Journal of Engineering Innovation and Global Impact (JEIGI)
Volume 1, Issue 1 | PP 1-14 | https://jeigionline.com/ |

Submitted: 13-05-2026 | Revised: 31-05-2026 | Accepted: 28-06-2026 | Published: 30-06-2026

Explainable AI for Human-Centric Smart Grids: A Conceptual Framework for Trustworthy Renewable Energy Systems

Dr Neelam Goyal

Dean - Behavioural Sciences
Sardar Patel University, Sikkim, India
Email: dr.neelamgoyal15@gmail.com

Abstract

From probabilistic solar forecasting and wind prediction, to real-time dispatch, fault detection and demand response orchestration, the use of artificial intelligence (AI) in the modern smart grid is an indispensable necessity. But these predictive benefits of deep and ensemble learning have come at the expense of interpretability, leading to a legitimacy gap in a publicly accountable and safety-critical infrastructure that requires operators to defend switching decisions, regulators to examine market performance and prosumers to agree to automated control of their assets. The current power systems explainable AI (XAI) literature is mostly technique-driven and assesses the quality of explanations using algorithmic fidelity, but not in relation to the impact of explanations on humans who have to act on them. In this paper, a conceptual framework is proposed called Human-Centric Explainable Grid (HX-Grid) that proposes a shift of explainability from being a post-hoc property of a model to being a sociotechnical property of the grid. Adopting a design science and theory-synthesis approach, the study brings together four streams of literature (AI in renewable-integrated grids, foundations of XAI, trust and human factors, and energy governance) and proposes an architecture that encompasses the physical context and modelling, the generation of the explanation, the adaptation to the perspective of stakeholders, and the feedback from governance. The framework is expanded using a matrix of stakeholders and an explanation, six propositions, and a protocol for assessing the constructs and instruments for later empirical testing, which will be based on placeholders. The contribution reflects a theoretical and design aspect, without any empirical data being reported. The implications for grid operators, technology vendors and regulators of new AI accountability frameworks are explored, as well as a research agenda on trust calibration in the energy sector.

Keywords: explainable artificial intelligence; smart grids; human-centric design; renewable energy integration; trust calibration; sociotechnical systems

1. Introduction

Electricity supply has evolved from a centrally dispatched deterministic machine to a decentralised stochastic and data saturated network, due to the decarbonisation of electricity supply. PV and wind generation are non-dispatchable and weather dependent, storage assets are time arbitragers, EVs are loads, resources and mobile constraints and millions of prosumer assets are at the distribution edge with ability to respond to price or control signals (Gungor et al., 2011; Tuballa & Abundo, 2016). This

complexity in real time is beyond rule-based control and the sector has taken to machine learning very fast. In short-term renewable forecasting, deep neural architectures are now prevalent and in voltage regulation and market bidding, reinforcement learning is used and in fault detection and non-technical loss identification, ensemble classifiers is used (Ahmad et al., 2022).

This absorption has made a paradox. The best models are the most "hard to read". A forecaster using gradient boosting or a recurrent network could make a significant reduction in error compared to a persistence forecaster, but not explain why it thinks a curtailment event is likely at 14:00. Opacity is acceptable in a recommendation where there is not a high stakes. In an electricity network it isn't, for three reasons. First, consequences are physical, not in a user interface, and irreversible: a mis-timed switching action cascades through hardware. Second, that all tariffs, curtailments and connection refusal decisions are subject to timely contestation in regulatory and legal proceedings and that the defence "the model said so" is not a defence (Machlev et al., 2022). Third, the system now relies on the voluntary participation of households and small firms whose assets are being coordinated, which is dependent on the perceived fairness and comprehensibility of the system, but not on accuracy of the validation set.

Explainable AI has been suggested as a solution. It has grown very quickly, yielding model-agnostic methods for attribution (Adadi & Berrada, 2018), surrogate models, counterfactual generators, and even inherently interpretable architectures (Arrieta et al., 2020). Its introduction into energy research has, however, mostly retained the original thrust of the field: the generation of explanations for a model, and testing this explanation against properties of that model fidelity, stability, sparsity, computational cost. The recipient is used as a generic consumer of information and not as a specific actor with a role, a decision, a time budget and a liability exposure. Different explanations of the same, underlying model are needed for a control-room engineer who is trying to solve a contingency in 90 seconds, a distribution planner evaluating a 5-year reinforcement scenario, a regulator who is on an audit of a year of dispatch, and a householder who wants to give up his thermostat control. The one thing that is wrong with current practice, and that is the feeling that they are interchangeable, is the reason for a field observation that occurs repeatedly: explanations are created and left out of the field.
This paper, then, poses three questions. RQ1: What needs to be included in the structure of the smart grid AI system so that explainability becomes a characteristic of the sociotechnical system and not just a characteristic of an AI system in isolation? RQ2: What are the differences in content, form and timing of the explanations for the various heterogeneous stakeholders of a renewable-integrated grid? RQ3: How does explainability become a form of calibrated trust reliance according to the actual system competence instead of trust?
The paper has three contributions. It extends the HX-Grid conceptual framework, which is an architecture that combines technical and human aspects. It outlines a stakeholder–explanation matrix which translates differentiation in terms of 5 classes of actors. It also produces six propositions, along with an accompanying evaluation protocol, making the framework more than just descriptive and empirically testable. The paper is clearly conceptual – no experimental or survey data is presented.

2. Literature Review

2.1 Artificial Intelligence in Renewable-Integrated Smart Grids

AI applications in power systems are grouped into four categories: The biggest forecasting models are statistical and deep sequence models, which provide estimates of irradiance, wind speed, and net load, ranging from minutes to days, and with probabilistic outputs to aid in reserve procurement. The second is operational control (RL and model-predictive hybrids) in which voltage, frequency, and storage dispatch are managed under uncertainty. The third is asset and anomaly management, which relies on classifiers for equipment failure detection, for cyber-intrusion detection and for theft

detection. The fourth is market and demand-side coordination, which involves learning agents bidding, aggregation and flexibility scheduling (Ahmad et al., 2022; Tuballa & Abundo, 2016). All four have reported a trend of growing model capacity and decreasing transparency, and evaluation is still dominated by point-error metrics, which are uninformative regarding the ability of a human to act on the output.

2.2 Foundations and Taxonomies of Explainable AI

The XAI literature mainly has two types of models: intrinsically interpretable models and post-hoc explanation of opaque models, and, within opaque models, post-hoc explanation of global models and local models (Adadi & Berrada, 2018; Arrieta et al., 2020). Examples of dominant instruments are local surrogate approximation (Ribeiro et al., 2016), cooperative game theory-based additive feature attribution (Lundberg & Lee, 2017), counterfactual explanation: the smallest input change that would alter the output (Wachter et al., 2017), and gradient-based saliency for network architectures (Selvaraju et al., 2017). Another criticism is that post-hoc explanation of an opaque model is an approximation and is not a true account of the model; and in high-stakes domains, models that are inherently interpretable should be favored if they perform equally well (Rudin, 2019). The parallel criticism focuses on the notion of evaluation and its lack of definition, as well as its lack of reference to any task a person has to perform (Doshi-Velez & Kim, 2017).

2.3 Explainable AI in Power and Energy Systems

The use of XAI in the energy sector has been increasing rapidly. Reviews map attribution and surrogate techniques to forecasting, stability assessment and fault diagnosis and consistently conclude that the same issues are identified: Lack of domain-specific evaluation criteria, limited attention to physical constraints, and no attention to the operator (Machlev et al., 2022). Illustrative studies use the attribution methods of PV generation forecasting to reveal the meteorological causes of model predictions and show the technical viability of the method without evaluating the decision impact (Kuzlu et al., 2020). In the medical field, medicine-related metrics are used to assess the quality of the explanations, as in a cross domain review (Tjoa & Guan, 2021). The outcome for energy research is technique without consequence, literature.

2.4 Trust, Human Factors, and Explanation as a Social Process

Human factors research provides the antidote. The design goal is not to increase trust in automation, but to design the system so that it is calibrated (aligned) with the trust placed in it; miscalibration occurs when a reliable system is not used, or when an unreliable system is used complacently (Lee & See, 2004; Parasuraman & Riley, 1997). Calibration is one of the main levers, and in this case it is explanation providing boundaries of competence rather than an impression of uniformity of authority. The contrastive, selective, and social nature of explanation is further understood as people who ask why this rather than that, challenge limited number of causes, and understand explanation as an exchange between parties (Miller, 2019). There are proposed and validated measurement instruments for explanation satisfaction and trust in XAI contexts (Hoffman et al., 2018) and formal treatments have separated warranted from unwarranted trust and insisted that the trustworthiness be established apart from the illusion of transparency (Jacovi et al., 2021). This is a stream that is not prevalent in the energy sector XAI work.

2.5 Governance and Regulatory Drivers

In the regulatory era, explainability is a compliance matter and no longer a design choice. Transparency, human oversight, logging and documentation requirements for AI systems deployed in

critical infrastructures (European Commission, 2021; Goodman & Flaxman, 2017), while data protection regimes also give rights associated with automated decision-making. Justifications for curtailment, connections and tariffs are required separately by the energy regulators. Explainability, therefore, is finally becoming an evidence requirement: When an automated action is made, an operator shall have the ability to reconstruct, a posteriori, the basis for that action.

2.6 Synthesis and Research Gap

The four streams are combined in Table 1. The gap is not an incremental one, but a structural one, in that the technical stream is explanation supply, not demand; the human factors stream is demand, not technical; the governance stream imposes obligations, not a design vocabulary. There is no integrative framework that links them for networks integrated with renewable energy.

Table 1. *Synthesis of literature streams and identified gaps*

Stream	Core contribution	Principal limitation for smart grids	Representative sources
AI in smart grids	Accuracy gains in forecasting, control, and diagnosis	Transparency treated as external to the modelling task	Gungor et al. (2011); Ahmad et al. (2022)
XAI foundations	Attribution, surrogate, counterfactual, and interpretable-by-design methods	Recipient modelled generically; evaluation proxy-based	Arrieta et al. (2020); Rudin (2019)
XAI in energy	Domain feasibility demonstrations	Technique-centric; no decision-impact evidence	Machlev et al. (2022); Kuzlu et al. (2020)
Trust and human factors	Calibration, contrastive explanation, validated instruments	Not connected to physical or market constraints of grids	Lee & See (2004); Miller (2019)
Governance	Transparency and oversight obligations	No design vocabulary for compliance	European Commission (2021)

3. Research Methodology

3.1 Research Design

The study is conceptual and design based study. It abides by the design science research paradigm – for which knowledge production is not observed but constructed and substantiated with artefacts (Hevner et al., 2004; Peffers et al., 2007). The paper belongs to the taxonomy of conceptual article types, in which the paper follows a theory synthesis approach that brings together previously disjoint literatures in an integrated conceptual structure (Jaakkola, 2020). This design is suitable as the research problem is not one of absence of measurement, but one of conceptual disorganisation: the constituent knowledge is present but uninstitutionalised, and early empirical testing would simply codify the disorganisation the paper is about.

3.2 Procedure

The framework was developed in five stages, summarised in Table 2.

Table 2. *Stages of framework development*

Stage	Activity	Output
1	Problem identification and motivation	Statement of the legitimacy deficit; RQ1–RQ3
2	Literature synthesis across four streams	Corpus of conceptual sources; stream-level gap analysis (Table 1)
3	Construct extraction and harmonisation	Vocabulary of layers, actors, explanation attributes
4	Framework synthesis	HX-Grid architecture (Figure 1); stakeholder matrix (Table 3, Figure 2)
5	Specification of testable content and evaluation strategy	Propositions P1–P6 (Table 4); evaluation protocol (Table 5)

For stage 2, the peer reviewed journal and conference literature, review articles and regulatory instruments were selected purposively, not exhaustively, to achieve conceptual saturation of the constructs rather than theoretical saturation. At Stage 3, recurring constructs were extracted and inconsistencies in terminology within streams, such as the different uses of ‘transparency' in the machine learning and regulatory literatures were resolved. Each of the two principles of ordering assemblies to layers (stage 4) – separation of concerns (one layer has one class of responsibilities) and directionality (models are explained toward actors, actors are feedback given to models) – will be discussed. Stage 5 made the framework falsifiable by stating propositions and stating how they would be tested.

3.3 Evaluation Strategy

Evaluation will follow a continuum from artificial towards naturalistic and formative towards summative, following known design science evaluation frameworks (Venable et al., 2016). There are three scenarios defined: (i) ex ante expert evaluation of the completeness and internal consistency of the framework, with the grid operators, the researchers from XAI and the regulators, (ii) artificial summative evaluation through instantiation on a simulated distribution feeder and synthetic renewable/load profiles, in which the configurations of explanations are changed and the quality of the decisions is monitored, and (iii) naturalistic evaluation through field instantiation in a control room or a prosumer trial. Specifications of construct and instrumentation are provided in Table 5.

3.4 Scope and Boundary Conditions

This paper does not include any empirical data, nor any performance figures, except where these quantities are used in the evaluation protocol, and are to be estimated in further work. The framework does not cover the design of generation assets, the wholesale market design or design of energy management at the building level (apart from where these touch on grid decisions).

4. Results and Findings: The HX-Grid Framework

4.1 Architecture Overview

The HX-Grid framework consists of four consecutive layers, together with a cross-cutting governance layer (Figure 1). The organising claim of the paper is that there is no module for explainability that can be added onto a model, but explainability emerges when all five are aligned; a grid that produces faithful explanations that nobody can use, or explanations that are usable that nobody has to act on, is not explainable in any operationally meaningful sense.

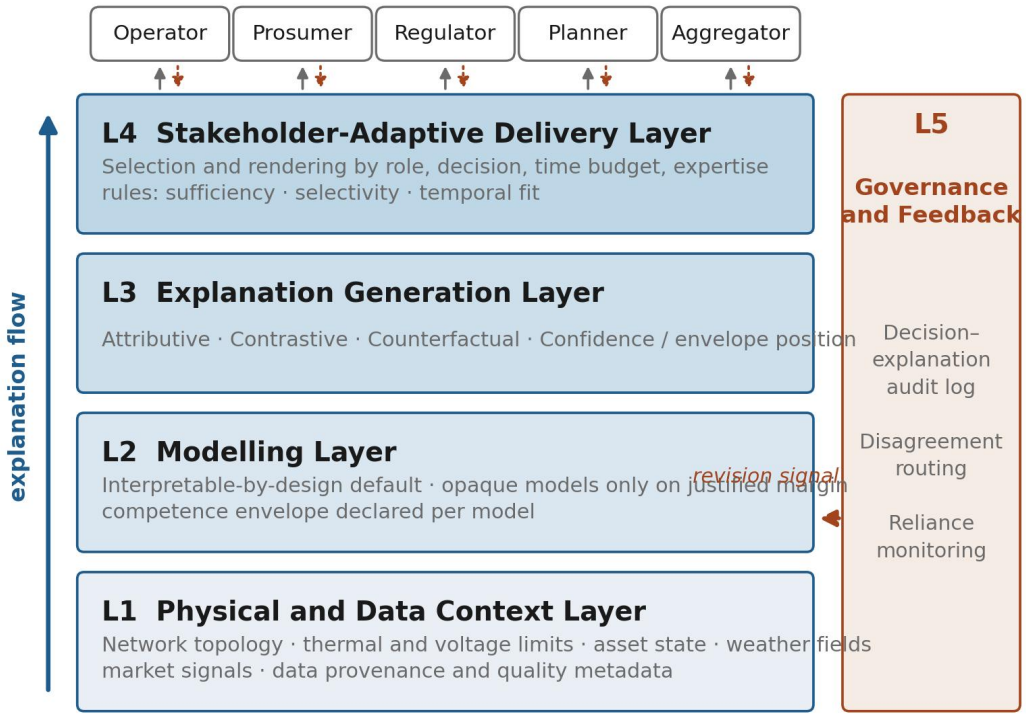


Figure 1. The HX-Grid framework: four sequential layers with a cross-cutting governance and feedback layer, serving five stakeholder classes.

Figure 1. *The HX-Grid framework: four sequential layers with a cross-cutting governance and feedback layer, serving five stakeholder classes.*

Layer 1 Physical context and data context. Only grid explanations on the basis of physics are considered explanations. This layer provides the network topology, thermal and voltage limits, asset status, weather fields and market signals to which any model output is read. It also contains data-provenance and quality metadata: if the explanation is giving a forecast for an irradiance sensor, but it was actually drifting, then that forecast has no value. First-class explanatory context, but not post-processing filters, are the physical constraints that enter the framework here.

Layer 2 Modelling. This layer contains the predictive and control models, and is subject to a design rule that interpretable by design architectures are required by default, with opaque models being acceptable only if there is a documented performance margin (Rudin, 2019). Each model comes equipped with an uncertainty envelope that is a clear definition of the operating conditions for which the model has been validated—the material for Layer 3's uncertainty communication and the precondition for calibration, rather than blanket trust.

Layer 3 Explanation generation. This layer defines the explanatory content in four registers: attributive (inputs that led to this output), contrastive (why this action and not the other the operator would have hoped for), counterfactual (what would have to change in order for the output to be different), confidence (where this instance falls in comparison to the competence envelope). The

contrastive register is highlighted since human explanations are always contrastive, and, as control-room reasoning is always based on a baseline (Miller, 2019), it is naturally oriented in contrast to that baseline.

Layer 4 Stakeholder-adaptive delivery. The key aspect of the framework that distinguishes it from the "traditional" approach. The content of explanation is picked up, compressed and rendered based on the role, decision and time budget and expertise of the recipient as given in Table 3. Delivery follows three principles: sufficiency (delivery contains enough causes to support the decision), selectivity (not all causes are delivered, only enough to support the decision), temporal fit (delivery latency is not more than decision window).

Layer 5 Governance and feedback. This cross cutting layer records decisions, explanations and human responses as an auditable record, fulfilling regulatory reconstruction requirements (European Commission, 2021); sends disagreement between operator judgement and model output up to Layer 2 (model output corrected back to Layer 2 as a revision signal); and tracks reliance patterns to detect miscalibration. Trust calibration is dynamic, rather than a design assumption, because of the feedback loop as described in Section 4.4.

4.2 The Stakeholder–Explanation Matrix

Table 3 differentiates explanation requirements across five actor classes, and Figure 2 renders the same relationship graphically: explanation depth demand rises with the decision window, and the temporal-fit rule bounds what can be delivered within it.

Table 3. *Stakeholder–explanation matrix for renewable-integrated smart grids*

Stakeholder	Primary decision	Decision window	Dominant register	Delivery form	Failure mode if mismatched
Control-room operator	Accept, modify, or override an automated dispatch or switching action	Seconds to minutes	Contrastive + confidence	Ranked causes with competence flag, embedded in SCADA view	Cognitive overload; reversion to manual heuristics
Distribution planner	Reinforcement and hosting-capacity decisions	Weeks to years	Global + counterfactual	Scenario reports with sensitivity narratives	Scenario opacity; unjustifiable capital plans
Regulator / auditor	Verify legality and non-discrimination of automated outcomes	Retrospective	Global + provenance record	Immutable decision–explanation log	Unauditable decisions; contested enforcement
Prosumer / household	Consent to and continue automated control of assets	Hours to days	Counterfactual + fairness account	Plain-language, comparative statements	Withdrawal from flexibility programmes
Aggregator /	Portfolio	Minutes to	Attributive +	Machine-readable	Systematic

Stakeholder	Primary decision	Decision window	Dominant register	Delivery form	Failure mode if mismatched
market agent	bidding and flexibility offers	hours	uncertainty	explanation payloads	mispricing of risk

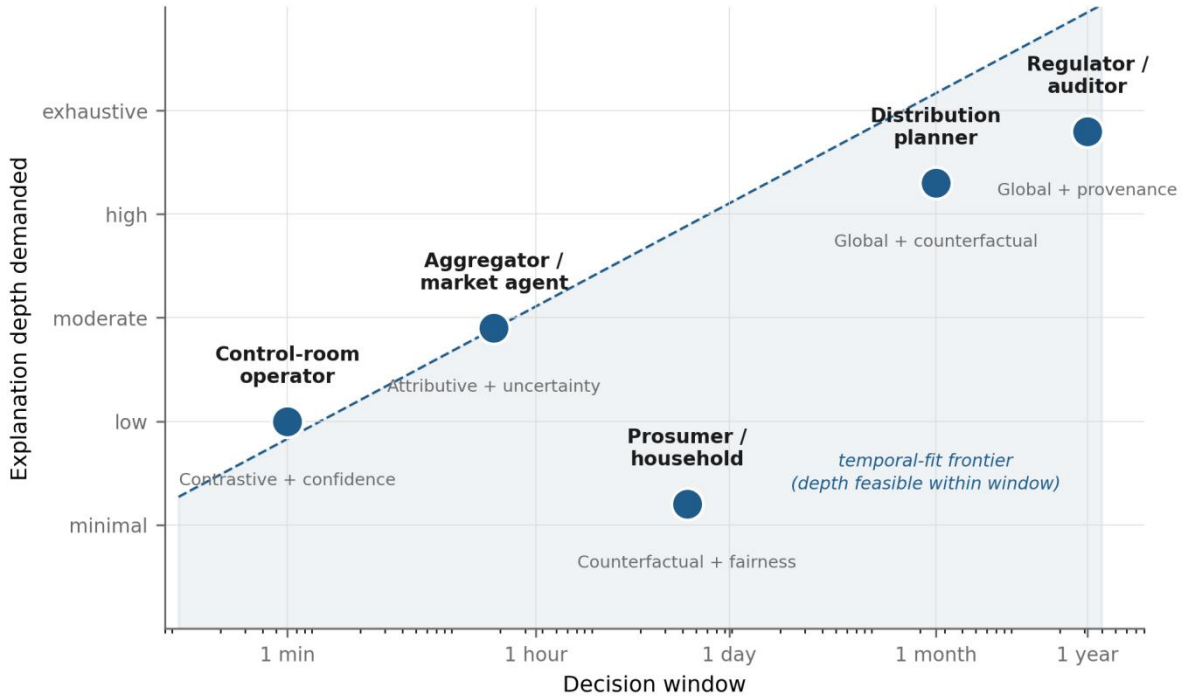


Figure 2. Stakeholder–explanation alignment map: explanation depth demand plotted against decision window, with the temporal-fit frontier.

Figure 2. *Stakeholder–explanation alignment map: explanation depth demand plotted against decision window, with the temporal-fit frontier.* Positions are conceptual placements derived from Table 3, not measured values.

The matrix makes the framework's core proposition concrete: a single explanation artefact cannot serve all five rows, and attempts to do so degrade toward the least demanding recipient.

4.3 Propositions

Table 4. *Propositions derived from the HX-Grid framework*

ID	Proposition
P1	Explanations aligned to a stakeholder's decision and time window improve decision quality relative to uniform explanations of equal technical fidelity.
P2	Communicating a model's competence envelope improves trust calibration more than increasing explanation detail within the envelope.

ID	Proposition
P3	Contrastive explanations reduce operator override latency relative to attributive explanations under contingency conditions.
P4	Explanations incorporating physical-constraint context are judged more credible by expert users than purely statistical attributions.
P5	Counterfactual, fairness-oriented explanations increase prosumer retention in flexibility programmes relative to attributive explanations.
P6	An active feedback loop (Layer 5) attenuates the growth of miscalibrated reliance over repeated interactions, relative to a static explanation regime.

4.4 The Trust Calibration Loop

Now, Layer 5 is unfolded in the mechanism in which explanation is proposed to act on trust, as in Figure 3. Model output is transmitted along with its competence envelope; Layers 3 and 4 provide an explanation suitable to the recipient; the human accepts, adjusts, or rejects that explanation; the set of accepted, adjusted or rejected explanations are compared to the competence envelope that the model has declared for that instance, and the differences are recorded and transmitted as a calibration signal, updating the model, its declared envelope, or the explanation policy. Unlike statically-deployed XAI, where an explanation interface is defined at design time and then never measured, or overridden by the user when the model is incorrect, or ignored by the user when it is correct, in the loop, miscalibration accumulates.

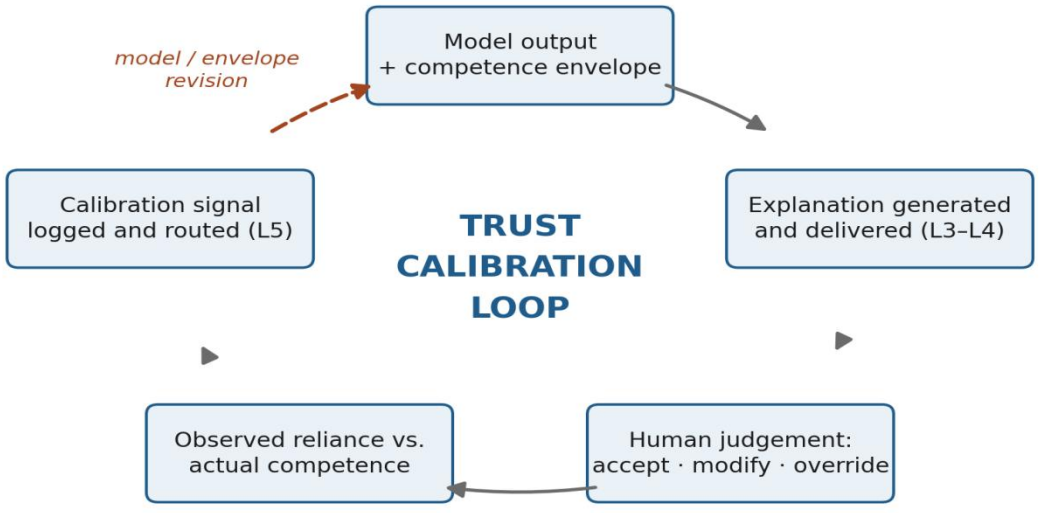


Figure 3. The trust calibration loop operationalised by Layer 5 of the HX-Grid framework.

Figure 3. *The trust calibration loop operationalised by Layer 5 of the HX-Grid framework.*

4.5 Evaluation Protocol

Table 5. *Evaluation constructs, instruments, and placeholder estimands*

Construct	Instrument	Estimand (placeholder)
Trust calibration	Reliance–competence divergence over repeated trials	δ = [to be estimated]
Explanation satisfaction	Validated XAI satisfaction scale (Hoffman et al., 2018)	μS = [to be estimated]
Decision quality	Error rate against ground-truth optimal action in simulated contingencies	ε = [to be estimated]
Override appropriateness	Proportion of overrides occurring outside the competence envelope	π = [to be estimated]
Temporal fit	Explanation latency relative to decision window	τ = [to be estimated]
Prosumer retention	Continuation rate in a simulated flexibility programme	ρ = [to be estimated]

All estimands remain unspecified pending empirical work; the protocol's function here is to fix constructs and instruments in advance, thereby constraining subsequent analytic flexibility.

5. Discussion

5.1 Theoretical Implications

The framework relocates explainability from the model to the sociotechnical system. There are three theoretical implications of this. First, it turns explanation in the direction of decision making: The unit of analysis is no longer the prediction, but the decision.First, it turns the unit of analysis back to that of the decision, not the prediction, redirecting the energy XAI research away from method benchmarking. Second, it brings in calibration as the norm, rather than the maximisation of trust that is implicit in the writing of applied XAI, which has failed, not succeeded (Lee & See, 2004; Jacovi et al., 2021). Third, it focuses on the physical constraint as the explanatory content instead of as a domain detail, as is done in medical and financial energy XAI, and this indicates that there will be systematic under-specifications of energy grid requirements when using borrowed evaluation criteria.

5.2 Practical Implications

For operators, the framework means that the competence envelopes and interfaces for explaining to stakeholders must be given as a contractual deliverable and not left to the vendor's discretion. For VENDORS, Layer 4 means a separation between the generation of explanation and the rendering of explanation at an architectural level, where one explanation substrate can be used by different audiences without any duplication of explanation modelling. The prosumer row of Table 3 highlights

one overlooked risk for distribution utilities seeking to open up flexibility markets: programme attrition that isn't caused by the design of the tariff, but by unintelligible automation. Loggers at Layer 5 must include the explanation displayed, and the human response, not just the model output, as is currently the case in most SCADA/historian architectures.

5.3 Policy Implications

Human oversight is essential for effective risk tiered AI regulation, not just nominal (European Commission, 2021). The framework provides a design word for showing effectiveness: oversight is effective if the explanation given is within the decision and window of the individual doing the oversight; if the audit record allows for reconstruction. On the other hand, it finds a compliance failure mode that is not detectable through a checklist audit of the system that fulfills all the documentation obligations but provides explanations that are not usable by any operator in the decision window of the system.

5.4 Limitations

The contribution is limited by four conditions. The framework is theoretical, has not yet been validated, its propositions are plausible but not tested, and the evaluation protocol is a proposal, not a result. The literature synthesis was not systematic, but rather based on purpose, allowing for construct extraction bias. The five classes of actors are analytical simplifications as real networks are hybrid actors—actors who are a prosumer and an aggregator, or clients who have an explanation that is internally contradictory. In addition, the framework recognizes that while explanation may be harmful in adverse contexts, where it is offered to give market participants a chance to manipulate model behaviour, it is helpful in a well-targeted context.

5.5 Future Research

The work priority is empirical. The feasibility of studying the testable condition of varying explanation alignment while maintaining fidelity in the control room is proposed to be achieved by conducting simulation studies. Longitudinal reliance measurement is needed for energy contexts which is not supported by the field in P2. P5 can be assessed via field trials with distribution utilities with flexibility programmes. In addition to the proposition testing, there are three avenues of investigation that are worth consideration: (1) the creation of metrics for the quality of explanations that are specific to energy domain; (2) the extension of the framework to a multi-agent setting, in which explanation must be exchanged among autonomous agents, as well as from an agent to a human; and (3) the examination of explanation under adversarial or market conditions.

6. Conclusion

AI's role in the operational imperative of transitioning to renewable based electricity systems, and its exposure as an opacity that is no longer an engineering externality, is the result of this transition.This transition is a consequence of the operational necessity of electricity systems transitioning to renewable energy and of the fact that opacity is no longer an engineering externality. Despite the technical advances in XAI research in power systems, it has provided an answer to a question that the sector has not asked: how to optimise the supply of explanation given no consideration of its demand. The HX-Grid is a corrective in terms of organizing explainability as a sociotechnical architecture made up of the inter-constitutive relations between physical context, competence-bounded modelling, multi-register explanation generation, stakeholder-adaptive delivery and governance feedback. It makes the following key points: explanation's purpose needs to be differentiated, calibration must be the normative target not trust, and over time it is the process of feedback that maintains calibration. At

the time the value of the framework is only heuristic and prospective, and the propositions and the evaluation protocol presented here are made available just so that they can be tested and, if they do not pass, falsified. The design of explainability must be a part of the grid's system properties if it is to be responsible and yet independent.Explainability needs to be designed into the system property of the grid, and, most of all, designed for those who must hold it accountable for what it does.

Declarations

Conflict of Interest
The authors declare no conflicts of interest related to this study.

Funding Statement
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical Approval
This study did not involve any human or animal subjects and, therefore, did not require ethical approval.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahmad, T., Zhu, H., Zhang, D., Tariq, R., Bassam, A., Ullah, F., AlGhamdi, A. S., & Alshamrani, S. S. (2022). Energetics systems and artificial intelligence: Applications of industry 4.0. *Energy Reports, 8*, 334–361. https://doi.org/10.1016/j.egyr.2021.11.256

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning (arXiv:1702.08608). *arXiv*. https://doi.org/10.48550/arXiv.1702.08608

European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gungor, V. C., Sahin, D., Kocak, T., Ergut, S., Buccella, C., Cecati, C., & Hancke, G. P. (2011). Smart grid technologies: Communication technologies and standards. *IEEE Transactions on Industrial Informatics, 7*(4), 529–539. https://doi.org/10.1109/TII.2011.2166794

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects (arXiv:1812.04608). *arXiv*. https://doi.org/10.48550/arXiv.1812.04608

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review, 10*(1–2), 18–26. https://doi.org/10.1007/s13162-020-00161-0

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624–635). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445923

Kuzlu, M., Cali, U., Sharma, V., & Guler, O. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access, 8*, 187814–187823. https://doi.org/10.1109/ACCESS.2020.3031477

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.

Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI, 9*, 100169. https://doi.org/10.1016/j.egyai.2022.100169

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Tuballa, M. L., & Abundo, M. L. (2016). A review of the development of smart grid technologies. *Renewable and Sustainable Energy Reviews, 59*, 710–725. https://doi.org/10.1016/j.rser.2016.01.011

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289