

Addressing Reliability Challenges of Deep Learning Models in Nuclear Applications

Jeong-ho LEE, Myungsoo KIM

KINAC, 1418 Yuseong-daero, Yuseong-gu, Daejeon, Korea 34101

E-mail: friend25kr@kinac.re.kr, myungsookim@kinac.re.kr

**Keywords: Nuclear Energy, Deep Learning, Reliability and Safety, Explainable AI (XAI), Physics-Informed Neural Networks (PINNs)*

1. Introduction

As global demand for low-carbon energy grows, the nuclear power sector is exploring digital technologies to support the long-term operation of existing plants and the development of economically viable small modular reactors (SMRs). In this context, deep learning (DL) may enhance operational efficiency by analyzing large and heterogeneous datasets. Potential applications include predictive maintenance, fuel-management optimization, anomaly detection, and operator decision support during abnormal or emergency conditions [2,4].

Despite these potential benefits, integrating DL into nuclear power plants raises important regulatory and safety concerns. The opaque, or "black-box," nature of multilayer neural networks can hinder verification, validation, and safety justification. Additional concerns include variability introduced during model development, performance degradation under distribution shift, and vulnerability to adversarial manipulation. These issues could erode safety margins if DL outputs are used without appropriate assurance measures and human oversight [2,15,21].

International organizations, including the IAEA and the OECD Nuclear Energy Agency (NEA), are examining the opportunities, safety implications, and governance needs associated with AI in nuclear applications [13,14,23,24]. This paper reviews relevant regulatory considerations and discusses technical measures for improving reliability, including explainable AI (XAI), physics-informed neural networks (PINNs), adversarial robustness, data governance, and lifecycle monitoring. By integrating regulatory and technical perspectives, the paper outlines practical considerations for the safe and reliable use of AI in nuclear energy systems.

2. Applications and Benefits of Deep Learning in Nuclear Fields

The adoption of DL in the nuclear sector is motivated by the need to improve the performance of existing facilities and to support the economic and safe operation of next-generation systems, including SMRs. Analysis of large operational datasets may contribute to condition monitoring, predictive maintenance, anomaly detection, fuel management, and decision support [2,4].

2.1. Enhancement of Operational Efficiency

Predictive maintenance is among the most widely discussed applications of DL in the nuclear sector. Models trained on historical and real-time condition-monitoring data can help identify degradation patterns and estimate the likelihood of component failure, thereby supporting risk-informed maintenance planning and reducing avoidable outages. However, the magnitude of these benefits depends on data quality, plant configuration, model validation, and integration with existing maintenance programs [5,6].

DL can also support operational optimization by analyzing complex relationships among core configurations, fuel assemblies, control-rod patterns, and operating constraints. AI-assisted methods have a long history in reactor-core design and fuel-management research; any proposed loading pattern must nevertheless be independently verified against applicable physics, safety limits, and licensing requirements [7].

2.2. Safety Enhancement and Accident Prevention

DL may strengthen nuclear safety through real-time monitoring of sensor and inspection data. Properly validated models can help detect incipient anomalies, including equipment degradation or abnormal thermal signatures, and can provide additional information to plant personnel. Such systems should supplement, rather than replace, qualified monitoring and protection functions [6].

DL may also support operator decision-making by processing plant parameters and procedure-related information during abnormal events. To be useful in safety-significant contexts, decision-support systems must present timely, understandable, and traceable information while preserving operator authority. Korean research on AI-based autonomous-operation platforms and the PRISM decision-support concept illustrates this direction of development [8,9].

3. Technical Challenges and Risks of Deep Learning Models

Although DL offers substantial potential for the nuclear industry, its use in safety-critical environments is constrained by technical and assurance challenges that must be addressed before deployment [2,10,15].

3.1. The Black-Box Dilemma and Transparency

A major obstacle to the adoption of DL in nuclear applications is the limited interpretability of many deep neural networks. Unlike conventional rule-based software, these models can contain millions of learned parameters, making it difficult to trace how a particular input produced a particular output. In a safety-critical domain, this opacity complicates safety justification, independent review, regulatory assessment, and operator trust [2,15,21].

3.2 Stochastic Development and Reliability (V&V and VVUQ)

DL development often involves stochastic elements, including random initialization, data shuffling, and stochastic optimization. Consequently, independently trained instances of the same architecture may not be identical, even when trained on the same dataset. Once trained and placed in a controlled inference configuration, however, many models can produce repeatable outputs for identical inputs. The principal assurance challenge is therefore not simply "non-determinism," but demonstrating adequate and reproducible performance across the intended operating domain [15,21].

Verification and Validation (V&V): The scale, nonlinearity, and data-dependent behavior of deep networks make exhaustive verification impractical. V&V should therefore combine requirements-based testing, independent datasets, robustness testing, traceability, configuration control, and evidence that the model remains within its approved operating envelope [15,21].

Uncertainty Quantification (UQ): Safety-relevant applications require methods for characterizing predictive uncertainty and identifying out-of-distribution inputs. The adequacy of these methods should be evaluated as part of verification, validation, and uncertainty quantification (VVUQ), with acceptance criteria linked to the intended use [15,21].

3.3 Data Quality and Model Drift

The performance of a DL model depends strongly on the quality, representativeness, provenance, and governance of its training and evaluation data [2,10,15]. Two recurring concerns are data imbalance and distribution shift:

Data Imbalance: Normal operating data are abundant, whereas abnormal and accident-related data are scarce. A model trained on such an imbalanced dataset may perform well on routine conditions yet fail to recognize rare safety-significant events. Biases, labeling errors, or gaps in the dataset can also propagate into model outputs [2,10].

Distribution Shift and Model Drift: Model performance can degrade when the operational data distribution differs from the training distribution or changes over time. Without monitoring and predefined response measures, this degradation may

lead to inaccurate predictions or unexpected behavior [15,21].

3.4 Cybersecurity and Integrity Risks

Integrating DL into nuclear power systems creates cybersecurity concerns associated with training data, model artifacts, software dependencies, and deployment infrastructure. These concerns extend conventional industrial-control-system security by adding attack paths that target the AI lifecycle itself [12,16,21].

Data and Model Manipulation: Attackers may poison training data, manipulate model parameters, or craft adversarial inputs to influence model behavior. In nuclear applications, such manipulation could generate misleading maintenance indications or decision-support outputs; therefore, data provenance, integrity checks, access control, and independent testing are essential [12,16].

Supply-Chain Vulnerabilities: DL systems depend on third-party libraries, pretrained models, development tools, and hardware. Compromised dependencies or hidden backdoors can create vulnerabilities throughout the development and deployment lifecycle [12,21].

Defense in Depth: Network-boundary controls remain necessary but are not sufficient for AI-enabled systems. Security measures should also address model robustness, data integrity and availability, secure development practices, controlled software and model updates, protected communications, and hardware security [12,16,21].

Lifecycle Management: Maintaining system integrity requires continuous performance and security monitoring, configuration control, periodic reassessment, and predefined actions when drift, anomalous inputs, or cyber threats are detected [12,16,21].

4. Regulatory Frameworks for Nuclear AI Safety

Deploying AI in the nuclear sector requires a safety-assurance framework that combines technical evidence with organizational capability and regulatory readiness. Because DL introduces data-dependent behavior, limited interpretability, and new lifecycle dependencies, existing regulatory processes may need to be supplemented with AI-specific guidance and evidence [13,15,21].

4.1 International Collaborative Efforts and Proactive Regulation

International organizations are developing collaborative approaches to AI in nuclear applications. The IAEA has convened activities on AI deployment, safety implications, and computer security, while the OECD NEA has established initiatives that bring regulators, operators, researchers, and vendors together. The NEA's International RegLab project promotes early regulatory engagement so that safety and licensing

considerations can be identified during technology development [13,14,23,24].

4.2 Standardization of Verification and Validation (V&V)

A key regulatory need is a standardized, risk-informed approach to V&V and VVUQ for AI-enabled nuclear applications. Existing cross-sector AI standards can provide useful foundations, but they do not replace nuclear-specific safety requirements or application-specific evidence [21].

Evaluation and Robustness: IEEE standards and emerging guidance address topics such as deep-learning evaluation and robustness testing. Their applicability should be assessed against the safety classification, intended function, and operating environment of the nuclear application [21].

Transparency and Explainability: Standards and guidance on XAI can support documentation and assessment of explanation methods, but an explanation must be evaluated for fidelity, stability, usability, and relevance to the safety decision at hand [21].

Risk Management and Governance: ISO/IEC 23894:2023 provides guidance on AI risk management, while ISO/IEC 42001:2023 specifies requirements for an AI management system. These standards can support organizational governance when integrated with nuclear quality-assurance and safety-management processes [21].

4.3 Paradigm Shift in Trust and Oversight

Trust in an AI-enabled nuclear system should not be based on predictive accuracy alone. It should be supported by evidence concerning data quality, model performance, robustness, cybersecurity, configuration control, human factors, and lifecycle governance [15,21].

By integrating these international standards and proactive regulatory models, the nuclear industry can create a stable environment where deep learning technologies can be safely utilized to enhance operational efficiency and safety.

5. Technical Strategies for Risk Mitigation

Mitigating the risks of DL in nuclear applications requires multiple complementary controls. Relevant measures include interpretable modeling and post-hoc

explanation, incorporation of physical constraints, robustness testing, cybersecurity controls, and lifecycle monitoring [25,26,28].

5.1 Explainable AI (XAI) and Interpretability

XAI methods can help developers, reviewers, and operators examine the factors influencing a model's output. The selected method should match the intended user and safety decision, and its fidelity and limitations should be evaluated rather than assumed [25].

Interpretable and Surrogate Models: Where feasible, transparent models such as linear or logistic regression and decision trees can be used directly or as local surrogates for a more complex model. Surrogates can aid understanding, but they may not faithfully represent model behavior outside the region in which they were fitted [25].

Post-hoc Explanation Techniques: LIME estimates local behavior by fitting an interpretable model around a selected prediction, whereas SHAP uses Shapley-value concepts to attribute a prediction to input features. Visualization can further support analysis of feature importance, learned representations, and data bias. These methods can be computationally demanding and may yield unstable or misleading explanations if applied without appropriate validation [25].

5.2 Physics-Informed Neural Networks (PINNs)

Physics-informed neural networks incorporate governing equations, boundary conditions, or other physical knowledge into model training, commonly through the loss function. This approach can improve physical consistency and data efficiency in applications such as fluid flow, heat transfer, and reactor transient prediction. Nevertheless, PINN outputs are not automatically guaranteed to satisfy all physical laws; residual errors, numerical limitations, and domain coverage must still be quantified and validated [28].

5.3 Model Robustness and Operational Lifecycle Management

Long-term reliability requires coordinated measures for robustness, monitoring, configuration management, and cybersecurity [26,28].

Adversarial Defense: Adversarial training exposes a model to perturbed examples during training and can improve robustness against specified attack classes. It should be complemented by input validation, preprocessing, robustness testing, access controls, and monitoring because no single defense is effective against all attacks [26].

Lifecycle Monitoring: Organizations should monitor model performance, data quality, distribution shift, and out-of-range inputs throughout operation. Monitoring thresholds, responsibilities, retraining criteria, rollback procedures, and

reapproval requirements should be defined before deployment [15,21].

Cybersecurity Integration: Technical infrastructure should be protected through secure development practices, network segmentation, authenticated and encrypted communications, controlled updates, hardware protections where justified, and measures that prevent unauthorized access to data and model parameters [16,21].

6. Conclusion and Policy Recommendations

The integration of DL into the nuclear industry offers opportunities in predictive maintenance, operational optimization, anomaly detection, and decision support, while also introducing risks related to opacity, data dependence, distribution shift, and cybersecurity [2,4,12,15]. Safe adoption therefore requires a transition from performance-centered evaluation to reliability- and assurance-centered evaluation.

To ensure the safe deployment of DL-based systems in nuclear applications, the following policy and technical recommendations are proposed:

6.1 Technical Reliability and Standardization

Robust and Interpretable Architectures: Research should evaluate combinations of XAI and physics-informed learning for nuclear applications. These methods may improve interpretability and physical consistency, but their limitations and failure modes must be characterized through application-specific VVUQ [25,28].

Data Governance and Benchmarking: Nuclear organizations need controlled processes for data provenance, quality, labeling, access, retention, and change management. Shared or representative benchmark datasets, where security and proprietary constraints permit, could support independent comparison and cross-institutional evaluation [15,21].

6.2 Human Capital and Organizational Readiness

Interdisciplinary Capacity Building: AI developers, nuclear engineers, human-factors specialists, cybersecurity experts, and regulators need a shared understanding of system objectives, hazards, and assurance evidence. Joint training and early collaboration can reduce gaps between AI development and nuclear safety practice [13,14].

Human-Centered Design: AI should support, rather than displace, accountable human decision-making in safety-significant operations. Interfaces should communicate recommendations, uncertainty, limitations, and supporting evidence in ways that are compatible with operator tasks and workload [17,18].

6.3 Proactive Regulatory Frameworks

Early Regulatory Involvement: Regulators should engage with developers early enough to clarify the intended use, safety classification, evidence expectations, and applicable requirements. The OECD NEA International RegLab project provides one model for structured dialogue among regulators and technology stakeholders [22,23].

International Cooperation: Cooperation among regulators, technical support organizations, standards bodies, industry, and research institutions can support consistent terminology, shared lessons, and harmonized approaches to V&V and VVUQ while respecting national legal frameworks [13,14,19,24].

In conclusion, DL may improve the efficiency and safety of existing and next-generation nuclear systems, but deployment should proceed incrementally and in proportion to safety significance. A defensible approach combines technical VVUQ, high-quality data, interpretability, cybersecurity, human oversight, configuration control, and proactive regulatory engagement.

REFERENCES

- [1] Accelerant. (n.d.). AI in Nuclear Plants. <https://discoveraccelerant.com/ai-in-nuclear-plants/>
- [2] International Atomic Energy Agency (IAEA). (n.d.). Challenges in the Deployment of AI/ML Technologies in Nuclear Power. https://www-pub.iaea.org/MTCD/publications/PDF/p15866-PUB2119_web.pdf
- [3] 첨단정보기술연구소. (n.d.). AI 기반 다목적 최적화 및 보안 체계 강화. <https://www.cbist.or.kr/home/board/download.do?mncd=1284&fno=7866&bid=00000322&did=00005626>
- [4] Kubelwa, N. N. (n.d.). Enhancing Nuclear Power Production with Artificial Intelligence. IAEA Bulletin. <https://www.iaea.org/bulletin/enhancing-nuclear-power-production-with-artificial-intelligence>
- [5] techUK. (n.d.). AI-powered Predictive Maintenance for Nuclear. <https://www.techuk.org/resource/ai-powered-predictive-maintenance-for-nuclear.html>
- [6] Korea Hydro & Nuclear Power. (n.d.). Automatic Predictive-Diagnosis System: Operation and Expected Effects [Video]. https://www.youtube.com/watch?v=Vov3sulfi_g
- [7] Nissan, E. (2019). An Overview of AI Methods for in-Core Fuel Management: Tools for the Automatic Design of Nuclear Reactor Core Configurations for Fuel Reload, (Re)arranging New and Partly Spent Fuel. Designs, 3(3), 37. <https://doi.org/10.3390/designs3030037>
- [8] Park, J. H. (2024). Development Status of the AI-Based Operator Support System PRISM [NuPIC presentation summary]. <https://www.inpnews.kr/news/articleView.html?idxno=6103>
- [9] Korea Atomic Energy Research Institute. (2019). Development of AI-Based Autonomous Operation Algorithms and Platform [Report summary]. <https://scienceon.kisti.re.kr/srch/selectPORSrchReport.do?cn=TRKO201900027191>
- [10] Alabdulkarim, M., & Assi, A. (2025). AI Application in Nuclear Emergency Preparedness: Challenges and Future

- Outlook. *Frontiers in Mechanical Engineering*, 11. <https://doi.org/10.3389/fmech.2025.1594397>
- [11] OECD Nuclear Energy Agency. (2025). Nuclear Innovation and Newly Emerging Disruptive Technologies. https://www.oecd-neo.org/upload/docs/application/pdf/2025-11/nuclear_innovation_and_newly_emerging_disruptive_technologies_recent_initiatives_by_industry_research_laboratories_and_gove.pdf
- [12] IAEA. (n.d.). How Artificial Intelligence Will Change Information and Computer Security in the Nuclear World. IAEA Bulletin. <https://www.iaea.org/bulletin/how-artificial-intelligence-will-change-information-and-computer-security-in-the-nuclear-world>
- [13] IAEA. (2022). The Role of the IAEA in the Deployment of AI for Nuclear Power. https://www-pub.iaea.org/MTCD/publications/PDF/p15866-PUB2119_web.pdf
- [14] OECD Nuclear Energy Agency. (n.d.). Task Force on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering. https://www.oecd-neo.org/jcms/pl_77779/
- [15] U.S. Nuclear Regulatory Commission. (2024). Artificial Intelligence/Machine Learning Risk-Transfer Considerations. <https://www.nrc.gov/docs/ML2424/ML24241A252.pdf>
- [16] IAEA. (2024). New Research Project on Computer Security for Nuclear AI. <https://www.iaea.org/newscenter/news/new-research-project-on-computer-security-for-nuclear-ai>
- [17] Barbosa, M. G., & Vidal, M. C. R. (2018). Human-Centered Design for Nuclear Power Plant Control Rooms [Conference paper]. <https://www.semanticscholar.org/paper/a48280c4905132c97399e6bf73a966d9acb8503d>
- [18] Handa, K. (2007). Human-Machine Interaction in Nuclear Power Plants. https://www.researchgate.net/publication/228628218_Human-machine_interaction_in_nuclear_power_plants
- [19] IAEA. (n.d.). Technical Meeting on the Safety Implications of the Use of AI in Nuclear Power Plants. <https://aiforgood.itu.int/about-us/un-ai-actions/iaea/>
- [20] U.S. Nuclear Regulatory Commission. (2013). Secure Development and Operational Environment Guidance. <https://www.nrc.gov/docs/ML1307/ML13073A210.pdf>
- [21] U.S. Nuclear Regulatory Commission. (2024). AI Regulatory Gap Analysis: Review of Standards. <https://www.nrc.gov/docs/ML2429/ML24290A059.pdf>
- [22] OECD Nuclear Energy Agency. (n.d.). International RegLab Project. https://www.oecd-neo.org/jcms/pl_104274/
- [23] OECD Nuclear Energy Agency. (2025). AI in Nuclear Operations: RegLab Workshop Convenes Regulators, Operators and Vendors. https://www.oecd-neo.org/jcms/pl_111815/
- [24] OECD Nuclear Energy Agency. (n.d.). Harnessing AI: First Workshop of the New NEA Joint Project AIxpertise. https://www.oecd-neo.org/jcms/pl_111870/
- [25] Chaudhary, A., Han, J., Kim, S., Kim, A., & Choi, S. (2024). Anomaly Detection and Analysis in Nuclear Power Plants. *Electronics*, 13(22), 4428. <https://doi.org/10.3390/electronics13224428>
- [26] Pozdnyakov, V., Kovalenko, A., Makarov, I., Drobyshevskiy, M., & Lukyanov, K. (2024). Adversarial Attacks and Defenses in Fault Detection and Diagnosis: A Comprehensive Benchmark on the Tennessee Eastman Process. *IEEE Open Journal of the Industrial Electronics Society*, 5, 428-440. <https://doi.org/10.1109/OJIES.2024.3401396>
- [27] Argonne National Laboratory. (n.d.). AI-Assisted Safety Modeling and Analysis of Advanced Reactors. <https://www.anl.gov/ntns/aiassisted-safety-modeling-and-analysis-of-advanced-reactors>
- [28] Palmer, J. S., Stewart, R. H., & Shields, A. J. B. (2025). Physics-Informed Recurrent Neural Networks to Predict Reactor Operations of the AGN-201 Nuclear Reactor (INL/CON-25-82713-Revision-0). Idaho National Laboratory. https://inldigitallibrary.inl.gov/sites/sti/sti/Sort_151656.pdf