systems used in policy, forecasting, and operational planning	Fairness claims require institutional oversight because technical mitigation cannot define legitimate energy priorities	Floridi and colleagues [5]; Jobin, Ienca and Vayena [9]

Accountability and Responsibility

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Defining accountability in AI-enabled energy decisions

Accountability in AI-enabled energy decisions refers to the ability to identify who is responsible for model design, deployment, monitoring, correction, and consequences. In smart energy systems, accountability is complicated because decisions may emerge from interactions among utilities, software vendors, grid operators, aggregators, regulators, and automated optimisation systems. General AI accountability scholarship shows that responsibility cannot be reduced to technical transparency alone, because institutions must also define duties, liability, and enforceable oversight [24]. In the energy context, this means that explainable forecasting or control systems remain incomplete unless they are connected to accountable decision pathways.

Mechanisms for algorithmic auditing and redress

The review identified auditability as a recurring but underdeveloped requirement for trustworthy AI in energy systems. AI ethics implementation research shows that tools such as impact assessments, documentation practices, model cards, audit trails, and review procedures can help translate principles into practice [12]. However, energy-specific studies rarely specify how affected consumers, communities, or operators could challenge AI-supported decisions in demand response, building control, or smart-grid optimisation. This gap is particularly important because biased or opaque models may affect access to flexibility incentives, energy services, or operational priority without clear redress mechanisms [10].

Human oversight and the retention of meaningful control

Human oversight is central to accountability because energy AI systems often support decisions in contexts where full automation may be unsafe, inequitable, or institutionally illegitimate. The literature on interpretable AI warns that black-box explanations may not be sufficient in high-stakes settings and that inherently interpretable systems may be preferable where decisions require justification [17]. Energy XAI studies similarly show that explanations can support operator understanding, but they do not automatically ensure meaningful human control if operators lack authority, training, or time to intervene [3]. Therefore, trustworthy energy AI requires both technical interpretability and organisational arrangements that preserve accountable human judgment.

Table 4 synthesises the accountability mechanisms, responsible actors, and implementation gaps identified across the reviewed literature.

Table 4. Accountability Mechanisms in AI-Enabled Renewable and Smart Energy Systems: Actors, Instruments, and Identified Gaps

Accountability stage	Mechanism or instrument	Responsible actor(s)	Identified gap in literature
Design and development	Model documentation, impact assessment, fairness audit	AI developers, utilities, research institutions	Few energy-specific impact assessment frameworks exist; general AI ethics tools not adapted for grid operations [12]
Deployment and operation	Audit trails, performance monitoring, real-time oversight	Grid operators, energy service providers	Auditability requirements specified without operational procedures for energy contexts [3]
Decision review and contestation	Redress mechanisms, consumer complaint procedures, independent review	Regulators, consumer advocates, dispute resolution bodies	Affected users lack clear pathways to challenge AI-supported demand-response or flexibility decisions [10]

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Incident response and liability	Post-hoc investigation, liability attribution, corrective action	Regulatory authorities, legal systems, utilities	Attribution of responsibility diffused across vendors, operators, and automated systems [24]
Human oversight	Operator training, authority to override, escalation procedures	System operators, control room personnel	Meaningful control requires organisational authority, not just technical interpretability [17]

Attribution of liability in autonomous energy systems

Attribution of liability becomes difficult when AI systems act across distributed assets, forecasting pipelines, automated markets, and demand-side resources. Reviews of AI in sustainable energy and energy systems show that the field increasingly integrates machine learning into complex decision environments, but legal responsibility frameworks have not advanced at the same pace [25]. Bibliometric and strategic reviews also indicate rapid growth in AI energy applications, suggesting that responsibility questions will become more urgent as AI moves from analytical support to operational control [26]. The review therefore identifies liability attribution as a major gap linking accountability to governance.

Governance Challenges and Frameworks

Regulatory and policy gaps

Existing energy regulation was not designed for AI systems that learn from data, adapt over time, and influence distributed decision-making. General AI ethics research shows that many guidelines converge around principles but provide limited mechanisms for enforcement, monitoring, or sector-specific accountability [9]. In renewable and smart energy systems, this creates a regulatory gap because AI-enabled forecasting, demand response, and building control can affect reliability and equity without being governed as consequential algorithmic decisions. Energy AI reviews further suggest that policy frameworks have lagged behind the expanding technical literature on AI for forecasting and system optimisation [27].

Emerging governance frameworks for trustworthy energy AI

Emerging governance frameworks for trustworthy energy AI remain fragmented across AI ethics, smart-grid management, energy justice, and sustainability transitions. Some studies propose broad ethical principles for responsible AI, while others focus on operational XAI, fairness-aware modelling, or energy justice applications [5]. Research on AI ethics tools shows that implementation requires methods, procedures, and institutional workflows rather than abstract commitments alone [12]. In the energy domain, the review identified a need for governance frameworks that connect model documentation, stakeholder participation, fairness assessment, operator oversight, and post-deployment monitoring.

The role of standards and certification

Standards and certification could help make trustworthy AI requirements more consistent across renewable and smart energy systems, but the reviewed literature rarely evaluates how such instruments should be adapted to energy-specific applications. AI ethics critiques argue that principles alone cannot guarantee ethical outcomes when organisations lack enforceable procedures, accountability mechanisms, or independent evaluation [7]. Similarly, evaluations of AI ethics guidelines show that many frameworks emphasise transparency and fairness while providing limited operational detail for certification, auditing, or sector-specific compliance [11]. **Table 5** integrates these gaps by showing how weaknesses in explainability, fairness, and accountability create governance challenges requiring coordinated responses.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Table 5. Trustworthy AI Governance Gaps in Energy Systems and Corresponding Risks and Regulatory Responses

Trustworthiness gap	Governance implication	Energy-system risk created	Governance response needed
Explanations remain technical and model-centred	Operators or regulators may receive outputs that are difficult to interpret for real decisions	Misuse of forecasts, overconfidence in black-box models, or weak operational justification	Require stakeholder-specific explanation documentation and operational interpretability testing
Fairness is treated as a model metric only	Equity concerns may be separated from energy justice, affordability, and access	Unequal distribution of flexibility burdens, comfort impacts, or prediction errors	Link fairness audits to energy justice assessment and affected-user consultation
Accountability is not assigned across the AI lifecycle	Responsibility may be diffused among vendors, utilities, operators, and regulators	Harmful decisions become difficult to challenge or correct	Define lifecycle accountability roles for design, deployment, monitoring, and redress
Governance frameworks remain generic	AI ethics principles may not fit grid reliability, renewable uncertainty, or demand-response contexts	Trustworthy AI becomes symbolic rather than enforceable in energy operations	Develop energy-specific certification and audit criteria for high-impact AI uses
Human oversight is nominal rather than meaningful	Human operators may be responsible without practical capacity to intervene	Automation bias, deskilling, or delayed correction of harmful model behaviour	Establish authority, training, and escalation procedures for human intervention
Cross-dimensional integration is weak	Explainability, fairness, and accountability are assessed separately	Systems may pass one trust criterion while failing another	Use integrated trustworthy AI impact assessments before deployment and during monitoring

Multi-stakeholder and participatory governance

Participatory governance is necessary because renewable and smart energy AI affects diverse stakeholders, including utilities, regulators, technology providers, operators, consumers, and vulnerable communities. Energy justice research shows that bias in machine learning cannot be addressed only through technical correction because affected groups must also be recognised in data, design, and governance processes [10]. Fairness research in local energy systems similarly suggests that algorithmic coordination can reshape participation, benefits, and burdens across communities [20]. Trustworthy energy AI therefore requires multi-stakeholder governance that connects technical evaluation with democratic legitimacy.

Synthesis and Cross-Cutting Gaps

The fragmentation challenge

The review identified fragmentation as the central weakness in the literature on trustworthy AI in renewable and smart energy systems. Explainability studies are relatively mature but often focus on technical interpretation without linking explanations to fairness, accountability, or governance [16]. Fairness studies demonstrate the possibility of fairness-aware modelling, yet they remain concentrated in building energy and local energy contexts rather than the full range of

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

smart-grid and renewable applications [21]. Accountability and governance studies provide essential principles, but they often remain general and do not fully specify how trustworthy AI should be implemented in energy-sector institutions [24].

Toward an integrated trustworthy AI agenda for energy

An integrated trustworthy AI agenda for energy should treat explainability, fairness, accountability, and governance as mutually reinforcing design and oversight requirements. Renewable forecasting and energy resource management studies show that AI will continue to support increasingly complex operational decisions, making trustworthiness central to deployment rather than an external ethical add-on [28]. Future frameworks should therefore combine interpretable modelling, fairness assessment, auditability, human oversight, participatory governance, and sector-specific certification. Such integration would move the field from isolated technical improvements toward accountable and socially legitimate AI for the energy transition.

Evaluation of the Review's Scope and Limitations

Strengths of the systematic approach

The strength of this review lies in its structured selection process, defined 2017–2024 time window, and explicit classification of studies across four trustworthy AI dimensions. The PRISMA-like process made the pathway from identified records to the final 29 studies transparent, reducing the risk of an unstructured narrative synthesis. The review also combined energy-specific studies with foundational AI ethics and fairness literature, which was necessary because accountability and governance are not yet fully developed within energy journals [5]. This approach enabled the review to evaluate both technical and institutional aspects of trustworthy AI.

Inherent limitations of a 29-study sample

The 29-study sample allowed focused synthesis but necessarily limited the breadth of included perspectives. Some relevant debates on standards, regulation, and AI governance may appear in reports, policy documents, or technical standards, but these were excluded to maintain the peer-reviewed journal-only criterion. This exclusion is important because AI ethics tools and governance practices often develop outside conventional journal articles before being evaluated academically [12]. The findings should therefore be interpreted as a synthesis of peer-reviewed scholarship rather than a complete map of all policy and technical governance activity.

Subjectivity in thematic classification

The thematic classification involved interpretive judgment because many studies address more than one trustworthy AI dimension. For example, XAI studies can support accountability, fairness studies can imply governance needs, and AI ethics frameworks can inform all four dimensions simultaneously [1]. Bias and fairness surveys also show that concepts such as representation, discrimination, and accountability overlap across technical and social domains [23]. The classification was therefore used as an analytical device rather than a claim that each study belongs exclusively to one category.

Limitations of the Review

Temporal and disciplinary boundaries

The 2017–2024 time window captures the recent expansion of AI trustworthiness debates but excludes earlier work on energy justice, automation governance, and power-system accountability that may indirectly inform the topic. The focus on peer-reviewed articles also prioritised academic literature over standards, regulatory documents, and practitioner guidance, even though these sources are important for real-world governance. Reviews of AI in renewable energy and

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

sustainable energy show that the field has grown rapidly during this period, making the selected window appropriate but not exhaustive [27]. As a result, the review should be read as a structured synthesis of recent scholarship rather than a complete historical account.

Temporal and disciplinary boundaries also mean that the review captures only a snapshot of a field where regulatory frameworks are evolving concurrently with technical capabilities. Recent analyses of EU legislation applicable to smart-grid AI, including the AI Act, GDPR, and NIS2 Directive, demonstrate that legal requirements for transparency, data protection, and cybersecurity are still being interpreted for energy-sector implementation [29]. Governance scholarship emphasises that AI ethics principles must be translated into enforceable mechanisms, yet energy-specific regulatory guidance remains underdeveloped relative to the pace of AI deployment [30]. These regulatory dynamics suggest that trustworthy AI requirements for renewable and smart energy systems will continue to shift as both technical standards and legal frameworks mature.

Challenges of a rapidly evolving field

Trustworthy AI in energy is evolving quickly because both AI capabilities and energy governance challenges are changing. New foundation models, digital twins, autonomous agents, and distributed energy platforms may introduce trustworthiness concerns that were not fully represented in the 2017–2024 literature. Reviews of renewable forecasting and AI-based energy resource management already show that technical innovation is moving faster than governance frameworks [13]. This temporal instability means that future reviews will need to update the evidence base as regulatory expectations, AI tools, and energy system architectures continue to change.

Table 6 summarises the key governance barriers to AI adoption in smart energy systems identified in recent regulatory analyses.

Table 6. Regulatory and Governance Barriers to Trustworthy AI Adoption in Smart Energy Systems

Regulatory instrument	Applicability to energy AI	Compliance challenge identified	Energy-specific requirement
EU AI Act (risk-based classification)	High-risk applications including critical infrastructure, demand-response, grid management	Determining risk category for forecasting and optimisation models; conformity assessment procedures	Sector-specific guidance for high-risk energy AI applications needed [29]
General Data Protection Regulation (GDPR)	Smart-meter data, consumer behaviour profiling, load forecasting	Right to explanation for automated decisions; data minimisation vs. model accuracy trade-offs	Energy sector interpretation of automated decision-making rights for flexibility services [29]
NIS2 Directive (cybersecurity)	Grid monitoring, real-time control, distributed energy resource coordination	Incident reporting timelines; supply chain security for AI components	Integration of AI-specific vulnerabilities into cybersecurity risk assessments [30]
EU Data Act	Data sharing across energy platforms, interoperability requirements	Access rights to AI-generated data; switching between data-processing services	Governance frameworks for AI training data in multi-utility contexts [29]
EU Cyber Resilience Act	AI-enabled hardware and software in energy infrastructure	Security by design requirements for AI products; vulnerability handling	Energy sector conformity assessment for AI-enabled grid components [30]

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Conclusion

This review found that the four dimensions of trustworthy AI in renewable and smart energy systems have developed unevenly. Explainability is the most visible and technically mature dimension, while fairness, accountability, and governance remain less integrated into energy AI research. The literature shows a strong need to move beyond isolated model interpretation toward broader socio-technical trustworthiness.

The review also showed that energy systems require sector-specific trustworthy AI frameworks. Renewable and smart energy systems involve operational reliability, public welfare, infrastructure safety, affordability, and justice concerns that cannot be addressed by generic AI ethics principles alone. Explainability, fairness, accountability, and governance should therefore be treated as inseparable requirements.

Future work should develop and validate integrated trustworthy AI frameworks for renewable and smart energy systems. This requires collaboration among AI researchers, energy engineers, regulators, social scientists, utilities, and affected communities. Building trustworthy AI for the energy transition will depend not only on better models but also on accountable institutions, inclusive governance, and regulatory innovation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Jun 2023 Revised: 14 Aug 2023 Accepted: 30 Aug 2023
Published online: 15 November 2024

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

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

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

2. Antonopoulos I, Robu V, Couraud B, Kirli D, Norbu S, Kiprakis A, et al. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. Renew Sustain Energy Rev. 2020;130:109899.

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. Sun Y, Haghighat F, Fung BC. Trade-off between accuracy and fairness of data-driven building and indoor environment models: A comparative study of pre-processing methods. Energy. 2022;239:122273.

5. 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.

6. Sun Y, Fung BC, Haghighat F. The generalizability of pre-processing techniques on the accuracy and fairness of data-driven building models: A case study. Energy Build. 2022;268:112204.

7. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1(11):501-7.

8. Alsaigh R, Mehmood R, Katib I. AI explainability and governance in smart energy systems: A review. Front Energy Res. 2023;11:1071291.

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

10. Chen CF, Napolitano R, Hu Y, Kar B, Yao B. Addressing machine learning bias to foster energy justice. Energy Res Soc Sci. 2024;116:103653.

11. Hagendorff T. The ethics of AI ethics--an evaluation of guidelines. arXiv:1903.03425. 2019.

12. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In: Ethics, governance, and policies in artificial intelligence. Cham: Springer; 2021. p. 153-183.

13. Haupt SE, McCandless TC, Dettling S, Alessandrini S, Lee JA, Linden S, et al. Combining artificial intelligence with physics-based methods for probabilistic renewable energy forecasting. Energies. 2020;13(8):1979.

14. Chen Z, Xiao F, Guo F, Yan J. Interpretable machine learning for building energy management: A state-of-the-art review. Adv Appl Energy. 2023;9:100123.

15. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):1-35.

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

17. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.

18. Van Zyl C, Ye X, Naidoo R. Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. Appl Energy. 2024;353:122079.

19. Zhang C, Hoes PJ, Wang S, Zhao Y. Intrinsically interpretable machine learning-based building energy load prediction method with high accuracy and strong interpretability. Energy Built Environ. 2024;(in press).

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

20. Soares J, Lezama F, Faia R, Limmer S, Dietrich M, Rodemann T, et al. Review on fairness in local energy systems. Appl Energy. 2024;374:123933.

21. Sun Y, Fung BC, Haghighat F. In-Processing fairness improvement methods for regression Data-Driven building Models: Achieving uniform energy prediction. Energy Build. 2022;277:112565.

22. Sun Y, Haghighat F, Fung BC. Fairness-aware data-driven-based model predictive controller: A study on thermal energy storage in a residential building. J Energy Storage. 2024;87:111402.

23. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Min Knowl Discov. 2020;10(3):e1356.

24. McGregor L. Accountability for governance choices in artificial intelligence: afterword to Eyal Benvenisti's foreword. Eur J Int Law. 2018;29(4):1079-85.

25. Saheb T, Dehghani M, Saheb T. Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis. Sustain Comput Inform Syst. 2022;35:100699.

26. Entezari A, Aslani A, Zahedi R, Noorollahi Y. Artificial intelligence and machine learning in energy systems: A bibliographic perspective. Energy Strategy Rev. 2023;45:101017.

27. Zhang L, Ling J, Lin M. Artificial intelligence in renewable energy: A comprehensive bibliometric analysis. Energy Rep. 2022;8:14072-88.

28. Krishnamurthy S, Adewuyi OB, Luwaca E, Ratshitanga M, Moodley P. Artificial intelligence-based forecasting models for integrated energy system management planning: An exploration of the prospects for South Africa. Energy Convers Manag X. 2024;24:100772.

29. Rajca M. Privacy risks and regulatory challenges in smart grids and renewable energy systems. Internetowy Kwart Antymonop Regul. 2024;13(2):9-25.

30. Raihan A. A comprehensive review of artificial intelligence and machine learning applications in energy sector. J Technol Innov Energy. 2023;2(4):1-26.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press