AI in Critical Infrastructure – Explainability and Models

Jens Wala
Research Training Group KRITIS, Technical University of Darmstadt, Dolivostraße 15, 64293 Darmstadt, Germany
E-Mails: wala@kritis.tu-darmstadt.de

Abstract: Critical infrastructures such as power grids, transportation networks, and water systems are essential to national economies and societal well-being. Integrating Artificial Intelligence (AI) into these systems could enhance productivity and operational resilience. However, the adoption of AI in critical infrastructures necessitates a focus on explainability to ensure transparency, trust, and regulatory compliance. This paper explores inherently explainable models (IEMs) and post hoc explainable models (PHEMs) within the domain of critical infrastructures. By examining regulatory requirements, analyzing different AI models designed for explainability, and comparing these models, this paper provides a comprehensive overview of strategies for selecting AI systems that enhance transparency and compliance. The findings underscore the importance of choosing appropriate AI models to ensure safe, reliable, and legally accountable AI implementation in critical infrastructure, ultimately supporting societal functions and public safety.
Keywords: Explainable Artificial Intelligence, Critical Infrastructure, Interpretability, Resilience, Transparency

Erschienen in Tagungsband 35. Forum Bauinformatik 2024, Hamburg, Deutschland, DOI: 10.15480/882.13498

1 Introduction

Critical infrastructures such as power grids, transportation networks, and water systems form the backbone of national economies and societal well-being. These infrastructures are complex socio-technical systems that require robust monitoring and control mechanisms to ensure efficient and uninterrupted service [1]. Traditionally, automation in critical infrastructure control has relied on control theory and rule-based systems [2]. Control theory is a branch of mathematics dealing with the behavior of dynamic systems with inputs. It uses feedback loops to maintain a system's desired output despite disturbances. In a rule-based system, the logic is explicitly programmed, making the decision-making process easily understandable and verifiable.

Integrating Artificial Intelligence (AI) into these systems could revolutionize their operational capabilities, leading to significant enhancements in productivity and potential reductions in human

resource needs. This evolution is particularly pertinent in the ongoing global skilled worker shortage, which has placed additional pressure on maintaining operational resilience [3].

However, adopting AI technologies in critical infrastructures is not without challenges. Key among these is the need for explainability [3]. Explainability pertains to the ability of stakeholders to understand and trust the decisions made by AI systems [3]. This transparency is crucial for operational trust and compliance with legal standards [4]. Recent legislative developments, such as the European Union's Artificial Intelligence Act (AIA) and Canada's Artificial Intelligence and Data Act (AIDA), have set forth stringent requirements for AI explainability [4], [5], [6]. These regulations aim to ensure that AI systems are effective but also accountable and fair, mitigating risks associated with opaque decision-making processes.

This paper delves into the dual paradigms of inherently explainable models (IEMs) and post hoc explainable models (PHEMs) within the critical infrastructure domain. The paper outlines strategies for selecting AI systems that enhance transparency and compliance by exploring these models, thereby supporting robust and legally sound AI implementation in critical infrastructures. The ensuing sections will review regulatory requirements, analyze different AI models designed for explainability, and compare these models in terms of their practical applications, challenges, and benefits in real-world settings.

2 Related Works

The paper by Panigutti et al. [4] explores how the EU's AIA addresses the challenges of AI system opacity. It discusses the Act's transparency and human oversight requirements, particularly for high-risk AI systems. The authors argue that while the Act does not mandate the use of explainable AI (XAI) techniques or transparent-by-design models, it demands comprehensive documentation and effective oversight measures. The paper highlights the limitations of current XAI methodologies. It emphasizes that AI systems can achieve trustworthiness through appropriate transparency and oversight rather than relying solely on technical explainability.

Zschech et al. [7] compare five generalized additive models (GAM) to four black-box learning approaches, classic decision trees, and linear regression, given their performance and degree of interpretability. Both regression and classification tasks have been tested. On twelve datasets, they could show that GAMs outperform the other tested approaches on five of them and find that the discrepancy between the model's performances is minuscule. Tests have been conducted with the recommended parameters provided by the developers of the respective models. Nonetheless, they show that GAM-based models offer a promising direction for achieving both interpretability and competitive predictive performance. The study highlights the potential of intrinsically interpretable models as ethical and technically viable alternatives to black-box models, particularly in critical decision-making areas.

3 Regulatory Requirements for AI Explainability

The EU proposed a framework called the Artificial Intelligence Act (AIA) to regulate artificial intelligence in 2021 [5]. The European Parliament unanimously approved the Act on 13 March 2024 [5]. The Act classifies the use of AI into different risk stages depending on the application [3]. Minimal risk applications are those where the AI's impact is deemed negligible, such as spam filters or AI-enabled games. Limited risk applications involve systems that may pose some risk but are not likely to result in significant harm. These include chatbots and virtual assistants. High-risk applications involve AI systems that significantly impact the rights and safety of individuals [8]. These include AI used in healthcare, transportation, and power grids. Unacceptable risk applications are outright prohibited due to their potential for significant harm, such as social scoring by governments or systems that manipulate human behavior.

Critical infrastructure is categorized as high-risk due to the potentially severe consequences that failures or mismanagement in these systems can cause. High-risk systems must meet stringent requirements for transparency and interpretability [8]. This ensures that the decisions made by AI systems can be understood and trusted by human operators [4]. Interpretability is crucial for compliance with legal standards and maintaining operational trust, especially in sectors where AI decisions directly affect human lives and societal functions.

This leads to the following research questions:

- **RQ1:** How can AI models be designed to ensure interpretability and compliance with the regulatory standards set by the EU AIA in high-risk critical infrastructure applications?
- **RQ2:** What are the best practices for implementing and maintaining explainable AI systems in critical infrastructure to enhance transparency, trust, and operational efficiency?

4 AI Models for Explainability

Deploying AI models in critical infrastructure necessitates a balance between performance and interpretability. Given the high risk associated with critical infrastructure operations, stakeholders must be able to understand and trust the decisions made by AI systems. This chapter delves into two paradigms of AI explainability: Inherently Explainable Models (IEMs) and Post Hoc Explainable Models (PHEMs), exploring their applications, benefits, and challenges in the context of critical infrastructure.

4.1 Inherently Explainable Models (IEMs)

Inherently explainable or interpretable models (IEMs) are designed with transparency as a fundamental feature. These models are constructed so humans can easily understand their operations and decision-making processes. IEMs' characteristics include simplicity, transparency, and ease of interpretation. This allows stakeholders to trust and verify the outputs of these models, making them particularly suitable for high-risk environments like critical infrastructure.

Rule-based automation involves systems that follow predefined rules to perform tasks. These rules are typically "if-then" statements derived from expert knowledge. Rule-based models are easy to understand and implement because they rely on logical conditions that dictate the system's actions. For instance, in a water supply system, a rule might state that if the water level in a reservoir drops below a certain point, a pump should be activated.

Decision trees and linear regression are fundamental types of inherently explainable models. Decision trees are structured as a series of binary decisions, represented as branches, which lead to different outcomes, represented as leaves [9]. The tree's structure makes it straightforward to follow the decision-making process from the root to any leaf, providing clear insight into how a particular decision was reached [9]. Linear regression, on the other hand, models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Its simplicity and the direct nature of the relationships it models make linear regression highly interpretable.

Generalized Additive Models (GAMs) extend linear models by including non-linear relationships between the dependent and independent variables while maintaining interpretability [7]. GAMs achieve this by representing the model as a sum of smooth functions of the predictor variables [7]. The advantage of GAMs lies in their ability to provide transparent predictions that can capture complex patterns in the data without sacrificing interpretability [7]. In the context of critical infrastructure, GAMs can be particularly useful. For example, in water supply monitoring and control, a GAM could predict water demand based on temperature, time of day, and historical usage patterns. By understanding these relationships, an autonomous system can make resource allocation and system management decisions and provide operators with the relevant decision information.

These inherently explainable models are crucial for maintaining transparency and trust in critical infrastructure systems. They ensure that the decision-making processes are effective but also understandable and verifiable by human operators, aligning with the stringent requirements for transparency and accountability set forth by regulations like the EU AIA.

4.2 Post-Hoc explainable models

In contrast to inherently explainable models, black box models are complex, and their decision-making processes are not easily interpretable [10]. These models, such as deep neural networks and complex ensemble methods, often yield highly accurate predictions but at the cost of transparency [10]. To address this issue, various methods have been developed to make the decisions of these models more understandable.

Post-hoc explainability refers to techniques applied after a model has made its predictions to interpret and explain the results.

One prominent post-hoc explainability technique is Local Interpretable Model-agnostic Explanations (LIME) [10], [11]. LIME works by locally approximating the black box model around a specific prediction with a simpler, interpretable model. By perturbing the input data slightly and observing the

changes in the model's output, LIME constructs a local surrogate model that explains the behavior of the black box model for that particular instance [11]. This approach is valuable for providing instance-level explanations, making understanding why a model made a specific prediction easier.

Another widely used method is SHapley Additive exPlanations (SHAP) [10], [11]. SHAP values are derived from cooperative game theory and represent the contribution of each feature to the model's prediction [11]. By considering all possible combinations of feature values, SHAP provides a consistent and fair attribution of feature importance [11]. This method ensures that the explanations are accurate and theoretically grounded, making it a powerful tool for understanding complex models.

Integrated Gradients is another technique that attributes a neural network's prediction to its input features [12]. It works by computing the gradients of the model's output to its inputs, integrated along a path from a baseline (such as a zero vector) to the actual input [12]. This method highlights the features that contribute most to the prediction, offering insights into the inner workings of the neural network and helping to explain the model's decisions more intuitively.

Post-hoc explainability techniques are essential for making the decisions of black box models more transparent. They enable stakeholders to gain insights into the functioning of complex AI systems, ensuring that these systems can be trusted and their decisions can be validated.

5 Comparative Analysis

Following the exploration of Post-Hoc Explainable Models (PHEMs) and Inherently Explainable Models (IEMs), the following table provides a detailed comparison. This comparison uses the following aspects to highlight differences between the explainability approaches:

Design Philosophy: This aspect describes the fundamental AI model creation approach. It contrasts the built-in transparency of IEMs with the performance-first approach of PHEMs.

Transparency refers to how easily humans can understand the model's decision-making process. It compares the clear, interpretable process of IEMs with the initially opaque nature of PHEMs.

Compliance: This aspect addresses how well the models meet regulatory requirements for transparency and accountability, particularly in regulations like the EU AIA.

Complexity Handling: This aspect compares how well each model type can handle complex patterns and large datasets, highlighting the trade-offs between simplicity and performance.

Performance: This refers to the overall effectiveness of the models in completing their assigned tasks, especially for highly complex problems.

Trust and Verification: This aspect addresses how easily stakeholders can trust and verify the decisions made by the AI models, which is crucial in critical infrastructure applications.

Aspect	Inherently Explainable Models (IEMs)	Post-Hoc Explainable Models (PHEMs)
Design Philosophy	Built with transparency and simplicity from the ground up.	Focused on performance; explanations are derived after the model is built.
Transparency	High transparency; the decision-making process is clear and interpretable.	Limited initial transparency; relies on additional techniques for explanations.
Compliance	Meets regulatory requirements for transparency and accountability.	Requires supplementary methods to meet transparency standards.
Complexity Handling	May struggle with very complex patterns; simplicity can limit performance.	Excels in handling complex patterns and large datasets.
Performance	Generally lower than black-box models for highly complex tasks.	Higher performance due to complexity but at the cost of initial opacity.
Trust and Verification	High, as decisions can be easily traced and understood.	Trust needs to be established through post-hoc explanations.

The EU AIA emphasizes the need for transparency and accountability in AI systems, particularly those classified as high-risk, such as those used in critical infrastructure. In this context, inherently explainable models (IEMs) are uniquely positioned to meet these requirements due to their built-in transparency. By design, IEMs offer a clear and interpretable decision-making process that aligns with the stringent transparency standards mandated by the AIA. This intrinsic interpretability is crucial for critical infrastructure systems where decisions impact public safety, security, and overall societal well-being. Human operators and regulatory bodies can verify and understand the actions of IEMs, which fosters trust and ensures compliance with legal standards.

In contrast, post-hoc explainable models (PHEMs) provide explanations after the model has made its predictions. While PHEMs can offer high performance and handle complex tasks effectively, they require additional techniques to render their decisions interpretable. This adds a layer of complexity and potential uncertainty, making them less straightforward to deploy in highly regulated environments like critical infrastructure.

Therefore, within the framework of the EU AIA, IEMs are the preferred choice for operating critical infrastructure control and monitoring systems. Their inherent transparency provides a robust

foundation for compliance, making them reliable tools for maintaining operational trust and legal accountability.

6 Discussion

The integration of AI in critical infrastructure presents both opportunities and challenges. The analysis reveals a preference for Inherently Explainable Models (IEMs) in this context, primarily due to their alignment with regulatory requirements like the EU AIA. Generalized Additive Models (GAMs) offer a promising balance between performance and interpretability.

The skilled worker shortage in critical infrastructure sectors adds urgency to AI adoption. IEMs potentially allow for faster integration and training of existing staff. However, this must be balanced against possible performance trade-offs compared to more complex models.

The tension between performance and explainability remains a key challenge. While Post-Hoc Explainable Models (PHEMs) often offer superior performance in complex tasks, their lack of inherent transparency poses challenges for regulatory compliance and stakeholder trust. Future research should focus on narrowing this gap, potentially through hybrid approaches.

Looking ahead, evolving regulations will continue to shape AI development in critical infrastructure, potentially driving global standards towards greater emphasis on explainability. However, this focus should not compromise system performance or security.

In conclusion, while IEMs offer the most promising path for AI in critical infrastructure, the field remains dynamic. Successful implementation will require collaboration between AI developers, infrastructure managers, policymakers, and ethicists to ensure AI enhances critical infrastructure's resilience, efficiency, and trustworthiness.

7 Conclusion

This paper has explored the role of explainable AI models in critical infrastructure, highlighting the preference for Inherently Explainable Models (IEMs) due to their alignment with regulatory requirements and operational needs. While IEMs, particularly Generalized Additive Models, balance performance and interpretability, the tension between model complexity and explainability remains challenging.

The skilled worker shortage in critical infrastructure sectors underscores the urgency of AI adoption, with IEMs potentially offering easier integration and staff training. However, evolving regulations will continue to shape AI development in this field, emphasizing the need for transparency without compromising performance or security.

Moving forward, successful AI implementation in critical infrastructure will require collaboration among AI developers, infrastructure managers, policymakers, and ethicists. Future research should focus on enhancing IEM capabilities, standardizing explainability evaluation frameworks, and

studying the long-term impacts of explainable AI in critical systems. This approach is key to leveraging AI to enhance critical infrastructure's resilience, efficiency, and trustworthiness while meeting regulatory demands.

References

[1] A. Fekete, Kritische Infrastruktur und Versorgung der Bevölkerung: Klimawandel, Epidemien, digitale Transformation und das Risikomanagement. in essentials. Berlin, Heidelberg: Springer Berlin Heidelberg, 2022. doi: 10.1007/978-3-662-65047-9.

[2] Simrock, “Control Theory,” CERN. Accessed: Jun. 25, 2024. [Online]. Available: https://cds.cern.ch/record/1100534/files/p73.pdf

[3] P. Laplante and B. Amaba, “Artificial Intelligence in Critical Infrastructure Systems,” Computer, vol. 54, no. 10, pp. 14–24, Oct. 2021, doi: 10.1109/MC.2021.3055892.

[4] Cecilia Panigutti, Ronan Hamon, and Isabelle Hupont, “The role of explainable AI in the context of the AI Act,” presented at the Conference on Fairness, Accountability and Transparency, Jun. 2023. doi: 10.1145/3593013.3594069.

[5] “EU AI Act: first regulation on artificial intelligence | News | European Parliament.” Accessed: Jun. 16, 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[6] Innovation, Science and Economic Development Canada, “The Artificial Intelligence and Data Act (AIDA) – Companion document.” Accessed: Dec. 08, 2023. [Online]. Available: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

[7] P. Zschech, S. Weinzierl, N. Hambauer, S. Zilker, and M. Kraus, “GAM(e) changer or not? An evaluation of interpretable machine learning models based on additive model constraints.” arXiv, Apr. 19, 2022. doi: 10.48550/arXiv.2204.09123.

[8] “ANNEX III AI-Regulation (Proposal) - HIGH-RISK AI SYSTEMS REFERRED TO IN ARTICLE 6(2).” Accessed: Aug. 10, 2023. [Online]. Available: https://lexparency.org/eu/52021PC0206/ANX_III/

[9] J. R. Quinlan, “Simplifying decision trees,” Int. J. Man-Mach. Stud., vol. 27, no. 3, pp. 221–234, Sep. 1987, doi: 10.1016/S0020-7373(87)80053-6.

[10] A. B. Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.” arXiv, Dec. 26, 2019. Accessed: Dec. 11, 2023. [Online]. Available: http://arxiv.org/abs/1910.10045

[11] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions.” arXiv, Nov. 24, 2017. doi: 10.48550/arXiv.1705.07874.

[12] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks.” arXiv, Jun. 12, 2017. doi: 10.48550/arXiv.1703.01365.