ORIGINAL RESEARCH **Open access**

Resilient AI Architectures for Cyber-Physical Energy Systems: Robustness, Explainability, and Recovery under Disruption

Maria Gonzales[1*], Jose Santos[1], Anna Cruz[2], Carlo Reyes[3]

Abstract

AI is increasingly embedded in cyber-physical energy systems, where it supports forecasting, monitoring, control, optimization, and operational decision-making. However, this growing dependence also exposes critical energy infrastructure to new forms of disruption, including adversarial attacks, corrupted data streams, model failures, communication instability, and unforeseen operational faults. Many AI designs remain optimized primarily for accuracy under normal operating conditions, with insufficient attention to robustness during disturbance, explainability during failure, and recovery after disruption. This creates a fragile dependency in systems where degraded AI performance can affect grid reliability, asset protection, and operator decision-making. This article proposes a resilient AI architecture for cyber-physical energy systems that treats robustness, explainability, and recovery as co-equal design pillars rather than separate technical concerns. The architecture is developed through conceptual synthesis of literature on adversarial robustness, explainable AI, resilience engineering, and cyber-physical energy system design. No experiments, simulations, datasets, or empirical performance results are reported. The proposed architecture identifies resilience-enhancing components within each pillar, including adversarial testing, secure data governance, diagnostic explanation, cross-model validation, fallback procedures, post-incident learning, and human oversight. It also explains how these components interact to support fault containment, operational transparency, adaptive recovery, and institutional learning. The article contributes a design-oriented framework for embedding resilience into AI-enabled energy systems from the outset, shifting AI from a potential source of infrastructure fragility toward a governed, interpretable, and adaptive capability for secure energy operations.

Keywords Explainability, Resilient AI, Cyber-physical energy systems, Robustness, Recovery, Architectural resilience

*Correspondence:
Maria Gonzales
maria.gonzales@up.edu.ph
[1] Department of AI for Off-Grid and Island Energy Systems, Faculty of Engineering, University of the Philippines Diliman, Quezon City, Philippines
[2] Department of Smart Grid Data Science and Optimization, Faculty of Engineering, De La Salle University, Manila, Philippines
[3] Department of Intelligent Energy Infrastructure, Faculty of Engineering, Ateneo de Manila University, Quezon City, Philippines

Introduction

AI is becoming increasingly embedded in cyber-physical energy systems, where computational models interact with sensors, communication networks, distributed energy resources, and grid control infrastructures. This convergence can improve monitoring, forecasting, and operational coordination, but it also expands the range of failure pathways through which cyber, data, and physical disturbances may propagate [1]. Cyber-physical power systems are especially vulnerable because their digital components do not merely observe the grid but increasingly shape operational decisions that affect

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.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

physical stability [2]. A resilient AI architecture for this context should therefore treat disruption not as an exceptional event but as a normal design condition.

Conventional AI designs in energy systems often privilege predictive accuracy, computational efficiency, or automation performance under relatively stable operating conditions. However, adversarial machine learning research shows that models can be manipulated through carefully crafted inputs, corrupted data streams, or strategic attacks that exploit learned decision boundaries [3]. In energy applications, these weaknesses can be especially consequential because compromised predictions may influence dispatch, voltage control, demand response, or system protection decisions [4]. Robustness must therefore be understood as a core architectural requirement rather than a post-deployment security patch.

An architecture-oriented approach is needed because resilience cannot be achieved through a single algorithmic technique. Explainable AI research in energy and power systems suggests that operators need interpretable outputs to understand why a model is issuing warnings, recommending controls, or classifying faults during abnormal conditions [5]. At the same time, reinforcement-learning and self-healing approaches indicate that recovery requires coordinated mechanisms for reconfiguration, fallback, and restoration after disruption [6]. Robustness, explainability, and recovery should therefore be co-designed as interacting layers within the AI system.

This article contributes a conceptual architecture for resilient AI in cyber-physical energy systems by specifying how robustness, explainability, and recovery can be integrated into a common design logic. The architecture draws on work on adversarial attacks in smart grids [7], cyber-physical power system resilience [8], and explainability for energy operations [9] to articulate a framework that remains conceptual rather than experimental. It does not report empirical results, attack metrics, or simulation findings. Instead, it proposes design principles and governance considerations that could guide future implementation, validation, and certification.

Background

Cyber-physical energy systems and AI integration

Cyber-physical energy systems can be conceptualized as layered infrastructures in which physical grid assets, communication networks, control systems, data platforms, and AI models continuously interact. State-description approaches to cyber-physical energy systems emphasize that resilience depends on understanding not only the physical state of grid components but also the informational and computational states that shape control decisions [10]. AI integration increases this interdependence by enabling models to infer system conditions, detect anomalies, and recommend actions based on heterogeneous data streams. Yet the same integration can create new vulnerabilities when corrupted data, communication failures, or compromised models affect operational judgement.

Adversarial robustness in machine learning

Adversarial robustness refers to the capacity of machine learning systems to maintain reliable behavior when inputs are intentionally manipulated, distributionally shifted, or corrupted. General adversarial learning research demonstrates that deep models can be vulnerable to evasion, poisoning, and transfer attacks, even when they appear accurate under ordinary test conditions [11]. In power systems, adversarial concerns are intensified because attacks against forecasting or detection models may exploit operational dependencies between cyber data and physical responses [12]. A resilient architecture should therefore incorporate adversarial awareness into data ingestion, model training, runtime monitoring, and decision validation.

Explainable AI in critical operations

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Explainable AI is central to resilient operations because operators must understand model behavior when conditions are ambiguous, degraded, or rapidly changing. Reviews of XAI in energy systems show that interpretability can support trust, diagnostic reasoning, and operational accountability when AI models are used in high-stakes control environments [5]. Explainability is not merely a communication feature; it can help identify whether an alert reflects a physical fault, a data-quality problem, or a possible adversarial manipulation [13]. In resilient AI architectures, explanation interfaces should therefore be designed as operational tools rather than decorative transparency mechanisms.

Recovery and self-healing in energy grids

Recovery in cyber-physical energy systems involves restoring acceptable function after disruption while preventing cascading failures and unsafe automation. Deep reinforcement learning research for resilient power and energy systems suggests that AI can support adaptive response, reconfiguration, and operational decision-making under disturbed conditions [6]. However, recovery mechanisms should not be treated as purely autonomous because grid restoration often requires human judgement, institutional procedures, and safety constraints. A resilient AI architecture should therefore combine automated recovery proposals with fallback control, operator confirmation, and post-incident learning.

Gaps in current AI resilience architectures

Current approaches to AI resilience in energy systems remain fragmented across cybersecurity, robust learning, explainability, and recovery research. Reviews of cyber-physical attacks in future energy systems highlight the breadth of threat pathways, but they also show that attack detection and defence are often considered separately from operational recovery and governance [14]. Similarly, work on cybersecurity analysis in cyber-physical power systems often emphasizes detection and defence without fully integrating explainability and restoration into the same architecture [2]. The resulting gap is an absence of unified design models that specify how resilient AI components should interact before, during, and after disruption.

Conceptual Foundation

Defining AI resilience in cyber-physical contexts

AI resilience in cyber-physical energy systems can be defined as the capacity of AI-enabled functions to maintain safe operation under disruption, degrade gracefully when full functionality is no longer possible, and support rapid restoration without amplifying system instability. This definition extends conventional robustness by including diagnosis, interpretability, fallback, and recovery as part of the architecture. Research on cyber-physical power system resilience emphasizes that resilience involves preparation, absorption, adaptation, and restoration across interdependent cyber and physical layers [8]. For AI systems, these phases should be translated into model-level, interface-level, control-level, and governance-level design requirements.

The interplay of robustness, explainability, and recovery

Robustness, explainability, and recovery reinforce each other but may also create architectural tensions. Robust models can reduce the probability that corrupted inputs cause unsafe behavior, but without explanations operators may not know whether the model is reliable during an unfolding disturbance [15]. Explainability can improve situational awareness, yet explanations that are too slow, complex, or misleading may impede crisis response [9]. Recovery mechanisms can restore functionality, but they require robust state estimation and interpretable justification so that reconfiguration does not create additional risk.

Architectural principles for resilience

A resilient AI architecture should incorporate defence-in-depth, least privilege, redundancy, diversity, fail-safe defaults, and fault containment. These principles are consistent with broader cybersecurity and resilience thinking in cyber-physical

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

power systems, where layered protection and controlled degradation are needed to prevent local failures from escalating into systemic disruption [16]. For AI components, defence-in-depth could involve protected data pipelines, adversarially aware models, runtime anomaly detection, and independent validation before control actions are executed. The architecture should also separate monitoring, explanation, decision support, and actuation functions so that compromise in one layer does not automatically compromise the entire system.

Architecture Components: Robustness, Explainability, Recovery

Robustness layer

The robustness layer could be designed to protect AI components against adversarial inputs, corrupted data, out-of-distribution conditions, and cyber-physical inconsistencies. Adversarial training, formal verification, input denoising, and physics-constrained validation can help ensure that model outputs remain plausible under disturbance [15]. Stream-learning approaches for false data injection attack identification indicate that runtime adaptation may also be important when the attack environment changes over time [17]. In the proposed architecture, robustness functions operate before and during model inference to filter suspicious inputs, detect abnormal behavior, and constrain outputs within safe operational envelopes.

Explainability layer

The explainability layer translates AI outputs into operationally meaningful explanations for grid operators, engineers, and governance reviewers. Reviews of interpretable machine learning in smart grids show that explanation methods can clarify which features, measurements, or system conditions influenced a model's output [9]. In resilience contexts, this layer should provide real-time diagnostic explanations during events and more detailed post-event explanations for root-cause analysis. It should also distinguish between normal uncertainty, sensor degradation, possible cyber manipulation, and physical disturbance so that operators can act on explanations rather than merely observe them.

Recovery layer

The recovery layer supports graceful degradation, fallback control, automated reconfiguration, and incremental restoration after disruption. Research on reinforcement learning environments for cyber-resilient power distribution systems indicates that AI can be structured to explore recovery-oriented responses while respecting operational constraints [18]. In the proposed architecture, recovery should not mean immediate full automation; it should include safe fallback modes, human-in-the-loop approval, and restoration pathways that are explainable to operators.

Figure 1 presents the proposed resilience architecture, showing how robustness, explainability, and recovery operate as interacting layers that protect AI-enabled cyber-physical energy systems before, during, and after disruption.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

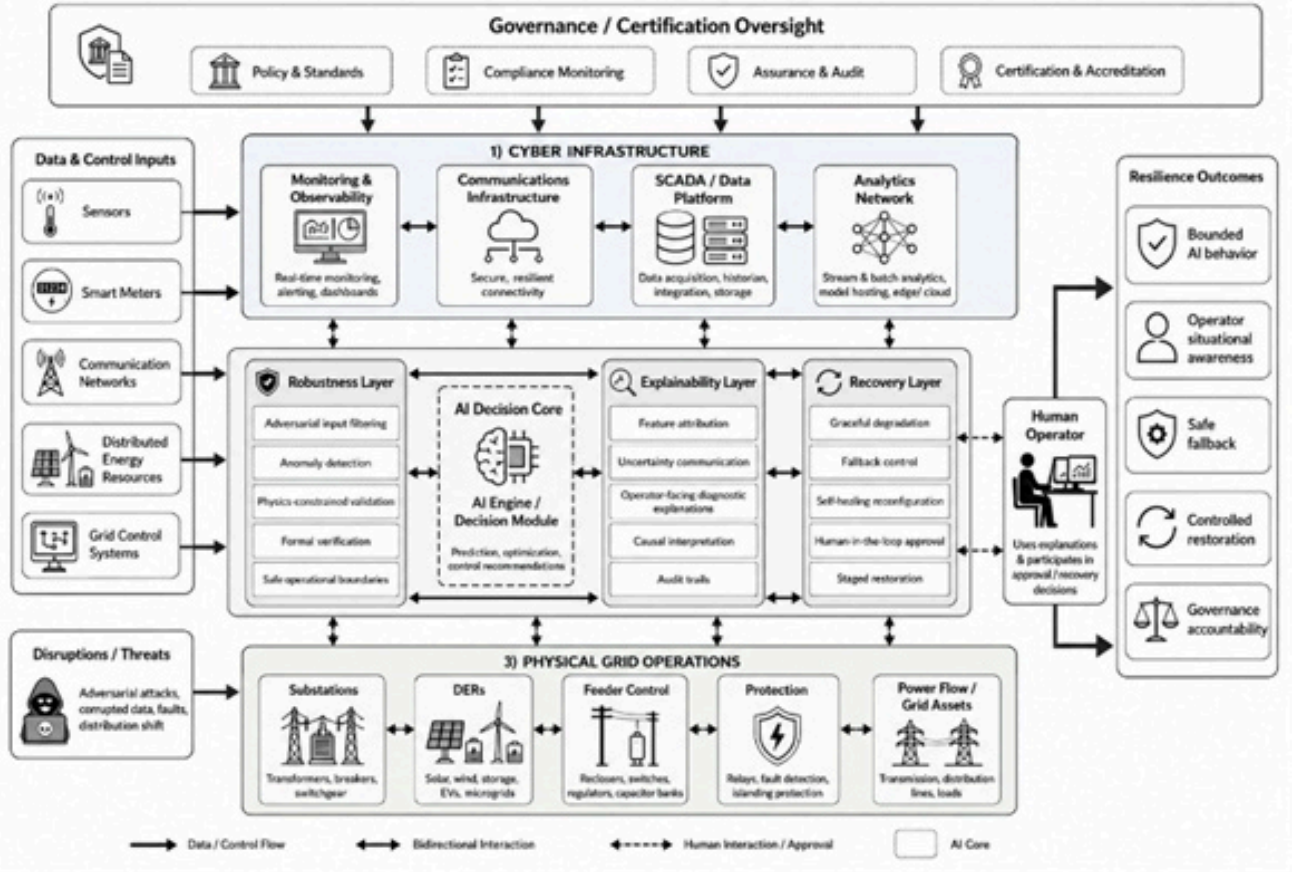

Figure 1. Resilience Architecture Diagram for AI-Enabled Cyber-Physical Energy Systems

Table 1 presents the three resilience pillars as architectural design commitments, linking each pillar to design principles, AI techniques, and expected resilience outcomes.

Table 1. AI-Driven Design Commitments for Resilient Power Distribution Systems Across Robustness, Explainability, and Recovery Pillars

Resilience Pillar	Design Principles	AI Techniques	Expected Resilience Outcome
Robustness	Defence-in-depth, physics consistency, adversarial awareness, fail-safe boundaries	Adversarial training, anomaly detection, input validation, physics-constrained learning, formal verification	AI outputs remain bounded, plausible, and operationally usable under corrupted data, attack, or distribution shift
Explainability	Operator intelligibility, diagnostic transparency, uncertainty communication, accountability traceability	Feature attribution, interpretable surrogate models, rule-based explanation layers, causal diagnostic interfaces, uncertainty-aware explanations	Operators can understand model behavior during disruption and distinguish faults, attacks, and uncertain recommendations
Recovery	Graceful degradation, redundancy, fallback control, staged restoration, human confirmation	Self-healing control, reinforcement-learning-assisted reconfiguration, ensemble fallback, digital-twin-supported restoration planning	Critical AI-supported functions can shift to safe modes, support restoration decisions, and recover without cascading failure

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Robustness Design Principles

Defending against adversarial inputs

Defending against adversarial inputs requires protection at the data, model, and decision-interface levels. General reviews of adversarial attack and defence technologies show that adversarial examples can exploit model sensitivity in ways that are not obvious from normal validation procedures. In energy contexts, adversarial attacks on power-system machine learning models could manipulate forecasts, anomaly detectors, or control recommendations if input validation and model hardening are absent [12]. A resilient architecture should therefore incorporate denoising, adversarial training, sensor cross-checking, and plausibility constraints before AI outputs influence operational decisions.

Fault detection and isolation

Fault detection and isolation should be designed as continuous architectural functions rather than occasional diagnostic procedures. AI-based detection can identify abnormal system behavior, but resilience requires that detection be linked to graceful degradation, containment, and operator notification [2]. Smart coordinated cyber-attacks on wind turbines show why detection should move beyond binary alarms toward identification of attack type, affected subsystem, and likely operational implications [19]. In the proposed architecture, fault isolation should separate compromised data sources, restrict affected models, and preserve minimum safe functionality through fallback pathways.

Diversity and redundancy in AI models

Diversity and redundancy can reduce the risk that a single model failure becomes a common-mode failure across the cyber-physical energy system. Adversarial machine learning research indicates that attacks may transfer across models, which means redundancy is insufficient if all models share similar architectures, training data, or vulnerabilities [3]. A resilient architecture should therefore combine heterogeneous model families, independent validation modules, and physics-based consistency checks rather than relying on repeated copies of the same AI component. Such diversity would be expected to improve fault containment, support cross-model disagreement detection, and create more reliable triggers for fallback or human review.

Figure 2 illustrates how resilience in AI-enabled cyber-physical energy systems can be strengthened by combining heterogeneous model families, independent validation modules, and physics-based consistency checks so that single-model failure or transferable adversarial vulnerability does not propagate as a common-mode system failure.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

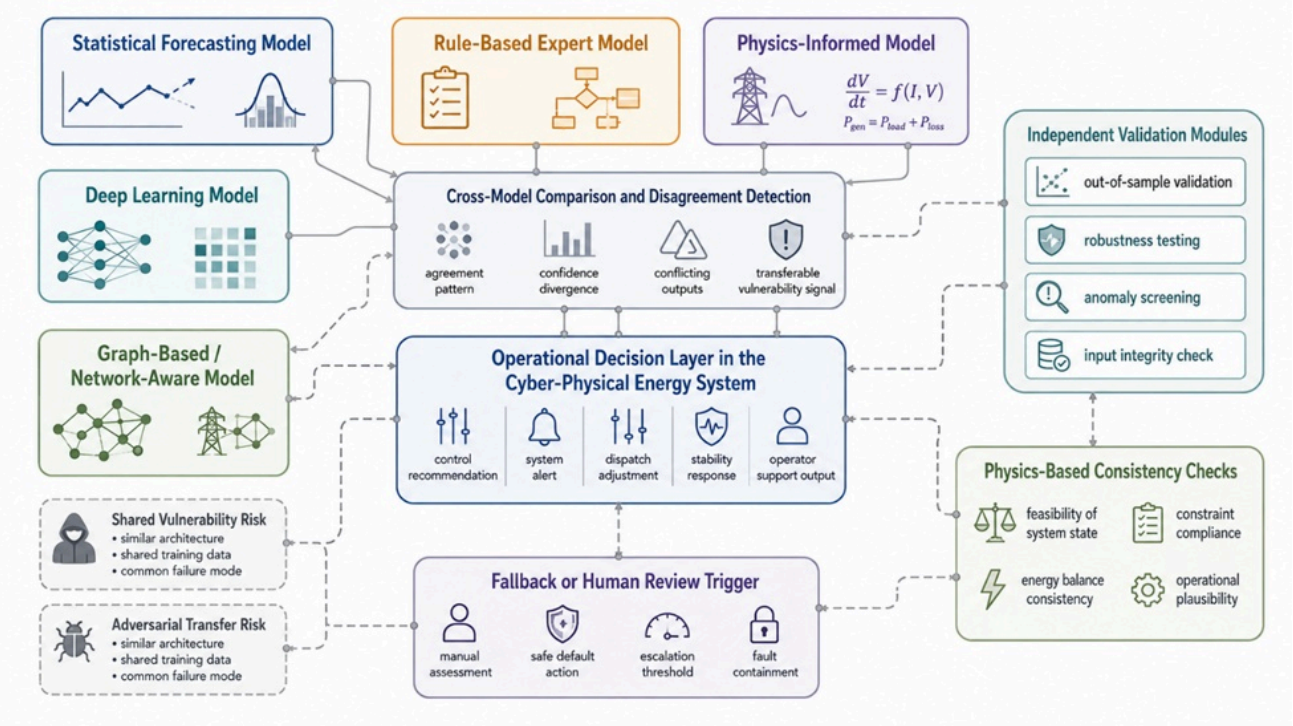

Figure 2. Diversity and Redundancy Architecture for Resilient AI-Enabled Cyber-Physical Energy Systems

Explainability for Resilient Operations

The role of explainability during disruptions

Explainability during disruptions should be understood as a resilience function that helps operators interpret whether AI behavior remains trustworthy under abnormal conditions. In cyber-physical energy systems, adversarial or corrupted inputs may produce outputs that appear technically plausible but are operationally unsafe, making explanation essential for informed human oversight [20]. Explainable AI can support operator reasoning by showing which signals, constraints, or historical patterns shaped a recommendation during a disturbance [5]. A resilient architecture should therefore treat explanation as part of crisis management rather than as a retrospective reporting tool.

Real-time diagnostic explanation

Real-time diagnostic explanation could be designed to translate AI alerts into concise operational interpretations under time pressure. For example, explainable fault detection in photovoltaic systems demonstrates how interpretable AI can help distinguish equipment faults from data irregularities or model uncertainty [21]. In grid operations, this principle can be extended to explanation interfaces that identify affected assets, influential measurements, confidence limits, and possible recovery pathways. Such explanations should be prioritized according to urgency so that operators receive decision-relevant information rather than excessive technical detail.

Explainability for post-incident learning

Explainability also supports post-incident learning by creating traceable links between AI inputs, model reasoning, alerts, and operational decisions. Explainable AI for energy systems maintenance shows that explanations can help identify recurring fault patterns, maintenance priorities, and weaknesses in monitoring design [22]. In a resilient architecture, post-incident explanations should feed into root-cause analysis, model review, and governance reporting. This would allow organizations to improve both technical models and institutional procedures after disruptions.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Avoiding cognitive overload

Although explainability is necessary, poorly designed explanations can overload operators during crises. Reviews of AI explainability and governance in smart energy systems highlight that transparency must be aligned with human decision contexts rather than delivered as unfiltered model detail [13]. A resilient architecture should therefore distinguish between real-time operational explanations, engineering-level diagnostics, and governance-level audit trails. This layered approach can preserve interpretability while avoiding excessive cognitive burden during urgent system events. **Figure 3** illustrates how explainability in AI-enabled energy systems should be differentiated across real-time operator interpretation, engineering diagnostic analysis, and governance-level post-incident learning to support resilience without creating cognitive overload.

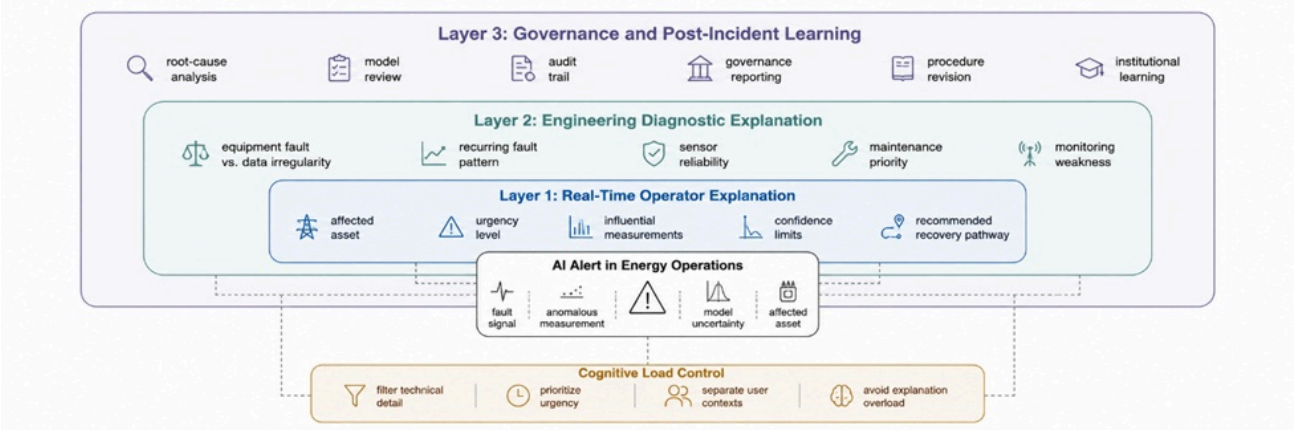

Figure 3. Layered Explainability Architecture for Real-Time Response and Post-Incident Learning in AI-Enabled Energy Systems

Recovery and Self-Healing Mechanisms

Automated reconfiguration

Automated reconfiguration could allow AI systems to support topology adjustment, isolation of compromised components, and restoration of service after localized disruption. Resilience enhancement of reinforcement-learning-based demand response under adversarial attacks suggests that AI recovery functions must account for both cyber manipulation and physical system constraints [20]. In the proposed architecture, reconfiguration should operate within predefined safety envelopes and should not bypass protection logic or operator authority. Automated recovery should therefore be constrained, explainable, and reversible.

Incremental restoration with human-in-the-loop

Incremental restoration with human-in-the-loop control recognizes that full automation may be inappropriate during uncertain or degraded conditions. Deep-learning systems for voltage stability assessment show how AI can support early recognition of system stress, but restoration decisions still require contextual judgement about operational priorities and acceptable risk [23]. A resilient architecture should allow AI to propose staged restoration options while human operators confirm, modify, or reject these options. This preserves human responsibility while using AI to reduce diagnostic and planning complexity.

Learning from disruptions

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Learning from disruptions requires mechanisms that update models, thresholds, explanations, and procedures after incidents without introducing uncontrolled adaptation. Adversarial attack and defence research in deep learning shows that defences must evolve because attack strategies and model vulnerabilities change over time [24]. In cyber-physical energy systems, post-event learning should therefore be governed through validation gates, version control, and review of whether new model behavior remains consistent with safety constraints. The architecture should support learning, but it should not permit opaque self-modification in safety-critical functions.

Testing and validation of recovery procedures

Recovery procedures should be tested before deployment through sandboxed environments, digital twins, stress scenarios, and structured resilience drills. Cybersecurity reviews for smart grid communication emphasize that intrusion detection and threat mitigation systems require validation across realistic communication and operational conditions [25]. For resilient AI architectures, recovery testing should examine whether fallback modes activate properly, whether explanations remain understandable, and whether restoration pathways avoid secondary failures. Such validation would be expected to strengthen confidence without implying that resilience can be guaranteed under all future disruptions.

Architecture Integration and Governance

System-level integration

System-level integration requires the robustness, explainability, and recovery layers to operate as mutually reinforcing components rather than isolated modules. Artificial intelligence techniques for smart-grid stability analysis and control show that AI functions often span prediction, assessment, and control, which means resilience must be coordinated across the full decision chain [26]. In the proposed architecture, robustness filters suspect inputs, explainability clarifies model reasoning, and recovery mechanisms activate fallback or restoration pathways when safe operation is threatened. Integration should be designed around controlled information flows, explicit handoff points, and governance visibility.

Governance, certification, and compliance

Governance should specify who is responsible for approving AI deployment, monitoring model drift, reviewing disruptions, and authorizing recovery behavior in cyber-physical energy systems. Reviews of cyber-physical power systems identify standards, technology drivers, and future perspectives that point toward the need for certification and compliance mechanisms suited to interconnected energy infrastructures [16]. A resilient AI architecture should therefore include audit trails, model documentation, incident review procedures, and escalation protocols. Governance is not external to the architecture; it is part of the architecture's control logic.

Evaluation Strategy for the Architecture

Resilience testing and red-teaming

Evaluation of the proposed architecture should include resilience testing and red-teaming focused on adversarial inputs, corrupted measurements, communication failures, and physical disturbances. Adversarial attacks on deep learning models in smart grids indicate that energy AI systems should be tested against disruption scenarios that challenge both data integrity and operational plausibility [7]. Such testing should evaluate whether robustness mechanisms detect abnormal inputs, whether explanations support operator judgement, and whether recovery functions activate safely. Because this article is conceptual, these evaluation strategies are proposed as future validation requirements rather than reported results.

Expert review and stakeholder feedback

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Expert review and stakeholder feedback should complement technical testing because resilience depends on organizational usability as well as algorithmic behavior. Explainable artificial intelligence for photovoltaic fault detection illustrates that explanation tools must be assessed in relation to the practical needs of users interpreting system faults [27]. In cyber-physical energy systems, expert panels could examine whether the architecture's explanations, fallback rules, and governance procedures are understandable to operators, engineers, regulators, and cybersecurity teams. This review process would help identify design weaknesses that purely technical validation might overlook.

Transferability across energy domains

The architecture should be evaluated for transferability across transmission systems, distribution networks, renewable plants, microgrids, and energy service platforms. Cyber-physical attack research on wind turbines shows that AI resilience concerns are not limited to central grid control but also apply to distributed and renewable energy assets [28]. At the same time, differences in communication structure, operational tempo, autonomy level, and regulatory responsibility mean that the architecture should be adapted rather than copied across domains. Transferability should therefore be treated as a design question requiring contextual validation.

Limitations

Conceptual nature and need for implementation

The proposed architecture remains conceptual and requires further implementation, testing, and refinement in operationally realistic environments. Work on physics-constrained robustness for machine learning in low-carbon smart grids suggests that resilience design can benefit from formal and domain-specific constraints, but translating such principles into deployable architectures requires additional engineering work [15]. This article does not claim that the architecture has been empirically validated or that its components will perform uniformly across all energy contexts. Its value lies in specifying an integrated design logic that future studies could operationalize.

Emerging threats and technological evolution

A further limitation is that the threat landscape for AI-enabled cyber-physical energy systems continues to evolve. Reviews of adversarial attacks and defences in deep learning show that new attack strategies can emerge as models, data infrastructures, and deployment settings change [24]. Energy systems are also undergoing structural transformation through decentralization, renewable integration, and increasing digital interconnection, which may create new vulnerabilities not fully captured by current resilience concepts [14]. The architecture should therefore be treated as adaptable and revisable rather than final.

Conclusion

This article proposed a resilient AI architecture for cyber-physical energy systems organized around three co-equal pillars: robustness, explainability, and recovery. The architecture conceptualizes AI resilience as more than resistance to attack; it includes operational intelligibility, graceful degradation, controlled restoration, and governance accountability.

The central argument is that resilience must be architected into AI systems from the beginning rather than added after deployment. AI components that influence energy operations should be designed with disruption, uncertainty, adversarial manipulation, and recovery responsibilities in mind.

Future research should translate this conceptual architecture into implementation pathways, validation protocols, regulatory guidance, and operational training practices. Collaboration among AI researchers, power engineers, cybersecurity specialists, regulators, and system operators will be necessary to ensure that AI strengthens rather than weakens the resilience of energy systems.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 18 Dec 2023 Revised: 23 Feb 2024 Accepted: 11 Mar 2024
Published online: 15 November 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Hao J, Tao Y. Adversarial attacks on deep learning models in smart grids. Energy Rep. 2022;8(Suppl 2):123-9. https://doi.org/10.1016/j.egyr.2021.11.026.

2. Zhang Z, Yang Z, Yau DKY, Tian Y, Ma J. Data security of machine learning applied in low-carbon smart grid: a formal model for the physics-constrained robustness. Appl Energy. 2023;347:121405.

3. Zeng L, Qiu D, Sun M. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. Appl Energy. 2022;324:119688. https://doi.org/10.1016/j.apenergy.2022.119688.

4. Luo J, Hong T, Fang SC. Benchmarking robustness of load forecasting models under data integrity attacks. Int J Forecast. 2018;34(1):89-104.

5. Shi Z, Yao W, Li Z, Zeng L, Zhao Y, Zhang R, et al. Artificial intelligence techniques for stability analysis and control in smart grids: methodologies, applications, challenges and future directions. Appl Energy. 2020;278:115733. https://doi.org/10.1016/j.apenergy.2020.115733.

6. Li Y, Zhang M, Chen C. A deep-learning intelligent system incorporating data augmentation for short-term voltage stability assessment of power systems. Appl Energy. 2022;308:118347.

7. Huang R, Li Y. Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system. IEEE Trans Smart Grid. 2023;14(3):2367-76.

8. Hallaji E, Razavi-Far R, Wang M, Saif M, Fardanesh B. A stream learning approach for real-time identification of false data injection attacks in cyber-physical power systems. IEEE Trans Inf Forensics Secur. 2022;17:3934-45. https://doi.org/10.1109/TIFS.2022.3216948.

9. Sahu A, Venkataramanan V, Macwan R. Reinforcement learning environment for cyber-resilient power distribution system. IEEE Access. 2023;11:127216-28. https://doi.org/10.1109/ACCESS.2023.3282182.

10. Mazare M, Ramezani H. Enhancing cybersecurity in wind turbines: a resilient reinforcement learning-based optimal control for mitigating FDI attacks. Appl Energy. 2024;373:123939. https://doi.org/10.1016/j.apenergy.2024.123939.

11. Wu H, Badihi H, Xue Y, Vilkko M. Smart coordinated cyber-attacks on wind turbines: from detection to identification via machine learning. Sustain Energy Grids Netw. 2025;43:101775.

12. 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.

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

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

15. Utama C, Meske C, Schneider J, Schlatmann R, Ulbrich C. Explainable artificial intelligence for photovoltaic fault detection: a comparison of instruments. Sol Energy. 2023;249:139-51.

16. Noura HN, Allal Z, Salman O, Chahine K. Explainable artificial intelligence of tree-based algorithms for fault detection and diagnosis in grid-connected photovoltaic systems. Eng Appl Artif Intell. 2025;139:109503. https://doi.org/10.1016/j.engappai.2024.109503.

17. Shadi MR, Mirshekali H, Shaker HR. Explainable artificial intelligence for energy systems maintenance: a review on concepts, current techniques, challenges, and prospects. Renew Sustain Energy Rev. 2025;216:115668. https://doi.org/10.1016/j.rser.2025.115668.

18. Guan X, Sun X, Cao X, Zhai Q, Jia QS, Wu J, et al. Cyber-physical energy system for the energy revolution: a comprehensive solution for system design and operation. Cyber-Phys Energy Syst. 2025;1(1):3-13. https://doi.org/10.1016/j.cpes.2025.09.003.

19. Klaes M, Narayan A, Patil AD, Haack J, Lindner M, Rehtanz C, et al. State description of cyber-physical energy systems. Energy Inform. 2020;3(Suppl 1):16. https://doi.org/10.1186/s42162-020-00119-3.

20. Du D, Zhu M, Li X, Fei M, Bu S, Wu L, et al. A review on cybersecurity analysis, attack detection, and attack defense methods in cyber-physical power systems. J Mod Power Syst Clean Energy. 2023;11(3):727-43.

21. Wu G, Li ZS. Cyber-physical power system (CPPS): a review on measures and optimization methods of system resilience. Front Eng Manag. 2021;8(4):503-18.

22. Diaba SY, Shafie-Khah M, Elmusrati M. Cyber-physical attack and the future energy systems: a review. Energy Rep. 2024;12:2914-32.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

23. Alvarez-Alvarado MS, Apolo-Tinoco C, Ramirez-Prado MJ, Alban-Chacón FE, Pico N, Aviles-Cedeno J, et al. Cyber-physical power systems: a comprehensive review about technologies drivers, standards, and future perspectives. Comput Electr Eng. 2024;116:109149.

24. Gautam M. Deep reinforcement learning for resilient power and energy systems: progress, prospects, and future avenues. Electricity. 2023;4(4):336-80.

25. Afrin S, Al Muttaki MR, Anil AI, Hasan S. AI-powered cybersecurity for smart grid communication: a systematic review of intrusion detection and threat mitigation systems. Energy Convers Manag X. 2026;29:101416. https://doi.org/10.1016/j.ecmx.2025.101416.

26. Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recognit. 2018;84:317-31. https://doi.org/10.1016/j.patcog.2018.07.023.

27. Yuan X, He P, Zhu Q, Li X. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst. 2019;30(9):2805-24.

28. Ren K, Zheng T, Qin Z, Liu X. Adversarial attacks and defenses in deep learning. Engineering. 2020;6(3):346-60.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press